mathematics

MDPI

Article

State of Charge Prediction for Electric Vehicles Based on Integrated Model Architecture

Min Wei [1,2,3,†], Yuhang Liu [1,2,†], Haojie Wang [1,2], Siquan Yuan [1,2] and Jie Hu [1,2,*]

[1] School of Automotive Engineering, Wuhan University of Technology, Wuhan 430070, China; minwei@whut.edu.cn (M.W.); 13050182710@163.com (Y.L.); wang13080139785@163.com (H.W.); siquan_yuan@whut.edu.cn (S.Y.)
[2] Hubei Key Laboratory of Advanced Technology for Automotive Components, Wuhan University of Technology, Wuhan 430070, China
[3] Technical Development Center, Shanghai Automotive Industry Corporation, General Wuling Automobile Co., Ltd., Liuzhou 545007, China
* Correspondence: auto_hujie@whut.edu.cn
† These authors contributed equally to this work as co-first authors.

Abstract

To enhance the accuracy of SOC prediction in EVs, which often suffers from significant discrepancies between displayed and actual driving ranges, this study proposes a data-driven model guided by an energy consumption framework. The approach addresses the problem of inaccurate remaining range prediction, improving drivers' travel planning and vehicle efficiency. A PCA-GA-K-Means-based driving cycle clustering method is introduced, followed by driving style feature extraction using a GMM to capture behavioral differences. A coupled library of twelve typical driving cycle style combinations is constructed to handle complex correlations among driving style, operating conditions, and range. To mitigate multicollinearity and nonlinear feature redundancies, a Pearson-DII-based feature extraction method is proposed. A stacking ensemble model, integrating Random Forest, CatBoost, XGBoost, and SVR as base models with ElasticNet as the meta model, is developed for robust prediction. Validated with real-world vehicle data across −21 °C to 39 °C and four driving cycles, the model significantly improves SOC prediction accuracy, offering a reliable solution for EV range estimation and enhancing user trust in EV technology.

Keywords: pure electric vehicle; battery state estimation; data-driven approach; remaining driving range prediction; stacking model

MSC: 68T07; 68T05

check for updates

Academic Editor: Jinfeng Liu

Received: 29 May 2025
Revised: 18 June 2025
Accepted: 3 July 2025
Published: 4 July 2025

Citation: Wei, M.; Liu, Y.; Wang, H.; Yuan, S.; Hu, J. State of Charge Prediction for Electric Vehicles Based on Integrated Model Architecture. *Mathematics* **2025**, *13*, 2197. https://doi.org/10.3390/math13132197

Copyright: © 2025 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Carbon dioxide is one of the key factors influencing global climate change [1]. In 2022, vehicle CO_2 emissions accounted for one-fifth of the world's total CO_2 emissions [2]. Studies show that EVs, as an effective solution to reduce air pollution and carbon emissions, have seen rapid development in recent years. However, during this process, range anxiety among drivers has become increasingly prominent due to bottlenecks in battery technology. As a critical metric for measuring driving range, accurately predicting the battery SOC can help drivers plan their trips more reasonably, thereby alleviating range anxiety [3].

Range anxiety refers to the degree of concern drivers have about whether their current battery SOC can reach their destination. This "anxiety" stems from factors such as

the prediction accuracy of electric vehicle SOC and remaining range, the distribution of charging stations, and charging power [4]. The prediction accuracy of SOC is influenced by factors like battery health status [5], driver's driving style, current and future road conditions, and environmental temperature [6]. To alleviate range anxiety and improve the prediction accuracy of SOC and remaining range, it is crucial to enhance these metrics. To achieve this goal, current research methods primarily include modeling approaches and data-driven methods [7]. Model-based remaining range models calculate from both energy consumption and energy storage perspectives and then use their ratio as the remaining range value. However, this method often results in significant cumulative errors due to calculation discrepancies between energy consumption and storage. Additionally, it fails to capture the complex nonlinear relationship between remaining range and environmental temperature. In contrast, data-driven methods focus on data, offering advantages in addressing this issue [8]. Therefore, this paper primarily explores data-driven approaches.

Zhao et al. [9] proposed a machine learning algorithm model that integrates XGBoost and LightGBM. This model predicts the remaining driving range of real-world vehicles by constructing features such as the cumulative output energy of motors and batteries, different driving modes, and battery temperature. Eagon et al. [10] introduced a new method using two RNNs to predict SOC and remaining driving range. Sun et al. [11] predicted future driving conditions based on online traffic information using Hidden Markov Models, achieving predictions for remaining driving range. De et al. [12] proposed a multiple linear regression model that combines neural network methods to predict unknown driving conditions and then forecasts RDR. Compared with model-based methods, data-driven approaches show certain advantages. Bustos et al. [13] utilized LSTM to predict vehicle speed and energy consumption, which were used as inputs to the LightGBM model, thus building a novel data-driven architecture for predicting the remaining driving range. Jain et al. [14] employed six ML algorithms: GPR, EBa, EBo, ANN, SVM, and LR to estimate the SOC in lithium-ion batteries. Sun et al. [15] proposed an ML method using a Gradient Boosting Decision Tree to predict the driving range of BEVs. The model considers weather, driving habits, and battery condition, outperforming traditional multiple linear regression models. The ML model shows a deviation between -1.41 km and 1.58 km, with an average deviation of 0.7 km, while the error for the linear model ranges from -3.6975 km to 3.3865 km. Yavasoglu H A et al. [8] used a data-driven approach to build a remaining range prediction model. They first employed a decision tree algorithm to identify and determine driving conditions and used a range estimation algorithm to identify driver styles. Then, they incorporated battery SOC, SOH, driving conditions, environmental temperature, and driver style as feature inputs into the model, constructing an ANN model.

In current research on SOC and remaining driving range prediction, issues such as poor quality of real vehicle data, insufficient model diversity, and a lack of practical validation hinder the application of data-driven methods in the cloud. First, data-driven methods rely heavily on extensive real-world driving data and have high requirements for data quality, which is the foundation for developing high precision SOC prediction models. Second, due to factors like driving conditions, driving style, and ambient temperature affecting the remaining driving range of vehicles during actual road travel, enhancing model diversity helps improve the accuracy and generalization performance of models when facing complex driving conditions and environmental factors. Finally, most studies often lack validation in real road scenarios, inevitably raising doubts about the applicability of these models in actual road environments. To address these challenges, this work primarily focuses on three aspects:

1. Based on the actual road driving data of a certain vehicle manufacturer, considering the influence of driving conditions and driving style on the prediction results of SOC,

a clustering method based on PCA-GA-K-Means for driving conditions and vehicle driving style was constructed, and energy consumption characteristics related to the prediction results of SOC were established.

2. Starting from the construction of model diversity, a Stacking model framework is proposed with RandomForest, CatBoost, XGBoost, and SVR as the base model and ElasticNet as the meta model.
3. The accuracy and generalization performance of the model of electric vehicles in different environmental temperatures, different driving styles, and different driving conditions were verified in real open scenarios.

2. Materials and Methods

2.1. Data Sources

The data used in this study were obtained from the cloud platform server of SAIC-GM-Wuling. The dataset consists of nearly one year of operational data from 100 new energy vehicles, specifically the 333 km range version of the Wuling Bingo model. The detailed vehicle specifications are shown in Table 1. All personally identifiable information in the dataset has been anonymized through rigorous desensitization procedures. The data collection followed China's national standard GB/T 32960 with a sampling frequency of 0.5 Hz. As shown in Figure 1, the data were collected by onboard terminals that acquire and store key status parameters of the vehicle and its components and were then transmitted to the cloud via the onboard T-BOX. The collected data includes vehicle status, speed, total voltage, etc. Detailed data descriptions are provided in Table 2.

Table 1. Vehicle specifications.

Component	Parameters	Value
Battery	Type Capacity	Lithium iron phosphate 31.9 kWh
Motor	Peak power Max torque	50 Kw 125 Nm
Vehicle	Weight Drag coefficient Front area	1155 kg 0.32 2.338 m^2

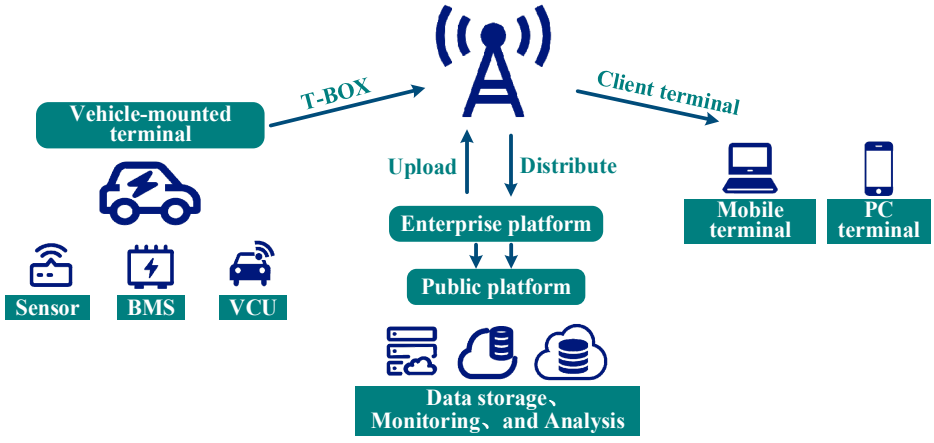


Figure 1. Schematic diagram of the data acquisition process.

Table 2. Data field descriptions.

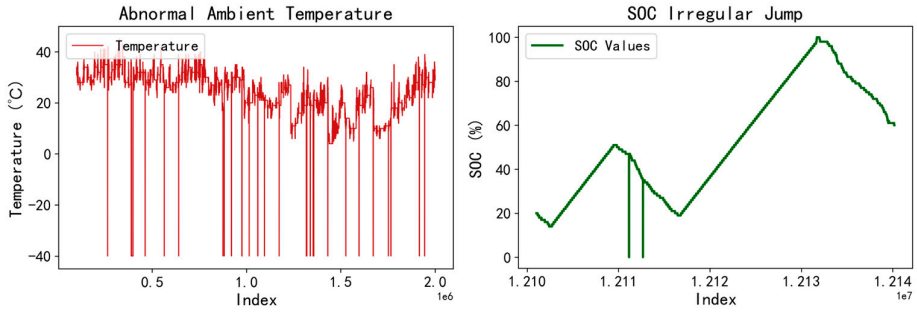
Order Number	Field Categorization	Field Name	Unit
1	Battery data field	Maximum voltage of battery cell	V
2		Minimum voltage of battery cell	V
3		Maximum temperature value	°C
4		Minimum temperature value	°C
5		Charged state	
6		SOC	%
7		Total current	A
8		Total voltage	V
9	Traffic data fields	Vehicle identification number	
10		Data collection time	
11		Speed of a motor vehicle	km/h
12		Accelerate the pedal stroke value	%
13		Vehicle status	
14		Brake pedal status	
15		Drive motor speed	r/min
16		Current motor torque	N·m
17		Drive motor controller temperature	°C
18		Drive motor temperature	°C
19		Cumulative mileage	km
20		The steering wheel angle	°
21	Main energy consuming components and environmental data	Air conditioning temperature	°C
22		Air conditioning heater relay control command	
23		Air conditioning compressor switch signal	
24		Ambient temperature	°C
25		Environmental pressure values	kPa

2.2. Outlier Treatment

During data acquisition, sensor signal transmission anomalies may lead to data loss and erroneous data, typically manifested as abnormal jumps, missing values, and data latency. These anomalies can adversely affect subsequent model training processes, necessitating preliminary data cleaning and correction procedures.

Analysis revealed that data anomalies primarily involved SOC missing values, abnormal jumps in ambient temperature, as well as latency in vehicle state transition boundary data, as illustrated in Figure 2. For time-lagged data, rule-based correction was directly applied. Regarding missing values and abnormal jumps given their substantial volume, outliers were first converted to null values before interpolation. Three interpolation methods were comparatively evaluated: Linear Interpolation, Nearest Neighbor Interpolation, and Lagrange Interpolation. The methodology proceeded as follows: 20,000 random data segments were selected and artificially nullified. The RMSRE (Equation (1)) metric was employed to evaluate imputation performance, with results detailed in Table 3. Based on comparative analysis, linear interpolation was ultimately adopted for SOC, accumulated mileage, and maximum and minimum temperature values, while Lagrange interpolation was applied to ambient temperature and total voltage data.

$$\mathrm{RMSRE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \quad (1)$$



(**a**) (**b**)

Figure 2. Data anomaly visualization: (**a**) ambient temperature anomaly diagram; (**b**) SOC anomaly diagram.

Table 3. Missing value imputation results.

Interpolation Variables	RMSRE Grade		
	Linear Interpolation	Nearest Neighbor Interpolation	Langevin Difference
SOC	0.0056	0.0069	0.0109
Cumulative mileage	0.0048	0.0076	0.0117
Ambient temperature	0.0097	0.0112	0.0086
Maximum temperature value	0.0082	0.0087	0.0102
Minimum temperature value	0.0125	0.0152	0.0197
total voltage	0.0089	0.0122	0.0079

2.3. Build the Remaining Driving Range Field

The original dataset only contained the accumulated mileage field, without a remaining range field. To facilitate subsequent model development and evaluation, it is necessary to construct the remaining range field here. Considering the complete SOC discharge interval of [100%, 0%], the remaining range field across the full interval can be obtained by calculating the difference between the accumulated mileage value at SOC = 0% and the accumulated mileage values at other SOC states during each discharge cycle, as shown in Table 4:

Table 4. Construction of remaining driving range field.

Label	Data Collection Time	Speed	...	SOC	Cumulative Mileage	Remaining Driving Range
25	2024-08-19 06:37:26	11.1	...	100	5518.8	308.1
25	2024-08-19 06:37:28	15.3	...	100	5518.9	308.0
25	2024-08-19 06:37:30	22.3	...	100	5519.0	307.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	2024-08-19 10:47:33	3.1	...	0	5734.5	0.1
25	2024-08-19 10:47:35	2.1	...	0	5734.6	0

2.4. Analysis of Influencing Factors

(1) The relationship between SOC and remaining driving range

SOC stands for the battery state of charge, which is a crucial indicator reflecting the remaining available capacity of the battery. It is also a key factor influencing the remaining driving range. Drivers typically use the SOC value to estimate the remaining driving range. The SOC reflects the current state of charge of the battery and has a positive correlation

with the remaining driving range. The higher the SOC value, the greater the remaining driving range, as shown in Figure 3a.

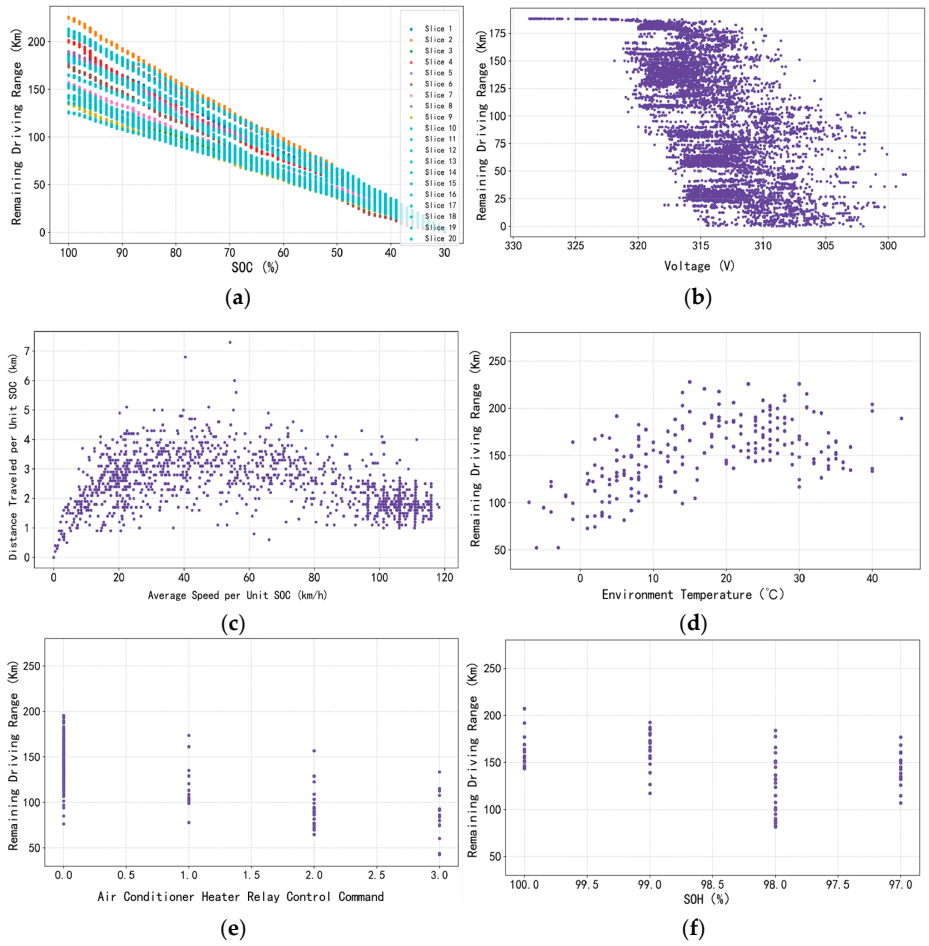


Figure 3. Influence factors and remaining driving range relationship diagram: (**a**) schematic diagram of the relationship between SOC and remaining driving range; (**b**) voltage and remaining driving range; (**c**) speed versus driving range per unit SOC; (**d**) ambient temperature versus remaining driving range; (**e**) relationship diagram between HVAC commands and remaining driving range; (**f**) correlation diagram between SOH and remaining driving range.

(2) The relationship between voltage and remaining driving range

The relationship between voltage and remaining driving range shows a certain positive correlation. Specifically, during vehicle operation, the size of the remaining driving range changes with variations in voltage. When the vehicle is fully charged, the SOC value is high, and so is the voltage; as the driving distance increases, the total voltage gradually decreases, as shown in Figure 3b.

(3) The relationship between vehicle speed and unit SOC mileage

When the vehicle speed is low, frequent acceleration and deceleration occur, leading to lower range. When the vehicle speed is too high, increased wind resistance also affects the driving range [16]. The graph showing the relationship between unit SOC mileage and vehicle speed indicates that the unit SOC mileage is higher when the vehicle speed is between 30 and 60 km/h, indicating better economic performance in this range, as shown in Figure 3c.

(4) The relationship between temperature and remaining driving range

Due to the heat released by the battery during operation, as the electric vehicle runs longer, it will lead to a gradual increase in battery temperature [17]. As shown in Figure 3d,

there is a strong correlation between battery temperature and remaining driving range in a single driving segment. Moreover, the remaining available capacity of the battery is affected by ambient temperature; it tends to be lower in cold environments, thus directly impacting the vehicle's driving range.

(5) The relationship between the main energy consuming parts and the remaining driving range

For pure electric vehicles, in addition to the energy consumption generated during driving, the energy consumption of key vehicle components is also a significant factor affecting the remaining range. For example, the energy consumption from the air conditioning and thermal management system, on-board entertainment and information systems, auxiliary driving and safety systems, lighting and signaling systems, as well as features like electric seats, electric tailgates, and wireless charging devices, is more severe compared to other components. This issue is particularly pronounced in cold winters and hot summers, as shown in Figure 3e.

(6) The relationship between SOH and the remaining driving range

The SOH of a pure electric vehicle's battery decreases as the number of vehicle cycles and the duration of use increase. SOH is a critical indicator reflecting the current battery capacity; the lower the SOH, the deeper the battery degradation. As shown in Figure 3f, although the remaining driving range of the current vehicle is influenced by factors such as ambient temperature, driving conditions, and driving style, it can still be observed from the graph that the remaining driving range decreases as the battery SOH declines.

In addition to the above factors, driving conditions and driving style will also affect the remaining driving range. The relationship between driving conditions, driving style, and remaining driving range will be analyzed below.

(1) The relationship between driving conditions and remaining driving range

When a vehicle is in motion, it encounters various road conditions, such as urban congestion, smooth traffic in the city, elevated highways, and expressways. These driving conditions affect the energy consumption of electric vehicles, directly impacting their remaining range, as shown in Figure 4. Therefore, considering the impact of driving conditions on the remaining range is of great significance.

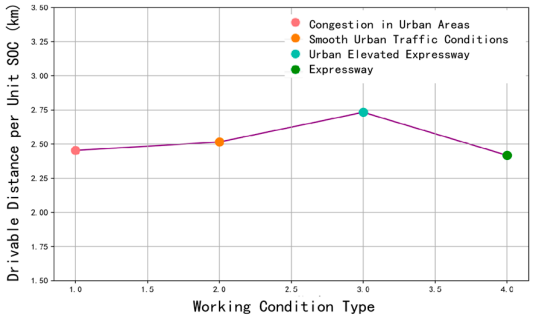


Figure 4. Relationship between working conditions and remaining driving range.

(2) The relationship between driving style and remaining driving range

By constructing a clustering model, the short-term driving styles of drivers are categorized into three types: gentle, calm, and aggressive. As shown in Figure 5, even under complex conditions where the vehicle is not traveling at a constant speed, differences in energy consumption performance can still be identified among different driving styles. Among these, the energy consumption performance of gentle drivers is better than that of calm drivers, while aggressive drivers have the poorest performance. The specific reason is that aggressive drivers frequently accelerate and decelerate, which can easily lead to energy loss.

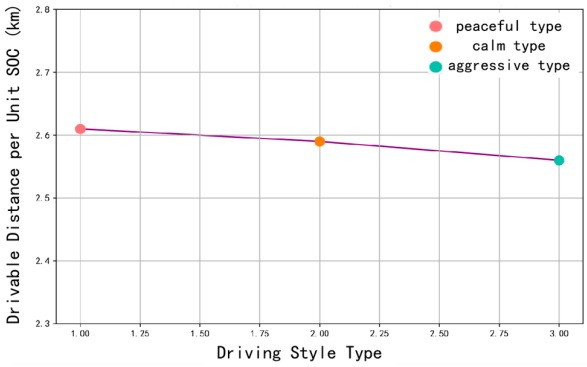


Figure 5. Relationship between driving style and remaining driving range.

3. Feature Analysis and Selection

3.1. Driving Condition Identification

The operating condition characteristic parameters, as key features for identifying driving conditions, need to accurately represent the status of the vehicle in this segment to obtain precise condition recognition results. These data generally include maximum speed, average speed, speed distribution, acceleration, acceleration time ratio, deceleration time ratio, and stop time ratio. Based on the acquired vehicle data, the following 12 characteristic parameters have been determined, as shown in Table 5.

Table 5. Characteristics of driving conditions.

Order Number	Parameter Symbols	Characteristic Parameter	Order Number	Parameter Symbols	Characteristic Parameter
1	a_{max}	maximal acceleration	7	$V_{0–20}$	Speed range 0–20 time ratio
2	b_{max}	maximum deceleration	8	$V_{20–40}$	Speed range 20–40 time ratio
3	T_a	Acceleration time ratio	9	$V_{40–60}$	Speed range 40–60 time ratio
4	T_d	Time reduction ratio	10	$V_{60–80}$	Speed range 60–80 time ratio
5	T_m	Uniform time ratio	11	$V_{80–120}$	Speed range 80–120 time ratio
6	T_p	Parking time ratio	12	a_x	Accelerate the pedal stroke value

Considering the large number of features and significant differences in indicator values, PCA is introduced to reduce the dimensionality of feature data. The PCA method first standardizes the data to minimize the impact of inconsistent scales among features. Based on the magnitude of eigenvalues, the feature vectors corresponding to the top few largest eigenvalues are selected to obtain the principal components. When the cumulative contribution rate of individual principal components reaches over 85%, it indicates that these features can reflect most of the characteristics of the original data [18]. The formula for calculating the contribution rate of principal components is as follows:

$$\omega_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \tag{2}$$

w_i is the contribution rate of the i-th principal component, λ_i ($j = 1, 2, \ldots, n$)represents the eigenvalue of the i-th principal component, and $\sum_{j=1}^{n} \lambda_j$ is the sum of all principal component eigenvalues.

The interpretation variance ratio and cumulative interpretation variance ratio of the operating condition are shown in Figure 6. When the number of principal components is 4, the cumulative contribution rate reaches 86.33%, exceeding 85% for the first time, so the number of principal components is selected as 4.

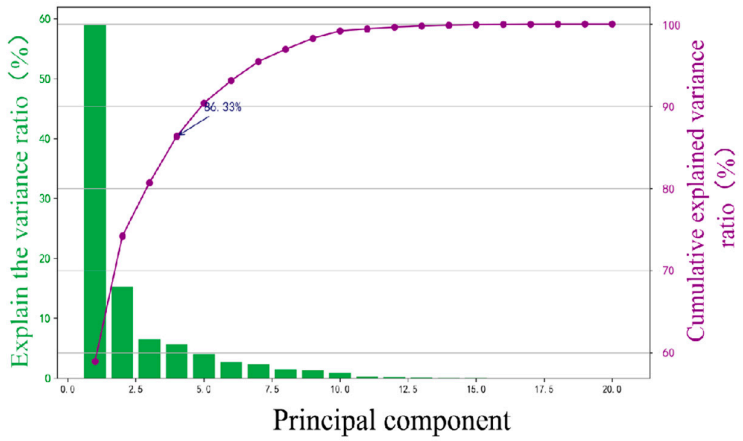
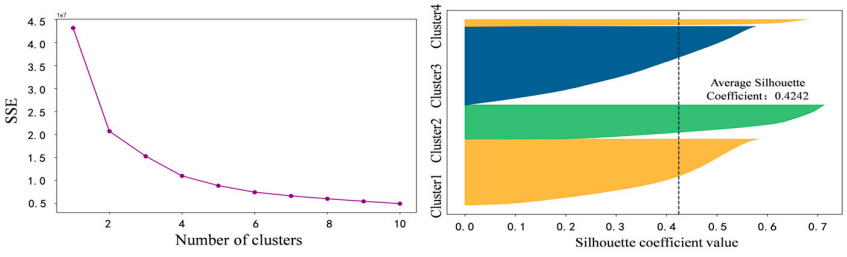


Figure 6. Cumulative explained variance ratio diagram.

Traditional K-Means algorithms are sensitive to initial cluster centers, and the initial values significantly impact the results. To enhance the robustness of clustering algorithms, a GA-optimized K-Means method for driving condition clustering has been proposed. The GA searches for the globally optimal initial cluster centers to achieve more accurate clustering results. The optimal number of clusters in the clustering algorithm is selected using a combination of the elbow method and the silhouette coefficient method. In the elbow method, the optimal number of clusters is where the curve flattens. In the silhouette coefficient method, the silhouette coefficient is an indicator of clustering performance, ranging from -1 to 1; the closer it is to 1, the better the clustering effect. Based on the results from the elbow method and the silhouette coefficient shown in Figure 7, the optimal number of clusters for this work is set to 4.



(**a**) (**b**)

Figure 7. Determination of the optimal number of clusters: (**a**) elbow method for determining the optimal number of clusters; (**b**) outline coefficient diagram.

The GA algorithm mainly obtains the optimal solution through crossover and mutation. In this paper, the optimal solution is the optimal clustering center, which is then input into the K-Means algorithm, as shown in Figure 8.

The K-Means clustering results optimized by the GA have a better profile coefficient (Silhouette Score) and lower error sum of squares, indicating that the initial centers optimized by the GA can better represent the real structure of data, thus improving the clustering effect. The K-Means clustering results optimized by the GA are shown in Figure 9, a 3D scatter plot illustrating the clustering of data points across three principal components. The plot uses distinct colors to represent four clusters, with the color gradient

indicating cluster assignments along the cluster index. This visualization highlights the separation of driving cycles, validating the PCA-GA-K-Means approach's ability to identify meaningful patterns.

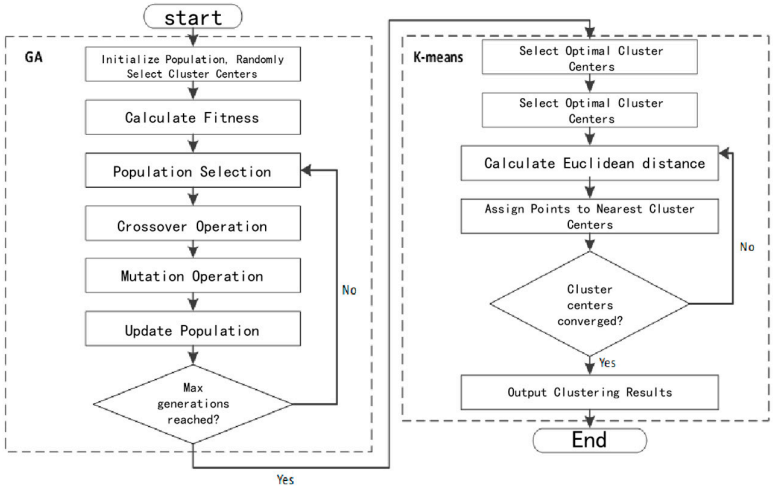


Figure 8. Flowchart of GA-K-Means algorithm.

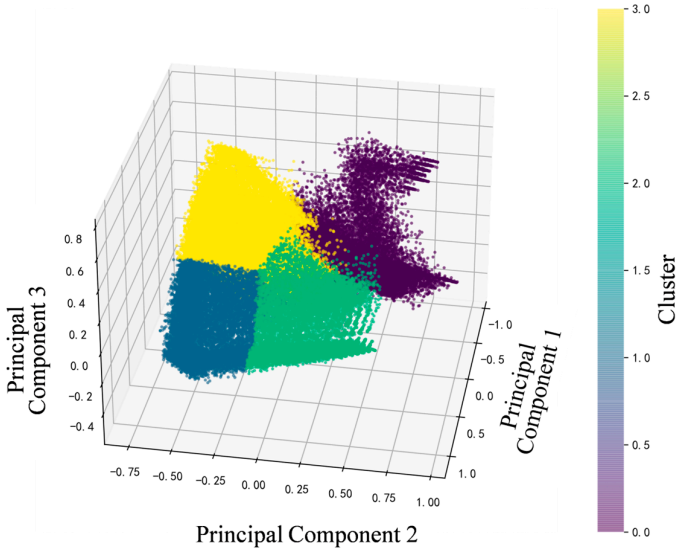


Figure 9. Schematic diagram of driving condition clustering results.

According to the obtained clustering results, the cluster centers can be identified. Since the number of clusters in this paper is 4, the obtained cluster centers are 4 in total. The K-Means identification is determined by calculating the distance between the input samples and the cluster centers and then classifying them into clusters based on the minimum distance, which serves as the clustering label. The formula for calculating the minimum distance is as follows:

$$\begin{cases} D_i = \sqrt{(data_{current} - C_i)^2}, j = 1, 2, 3 \ldots k \\ [\backsim, label] = min(D) \end{cases} \quad (3)$$

$data_{current}$ represents the input sample, C_i denotes the driving cycle cluster centroid, and $label$ is the identified label. After obtaining the label, the clustering results of driving conditions are obtained, as shown in Figure 10 for the representative segments and recognition results of four kinds of driving conditions.

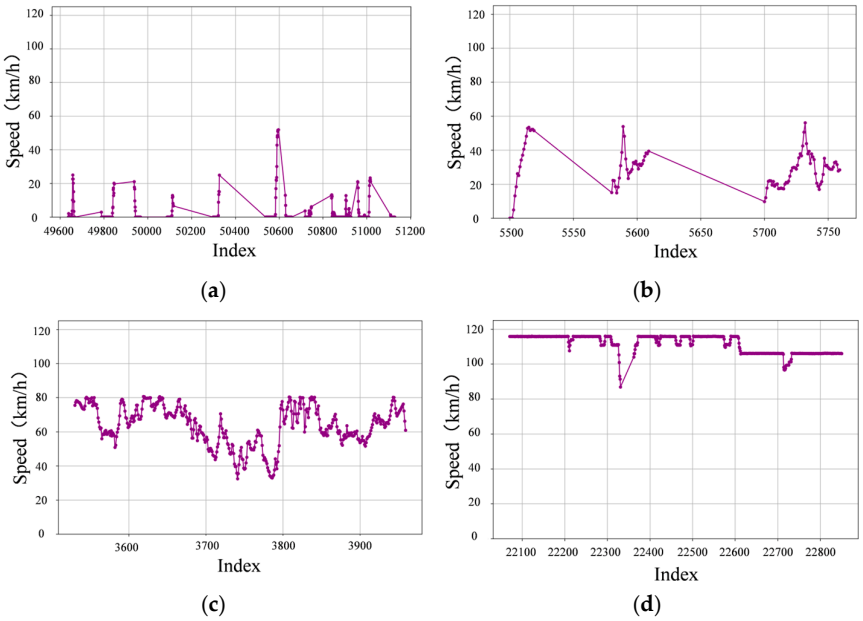


Figure 10. Schematic diagram of clustering results of four typical working conditions: (**a**) condition type 1: urban congestion condition; (**b**) condition type 2: urban smooth condition; (**c**) condition type 3: urban elevated expressway; (**d**) condition type 4: expressway.

3.2. Construction of Vehicle Driving Style Features Based on KL-GMM

During the process of vehicle driving, driving style is influenced by personal character on the one hand and driving conditions on the other. At present, a large number of scholars have divided driving style into three categories: peaceful type, calm type, and aggressive type [16]. In order to identify driving styles more accurately, the following driving style characteristic parameters are constructed, including vehicle acceleration, acceleration rate of change, etc. As shown in Table 6. The formula is as follows:

$$f = (a_std, J \ldots SW_a_mean, \mathrm{t}) \tag{4}$$

Table 6. Driving style characteristic parameters table.

Order Number	Parameter Symbols	Parameter Name	Order Number	Parameter Symbols	Parameter Name
1	*a_std*	Acceleration standard deviation	5	*SW_a_max*	Maximum steering wheel corner acceleration
2	*AP_std*	Accelerate pedal stroke standard deviation	6	*AP_a_std*	Accelerator pedal acceleration standard deviation
3	*J_std*	Standard deviation of acceleration change rate	7	*AP_a_mean*	Acceleration pedal average acceleration
4	*SW_a_max*	Maximum steering wheel corner acceleration	8	*SW_a*	Angular acceleration of steering wheel

During the actual clustering process, features in the dataset may be irrelevant to or redundant with the clustering results. The MIC can uncover significant relationships between features, effectively capturing nonlinear relationships between variables, which helps understand which features contribute the most to clustering. To obtain the few features most relevant to driving style labels, the MIC method is used to mark the MIC scores between each feature and the driving style label, as shown in Figure 11.

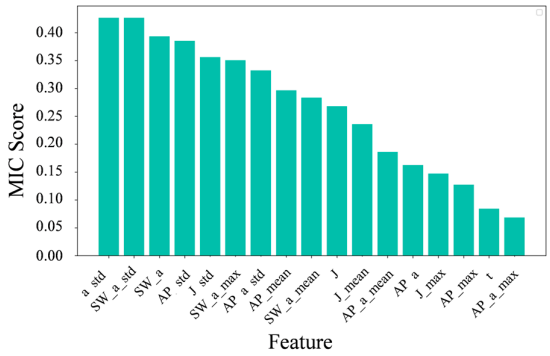


Figure 11. Features and MIC scores for driving style.

Before GMM clustering, it is necessary to determine the number of clusters first. Generally, the commonly used methods include BIC and AIC, which are used as statistical quantities for model selection and evaluate the model by punishing overfitting.

$$AIC = 2 * k - 2 * LL \tag{5}$$

$$BIC = log_e(N) * k - 2 * LL \tag{6}$$

k is the number of parameters in the model, N is the sample size, and the number of AIC/BIC is smaller, the model fit will be better. As shown in Figure 12, scores are plotted on the graph; the more clusters there are, the lower the score, which can easily lead to overfitting. Therefore, to avoid overfitting, we take the derivative of each point and find that when the number of cluster groups exceeds three, the slope begins to change slowly. Hence, the clustering number for the GMM model is set to three.

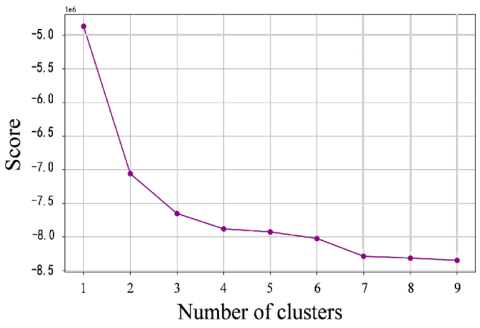


Figure 12. AIC/BIC Scoring.

The GMM assumes that the data comes from multiple Gaussian distributions, with each cluster modeled by a single Gaussian distribution [19]. Therefore, the clustering effect of the GMM typically depends on the shape, size, and position of the Gaussian distributions. When there is some overlap between clusters in the data, the GMM may result in over clustering or incorrect clustering, especially when the sizes and shapes of the clusters differ significantly. Figure 13 shows a schematic diagram of the driving style clustering results using a GMM. The plot displays three distinct clusters (yellow, purple, and green), reflecting different driving styles derived from behavioral data. The separation and overlap of these clusters indicate the model's ability to capture diverse driving patterns, though some overlap suggests potential challenges in distinguishing closely related styles.

To quantify the relationship between driving conditions, driving style, and remaining range, a driving condition–driving style coupling database was established. Driving styles were clustered based on each driving condition, resulting in twelve combinations of driving

conditions and driving styles. Table 7 presents the established driving condition–driving style coupling database along with their probability ratios.

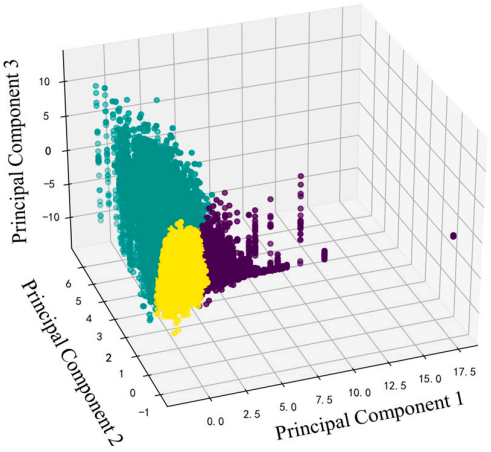


Figure 13. Schematic diagram of driving style clustering results. Different colors represent different clustering results.

Table 7. Driving conditions–driving style coupling library.

Order Number	Operating Conditions–Driving Style	Proportion (%)	Order Number	Operating Conditions–Driving Style	Proportion (%)
1	Urban congestion-peaceful type	7.77	7	Urban elevated-peaceful type	12.96
2	Urban congestion-calm type	10.73	8	Urban elevated-cool type	13.44
3	Urban congestion-aggressive	2.54	9	Urban elevated-radical type	6.17
4	Smooth and peaceful urban traffic	12.58	10	High speed-peaceful type	2.07
5	Smooth downtown-calm type	30.32	11	High speed-calm type	1.19
6	Smooth downtown-aggressive	0.02	12	High speed-aggressive type	0.21

3.3. Vehicle Energy Consumption Characteristics Considering Driving Style and Driving Conditions

The driving conditions and driving style of vehicles are influenced by various factors, such as future road environments, traffic volume, and uncertain factors that affect the driver's driving style. Therefore, these factors significantly impact the prediction of future vehicle driving conditions and driving styles. Currently, some scholars use API interfaces to access navigation software to obtain future road information for predicting future vehicle speeds [20], thereby obtaining future driving conditions and driving styles. Other scholars directly predict vehicle speeds using methods $P_1, P_2, P_3, \ldots, P_{12}$ like Markov chains and neural networks. Since this study focuses on the construction and prediction of range models, it is assumed that the distribution probabilities of future driving conditions and driving styles have been obtained through the aforementioned methods. Therefore, the energy consumption per unit distance for the current vehicle can be derived,

$$\begin{aligned} E = \sum_i^{12} p_i * V = {} & p_1 * 16.1 + p_2 * 18.7 + p_3 * 21.7 + p_4 * 12.43 \\ & + p_5 * 12.74 + p_6 * 13.21 + p_7 * 9.54 + p_8 * 11.34 + p_9 * 13.77 \\ & + p_{10} * 15.05 + p_{11} * 15.59 + p_{12} * 15.81 \end{aligned} \quad (7)$$

Among them, E is the current energy consumption per unit distance of the vehicle; $P_1, \ldots, P_{12}$ represent the probabilities of occurrence for driving cycles and driving styles over a future period; V_i denotes the energy consumption per unit distance for each driving cycle driving style combination.

For the total energy consumption E_{total} of a pure electric vehicle, calculate the total energy output E_{ex} from the battery and the energy recovery E_{in} obtained through regenerative braking. The calculation formulas are as follows:

$$E_{ex} = \sum_{i=1}^{N-1} U_i I_i (t_{i+1} - t_i) \tag{8}$$

$$E_{in} = \sum_{i=1}^{N-1} U_i I_i (t_{i+1} - t_i) \times \eta_{es} \times \eta_{md} \times \eta_{td} \times \xi \tag{9}$$

$$E_{total} = E_{ex} + E_{in} \tag{10}$$

Among them, U_i is the total voltage, I_i is the total current, η_{es} is the average energy utilization rate, η_{md} is the motor efficiency, η_{td} is the transmission system efficiency, and ξ is the braking energy contribution rate.

Both the energy consumption prediction model based on driving conditions and driving styles and the remaining range prediction model based on historical average energy consumption belong to energy consumption-based methods for predicting remaining range. The calculation process is similar, involving division of the current remaining battery energy by the energy consumption per unit distance.

The energy consumption per unit mileage of each driving condition–driving style cluster can be calculated from the historical driving data of vehicles, as shown in Table 8.

Table 8. Energy consumption per unit mileage of historical driving data.

Order Number	Operating Conditions–Driving Style	Energy Consumption (kW·h/100 km)	Order Number	Operating Conditions–Driving Style	Energy Consumption (kW·h/100 km)
1	Urban congestion-peaceful type	16.1	7	Urban elevated-peaceful type	9.54
2	Urban congestion-calm type	18.7	8	Urban elevated-cool type	11.34
3	Urban congestion-aggressive	21.7	9	Urban elevated-radical type	13.77
4	Smooth and peaceful urban traffic	12.43	10	High speed-peaceful type	15.05
5	Smooth downtown-calm type	12.74	11	High speed-calm type	15.59
6	Smooth downtown-aggressive	13.21	12	High speed-aggressive type	15.81

4. Feature Screening and Model Construction

4.1. Feature Extraction Method Based on DII

The core idea of DII is to optimize feature weights so that the distance relationships in the input feature space can effectively predict the distance relationships in the real feature space [21]. This is achieved through gradient descent optimization of the weights, while simultaneously aligning the units and scaling the importance of features. DII can be understood as differentiable information imbalance, where information imbalance (Δ) quantifies the ability of feature space A to predict the distance from feature space B, as expressed by the following formula:

$$\Delta(d^A \to d^B) = \frac{2}{N^2} \sum_{i,j:r_{ij}^A=1} r_{ij}^B \tag{11}$$

Among them r_{ij}^A and r_{ij}^B, respectively, represent the ranking of the distance between data point j and i in feature space A and B. Δ close to 0 means that A can perfectly predict B, while close to 1 means that there is no prediction ability.

DII supports gradient optimization based on information inequality (Δ), through which Δ is extended to continuously differentiable DII. The formula is as follows:

$$DII(d^A(\omega) \to d^B) = \frac{2}{N^2}\sum_{i \neq j} c_{ij}(\lambda, d^A(\omega)) \cdot r_{ij}^B \tag{12}$$

Among them, c_{ij} is the soft maximum coefficient, and the nearest neighbor constraint is approximated by exponential decay weight:

$$c_{ij} = \frac{e^{-d_{ij}^A(\omega)/\lambda}}{\sum_{m \neq i} e^{-d_{im}^A(\omega)/\lambda}} \tag{13}$$

Among them, when λ approaches 0, DII degenerates to Δ.

Introducing L1 regularization can automatically select the optimal number of features. Therefore, when facing the complex nonlinear dependency between target variables and features in remaining driving range prediction models, it can efficiently complete feature extraction. By adjusting the size of L1 regularization, the learning weight for feature selection can be adjusted; an increase in the regularization coefficient leads to more feature weights being set to zero, so the number of features in the following figure can be understood as the number of non-zero features. As shown in Table 9, which presents the optimization of regularization weights and determination of feature numbers, during the regularization adjustment process, the DII value also changes. An increase in the DII value indicates a decrease in feature similarity, suggesting the loss of feature information. As illustrated in Figure 14, which shows the relationship between feature numbers and DII, when the number of features is 8, the DII value is minimized, indicating the strongest similarity among features. When the number of features decreases, the DII value increases sharply, also representing the loss of key information.

Table 9. Regularization weight optimization and feature number determination table.

L1 Regularization	Number of Features	DII Price
not have	20	0.007
0.0001	12	0.005
0.0002	8	0.003
0.0012	3	0.009
0.0029	2	0.023
0.0023	1	0.079

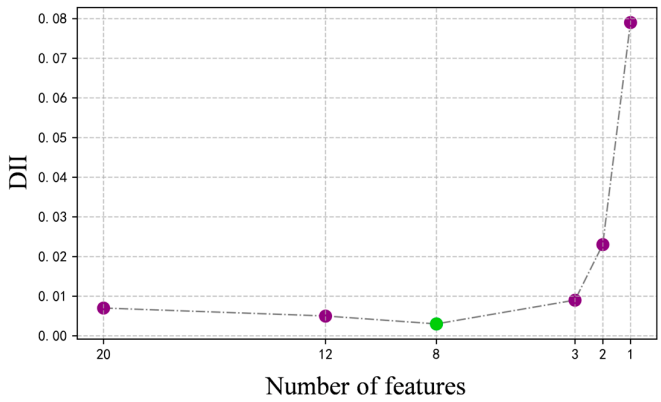


Figure 14. Relationship between feature number and DII. The green points represent the results corresponding to the number of features we selected.

Figure 15 is the feature importance ranking diagram, and these features are also input into the subsequent Stacking model. It can be seen from the figure that the key factors affecting the target variable include SOC, temperature, air conditioning energy consumption, driving conditions–driving style, SOH, etc.

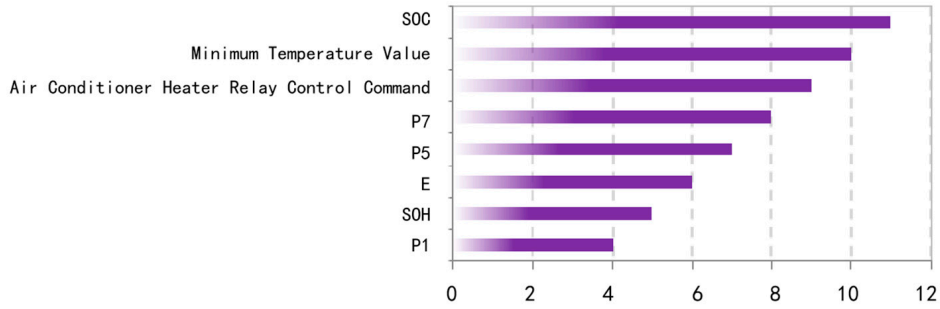


Figure 15. Feature importance ranking diagram.

4.2. Construction of Remaining Driving Range Model Based on Stacking Algorithm

Stacking enhances the overall performance of models by combining predictions from multiple models. Stacking typically consists of two layers: the first layer is the base model, which generates predictions through training multiple models. The diversity of base models is key to the success of stacked algorithms. Different types of base models have distinct assumptions and learning capabilities, allowing them to capture patterns in data from various perspectives. Figure 16 illustrates the principle of the Stacking algorithm.

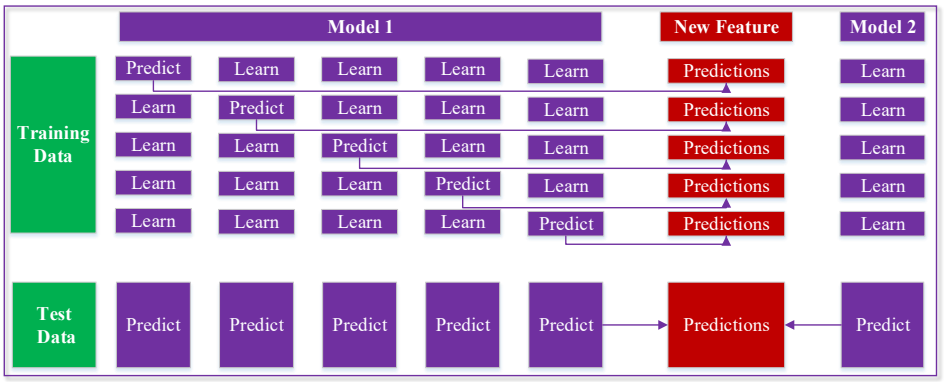


Figure 16. Stacking algorithm schematic.

In the Stacking algorithm, the selection of base models and meta model significantly impacts the model's generalization, robustness, prediction accuracy, and computational complexity. Base models should ideally maintain diversity, incorporating models with significant differences to enable them to complement each other in capturing data features, thereby enhancing overall predictive performance. The meta model should be kept as simple and robust as possible to avoid introducing complex nonlinear mappings on top of the base models' outputs, thus reducing the risk of overfitting.

The model results of Ridge, Lasso, ElasticNet, SVR, LightBGM, XGBoost, CatBoost, and RandomForest were compared individually, using five-fold cross validation. The RMSRE values of the prediction results for each model are shown in Figure 17. Based on the selection criteria for base models and meta models, RandomForest, CatBoost, XGBoost, and SVR, which have significant differences, were chosen as base models, while ElasticNet was selected as the meta model.

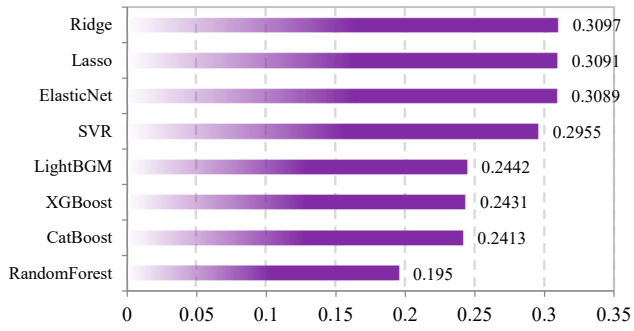


Figure 17. Single-model prediction results: RMSRE score.

The final Stacking model framework, as shown in Figure 18, consists of two layers: the base model and the meta model. First, the dataset is divided into a training set and a test set. The training set is then input into multiple base models in the first layer, and the results from each base model are obtained through five-fold cross validation. Next, the prediction results from each base model are used as new features and fed into the meta model to obtain the final prediction result.

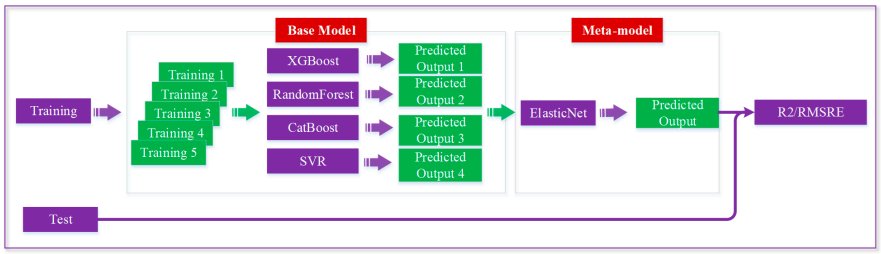


Figure 18. Stacking model framework.

5. Test Verification

R^2 and RMSRE were used as evaluation indexes to verify the prediction results of the single model and Stacking fusion model. The calculation formulas of R^2 and RMSRE are shown in (13) and (14). The comparison results of the accuracy of each model after parameter correction are shown in Table 10; the Stacking model achieved the best performance, with an R^2 of 0.9410 and an RMSRE of 0.0999. Compared to the best performing single model (CatBoost, RMSRE = 0.1124), the Stacking model reduced the prediction error by 11.1%, demonstrating the effectiveness of ensemble learning. Among the base models, SVR performed the worst (R^2 = 0.6391, RMSRE = 0.2820), which may be due to its limited ability to capture nonlinear feature interactions in complex scenarios. CatBoost and XGBoost outperformed Random Forest, possibly because gradient boosting methods can better fit subtle variations in the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y_i})^2} \tag{14}$$

$$\text{RMSRE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2} \tag{15}$$

In order to verify the improvement effect of the Stacking-based data-driven model compared with traditional model method in different ambient temperature and different driving conditions, the accurate comparison of the whole temperature region and the whole working condition were carried out, respectively.

Table 10. Comparison results of accuracy of each model.

	R^2	RMSRE
XGBoost	0.9137	0.1279
CatBoost	0.9233	0.1124
RandomForest	0.8990	0.1571
SVR	0.6391	0.2820
Stacking	0.9410	0.0999

To verify the robustness of the data-driven remaining driving range prediction model under the influence of environmental temperature and driving conditions, we first focus on environmental temperature. We select as much data as possible from all temperature ranges, choosing eight real vehicle data points with environmental temperatures in the ranges [27~39 °C], [23~33 °C], [15~25 °C], [6~13 °C], [−4~3 °C], [−10~2 °C], [−13~1 °C], and [−21~−3 °C] for analysis and validation, as shown in Figure 19. By comparing the predicted results with the actual values, we confirm the model's robustness across different environmental temperatures. The data shows that the data-driven model performs well across the entire temperature range, especially in low temperature environments, where its prediction accuracy significantly improves compared to the energy consumption model.

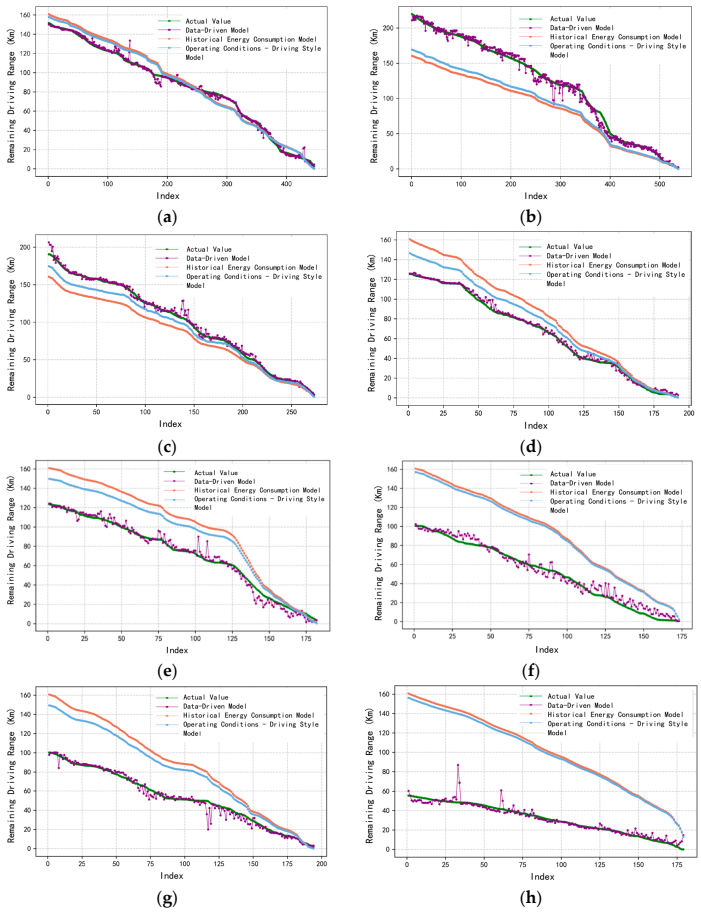


Figure 19. Comparison of model prediction results at [−2~39 °C] ambient temperature (**a**) ambient temperature range [27~39 °C]; (**b**) ambient temperature range [23~33 °C]; (**c**) ambient temperature range [15~25 °C]; (**d**) ambient temperature range [6~13 °C]; (**e**) ambient temperature range [−4~3 °C]; (**f**) ambient temperature range [−10~2 °C]; (**g**) ambient temperature range [−13~1 °C]; (**h**) ambient temperature range [−21~−3 °C].

To further evaluate the generalization ability of the model under complex driving conditions, this study selected four typical driving scenarios: urban congestion, urban smooth traffic, urban elevated roads, and expressways. For each condition, the predictions of three different models were compared with the actual values, as shown in Figure 20. It is evident that under all conditions, the predictions of the data-driven model are consistently the closest to the actual values.

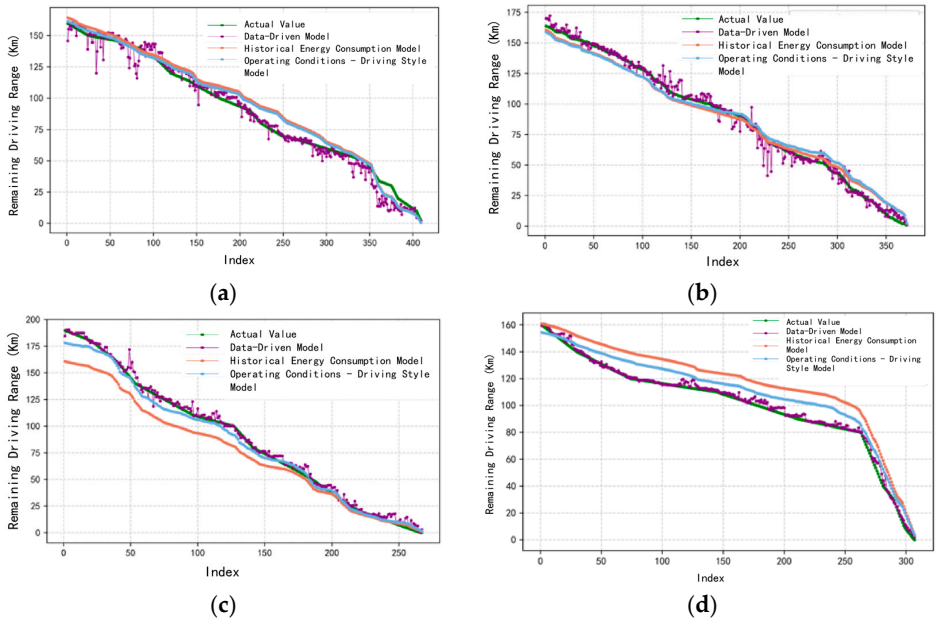


Figure 20. Comparison of prediction results of four typical driving conditions models: (**a**) congestion in urban areas; (**b**) smooth urban traffic conditions; (**c**) urban elevated expressway; (**d**) expressway.

The prediction errors of each model under the four typical driving conditions are presented in Figure 21. According to the results, the data-driven model achieves higher prediction accuracy than both the historical energy consumption model and the operating conditions–driving style model across all conditions, demonstrating strong generalization capability. Moreover, by controlling the ambient temperature within the range of [15~25 °C], the influence of air conditioning energy consumption on prediction results is eliminated, further validating the reliability of the model under typical driving conditions.

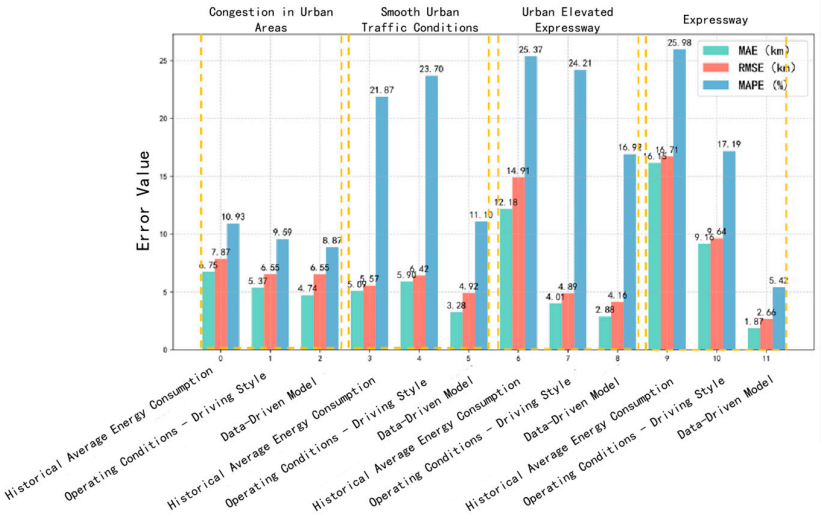


Figure 21. Prediction error results of each model under four typical driving conditions.

To further illustrate the overall prediction performance differences between the three models, a boxplot comparison of the prediction errors is provided, as shown in Figure 22. The boxplot effectively demonstrates the distribution, dispersion, and presence of outliers in the prediction errors of the data-driven model, the historical energy consumption model, and the operating conditions–driving style model. It can be observed that the data-driven model exhibits the lowest median error and the narrowest IQR among the three, indicating more accurate and stable predictions. In contrast, the historical energy consumption model and the operating conditions–driving style model show higher median errors and wider distributions, with more outliers, suggesting greater variability and less robustness. This statistical comparison further confirms the superior predictive performance and generalization capability of the Stacking-based data-driven approach under both varying environmental conditions and complex driving scenarios.

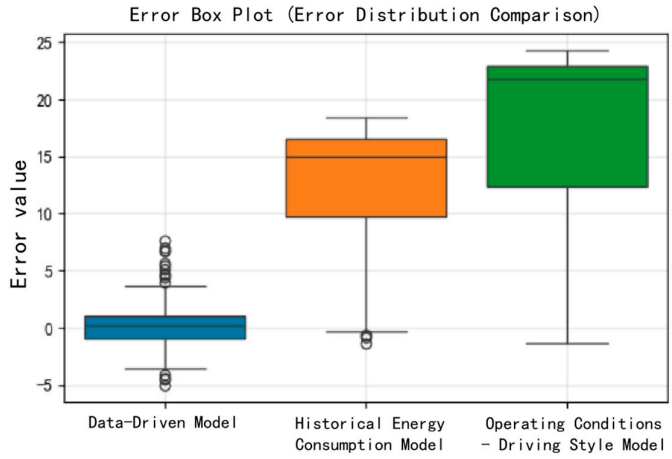


Figure 22. Error boxplot of the three models.

6. Conclusions

This study aims to enhance the prediction accuracy and robustness of EV remaining range by developing a data-driven model architecture guided by an energy consumption model, comprehensively considering factors such as battery aging, driving conditions, driving styles, ambient temperature, and air conditioning energy consumption, achieving high-precision range prediction. A PCA-GA-K-Means-based driving condition clustering method is proposed, optimizing K-Means initial cluster centers with GA to categorize 212,341 driving segments into four conditions: urban congestion, urban smooth traffic, urban elevated expressways, and expressways. A GMM-based driving style feature extraction method is introduced, constructing a coupled library of twelve typical driving condition–style combinations, revealing their complex correlations with range. A Pearson-DII feature extraction approach is proposed to overcome limitations of traditional methods in handling multicollinearity and nonlinear redundant feature relationships, reducing RMSRE by 18.1% compared to conventional methods. A Stacking ensemble model, with Random Forest, CatBoost, XGBoost, and SVR as base models and ElasticNet as the meta model, is developed, decreasing RMSRE by 11.1% compared to single models. Validated with real-world data from eight vehicles across −21 °C to 39 °C and four typical driving conditions, the Stacking model's RMSE is reduced by approximately 85.34% and 33.47% compared to the driving condition–style model, demonstrating superior prediction accuracy and robustness. However, this study still has the following limitations: the current dataset does not include high-temperature samples above 39 °C, so the model's generalization capability under extreme heat conditions has yet to be evaluated; the model was developed based on data from a specific vehicle model, and its cross model adaptability requires further experimen-

tal validation; in addition, since the data were collected from real-world vehicle operations, standard drive cycle data such as BJDST, DST, and US06 were not included. Future research will focus on expanding data collection to cover more extreme environmental conditions, developing a generalized model applicable to multiple vehicle types, and incorporating standard drive cycle tests to enhance the model's adaptability and interpretability. The modeling approach proposed in this study demonstrates strong generalizability and scalability. It offers a novel perspective for the academic community by integrating behavioral features and environmental modeling into SOC prediction, while also providing practical value for the industry by improving the accuracy of EV range display and alleviating range anxiety, thus holding significant theoretical and application potential.

Author Contributions: Original Draft Preparation, M.W.; Methodology, Y.L.; Software, H.W.; Validation, S.Y.; Supervision, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Guangxi Science and Technology Major Project, grant number 2023AA03009.

Data Availability Statement: The original contributions presented in this study are included in this article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: Min Wei was employed by the SAlC, General Wuling Automobile Co., Ltd. We declare that we have no financial or personal relationships with any other individuals or organizations that could improperly influence our work; there are no professional or other personal interests of any nature or kind in any products, services, and/or companies that could be interpreted as influencing the positions presented in this article or the review of this article's title.

Abbreviations

The following abbreviations are used in this manuscript:

SOC	State of Charge
GMM	Gaussian Mixture Model
EV	Electric Vehicles
XGBoost	Extreme Gradient Boosting Regression Tree
LightGBM	Lightweight Gradient Boosting Regression Tree
RNN	Recurrent Neural Networks
GPR	Gaussian Process Regression
EBa	Ensemble Bagging
EBo	Ensemble Boosting
ANN	Artificial Neural Network
SVM	Support Vector Machine
LR	Linear Regression
ML	Machine Learning
BEV	Battery Electric Vehicle
SOH	State of Health
RMSRE	Root Mean Square Relative Error
PCA	Principal Component Analysis
GA	Genetic Algorithm
MIC	Maximum Information Coefficient
BIC	Bayesian Information Criterion
AIC	Akaike Information Criterion
DII	Differentiable Information Inequality
IQR	Interquartile Range

References

1. Tracker, C.A. *Decarbonising the Indian Transport Sector Pathways and Policies*; Climate Action Tracker: Berlin, Germany, 2020.
2. Uddin, W. Mobile and Area Sources of Greenhouse Gases and Abatement Strategies. In *Handbook of Climate Change Mitigation and Adaptation*; Springer International Publishing: Cham, Switzerland, 2022; pp. 743–807.
3. Ghosh, A. Possibilities and Challenges for the Inclusion of the Electric Vehicle (EV) to Reduce the Carbon Footprint in the Transport Sector: A Review. *Energies* **2020**, *13*, 2602. [CrossRef]
4. Li, S.; Jiang, Z.; Zhu, Z.; Jiang, W.; Ma, Y.; Sang, X.; Yang, S. A Framework of Joint SOC and SOH Estimation for Lithium-Ion Batteries: Using BiLSTM as a Battery Model. *J. Power Sources* **2025**, *635*, 236342. [CrossRef]
5. Pevec, D.; Babic, J.; Carvalho, A.; Ghiassi-Farrokhfal, Y.; Ketter, W.; Podobnik, V. Electric Vehicle Range Anxiety: An Obstacle for the Personal Transportation (R)evolution? In Proceedings of the 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 18–21 June 2019; pp. 1–8.
6. Bage, A.N.; Takyi-Aninakwa, P.; Yang, X.; Tu, Q.H. Enhanced Moving-Step Unscented Transformed-Dual Extended Kalman Filter for Accurate SOC Estimation of Lithium-Ion Batteries Considering Temperature Uncertainties. *J. Energy Storage* **2025**, *110*, 115340. [CrossRef]
7. 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. [CrossRef]
8. Yavasoglu, H.A.; Tetik, Y.E.; Gokce, K. Implementation of Machine Learning Based Real Time Range Estimation Method without Destination Knowledge for BEVs. *Energy* **2019**, *172*, 1179–1186. [CrossRef]
9. 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. [CrossRef]
10. Eagon, M.J.; Kindem, D.K.; Panneer Selvam, H.; Northrop, W.H. Neural Network-Based Electric Vehicle Range Prediction for Smart Charging Optimization. *J. Dyn. Syst. Meas. Control* **2022**, *144*, 011110. [CrossRef]
11. 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. [CrossRef]
12. De Cauwer, C.; Verbeke, W.; Coosemans, T.; Faid, S.; Van Mierlo, J. A Data-Driven Method for Energy Consumption Prediction and Energy-Efficient Routing of Electric Vehicles in Real-World Conditions. *Energies* **2017**, *10*, 608. [CrossRef]
13. Bustos, J.E.G.; Baeza, C.; Schiele, B.B.; Rivera, V.; Masserano, B.; Orchard, M.E.; Burgos-Mellado, C.; Perez, A. A Novel Data-Driven Framework for Driving Range Prognostics in Electric Vehicles. *Eng. Appl. Artif. Intell.* **2025**, *142*, 109925. [CrossRef]
14. Jain, A.; Jha, V.; Alsaif, F.; Ashok, B.; Vairavasundaram, I.; Kavitha, C. Machine Learning Framework Using On-Road Real-Time Data for Battery SoC Level Prediction in Electric Two-Wheelers. *J. Energy Storage* **2024**, *97*, 112884. [CrossRef]
15. Sun, S.; Zhang, J.; Bi, J.; Wang, Y. A Machine Learning Method for Predicting Driving Range of Battery Electric Vehicles. *J. Adv. Transp.* **2019**, *2019*, 4109148. [CrossRef]
16. Zhang, C.; Wang, W.; Chen, Z.; Zhang, J.; Sun, L.; Xi, J. Shareable Driving Style Learning and Analysis with a Hierarchical Latent Model. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 11471–11484. [CrossRef]
17. He, K.; Song, Y.; Xie, H.; Cao, F. Winter Driving Range Optimization of Electric Bus Based on CO_2 Thermal Management System and Thermal Energy Cascade Utilization. *Energy* **2025**, *305*, 135668. [CrossRef]
18. Liang, K.; Zhao, Z.; Li, W.; Zhou, J.; Yan, D. Comprehensive Identification of Driving Style Based on Vehicle's Driving Cycle Recognition. *IEEE Trans. Veh. Technol.* **2022**, *72*, 312–326. [CrossRef]
19. Song, D.; Zhu, B.; Zhao, J.; Han, J.; Chen, Z. Personalized Car-Following Control Based on a Hybrid of Reinforcement Learning and Supervised Learning. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 6014–6029. [CrossRef]
20. Morlock, F.; Rolle, B.; Bauer, M.; Sawodny, O. Forecasts of Electric Vehicle Energy Consumption Based on Characteristic Speed Profiles and Real-Time Traffic Data. *IEEE Trans. Veh. Technol.* **2019**, *69*, 1404–1418. [CrossRef]
21. Wild, R.; Wodaczek, F.; Del Tatto, V.; Cheng, B.; Laio, A. Automatic Feature Selection and Weighting in Molecular Systems Using Differentiable Information Imbalance. *Nat. Commun.* **2025**, *16*, 270. [CrossRef] [PubMed]

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.