Energy Consumption Prediction of Electric Bus Based on CNN-XGBoost

Weidong Fang, Yufeng Chi*

School of Electronic, Electrical Engineering and Physics
Fujian Provincial Key Laboratory of Automotive Electronics and Electric Drive
Fujian University of Technology, Fuzhou 350118, China
wdfang@126.com, chiyufeng0103@163.com

Jeng-Shyang Pan

College of Computer Science and Engineering
Shandong University of Science and Technology, Qingdao 266590, China

Chaoyang University of Technology, Taichung 413310, Taiwan
jengshyangpan@gmail.com

*Corresponding author: Yufeng Chi

ABSTRACT. *The existing energy consumption prediction models for pure electric buses do not comprehensively consider various energy consumption influencing factors, and most models only use data from a single operating route. This paper proposes an electric bus trip energy consumption prediction model based on CNN-XGBoost to improve the prediction accuracy. Using the actual driving data of eight pure electric buses, firstly, we analyze the change rule of energy consumption under different ambient temperatures and propose the energy consumption influencing factors stratification to analyze the influence of driving conditions on energy consumption. Then, we innovatively input the speed-related three types of joint probability distribution matrix into the CNN model to predict the motor energy consumption and then input the predicted motor energy consumption and the influencing factors, such as ambient temperature and line information, into the XGBoost model to predict the energy consumption of bus trip. Finally, it is compared and analyzed with various prediction models such as SVR and LGBM, and the results show that the R^2 of the CNN-XGBoost model is 0.923, the prediction performance is significantly higher than other models.*
Keywords: Data-driven, Electric bus, Energy consumption prediction, XGBoost, Driving conditions clustering

1. **Introduction.** The proportion of pure electric buses in recent years has shown a year-on-year increase, and the green development of electric buses is an important part of realizing the low-carbon development of transportation [1]. Mao et al. experimentally showed that electric buses could decrease CO_2 emissions by 18.0-23.9% compared to diesel buses [2]. Electric buses to replace traditional diesel buses are a major trend for future development. However, the development of pure electric buses is currently constrained by shortcomings such as long battery charging time and lack of effective breakthroughs in key battery technologies [3]. In research on solving the practical problems of electric buses, the topics related to the energy consumption of electric buses are a major hotspot at present. Accurate prediction of energy consumption of electric buses enables efficient resource allocation, optimal route selection, and effective management of charging infrastructure,

ultimately improving operational efficiency and cost-effectiveness [4,5]. It is important to analyze what factors mainly affect the energy consumption of pure electric bus trips in order to achieve an accurate prediction of the energy consumption of such trips.

In previous studies, three categories can be used to classify the variables influencing how accurately electric car energy consumption predictions are made: external variables, speed-related variables, and vehicle variables [6]. Vehicle factors include vehicle weight and vehicle windward area, etc. Lajunen et al. analyze the effects of auxiliary equipment, aerodynamic resistance, and rolling resistance on the energy consumption of the whole vehicle through simulation software [7]. Speed-related factors include speed-derived characteristics such as average vehicle speed, acceleration, and the frequency of acceleration and deceleration, which can significantly affectfect on energy consumption [8]. Zhang et al. showed that the energy consumption of electric vehicles increases with the increase in acceleration and deceleration intensity [9]. Lin et al. analyzed the effect of driving behavior on energy consumption by constructing a random forest energy consumption analysis model [10]. External factors include ambient temperature, road type, and number of bus stops [11]. Yi et al. studied the distribution of energy consumption of EVs under different average travel speeds and ambient temperatures [12]. Hao et al. showed that whenever the environmental temperature is less than $10°C$, the power consumption enhances by 2.4kWh/100km for every five ambient temperature decreases [13]. For the road type factor, Wang et al. established an EV energy consumption model considering the effects of weather conditions and road rolling resistance to verify the validity of considering road type in the energy consumption model [14]. El-Taweel et al. added roadway level of service to the model to simulate different traffic conditions and increase the precision of estimates for energy consumption [15].

Therefore, in order to build the electric vehicle energy consumption prediction model, it is necessary to analyze the accurate operation data and examine the influencing factors of energy consumption from multiple perspectives. Based on this, an accurate energy consumption prediction model can be established. In this paper, to solve this problem, we will use accurate operating data to construct an energy consumption prediction model that considers all kinds of factors based on quantitatively analyzing the influencing factors of energy consumption to improve the prediction accuracy of the model.

2. **Literature Review.** There are two basic classes that can be used to categorize the existing energy consumption forecast methods for electric vehicles, which are physical model and data-driven model. Gallet et al. constructed a longitudinal dynamics-based model of energy consumption of electric buses, which takes into account air resistance, rolling friction, gradient, and auxiliary equipment [16]. Szilassy et al. established a road-based characteristics of electric bus trip energy consumption estimation model, which considers the effect of fluctuations in environmental temperature, terrain, parking spacing and passenger load on energy consumption [17]. Energy consumption prediction based on physical models requires accurate vehicle parameters and real-time operation status, but these parameters are more difficult to obtain in practical applications. Furthermore, the generalization ability of the physical model is weak, and the model needs to be re-modeled when the vehicle model is different.

Compared with physical modeling, using data-driven models eliminates the need to build specific mathematical models. It allows multiple features to fit well with labels through the black-box operation of machine learning. Nowadays, with the rapid development of the Internet of Vehicles (IoV), a large amount of real-time data from vehicles can be uploaded to servers for storage [18]. Therefore, the energy consumption prediction of electric vehicles can be realized by using the accurate operating data of pure electric

vehicles combined with big data technology and artificial intelligence algorithms. Pamula et al. built a deep neural network (DLN) that can predict the energy consumption of buses on a stop-by-stop basis, and the model can be applied to bus routes of different sizes [19]. Gao et al. used wavelet neural network (WNN) to train multiple energy consumption influencing factors together with battery energy consumption data to build a multifactor bus prediction model [20]. Fiori et al. construct a predictive model for the instantaneous energy consumption of buses, which includes the instantaneous power required for traction and the average power required for auxiliary systems [21]. Ji et al. constructed a log-linear regression-based model to estimate the trip energy consumption of an electric bus and divided the required energy into traction and battery pack thermal management systems [22]. Abdelaty et al. constructed a multivariate regression model to predict the energy consumption of electric buses, and the model considered the effects of road class, initial SOC, number of passengers, driving behavior, average speed, stopping frequency, and air conditioning system on energy consumption [23]. Chen et al. built long short-term neural network (LSTM) model to predict the instantaneous energy consumption of electric buses, and the model takes instantaneous speed, acceleration, and vehicle specific power weather as inputs [24]. Nan et al. developed an LSTM-XGBoost instantaneous prediction model for pure electric buses, and the proposed model has good time-series and regression prediction capabilities, which can effectively fit the complex volatility of instantaneous values [25]. Li et al. trained different combinations of features of Random Forest in the energy forecasting model, and the results show that the model with integrated consideration of the length of trip, dynamic traffic conditions, environmental conditions, and road characteristics has the best prediction effect [26]. Qin et al. established a Support vector machine regression (SVR), and the Gray Wolf optimization method was used to create a prediction model, and the Gray Wolf algorithm was utilized to optimize the hyperparameters of the SVR algorithm to boost the model's prediction accuracy [27]. Li et al. constructed a simplified physical model to simulate the energy consumption of buses, constructed a CatBoost model considering multiple factors, and finally fused the physical and data-driven models to realize the prediction [28].

In summary, there are two main problems in the existing studies. Firstly, the energy consumption prediction model for pure electric buses does not comprehensively consider energy consumption influencing factors such as bus route information, weather information and driving condition information. Secondly, there is a lack of more in-depth exploration of the factors affecting the energy consumption of pure electric buses. Therefore, in order to solve the limitations of the existing models and improve the accuracy of the electric bus trip energy consumption prediction, this paper uses the accurate driving data of pure electric buses to carry out the following work.

(1) Based on the data preprocessing, analyze the changing law of battery, motor, and auxiliary energy consumption under different ambient temperature intervals. Based on the K-means algorithm to realize driving conditions clustering, and use the energy consumption influence factors stratification method to analyze the influence of different driving conditions on energy consumption.

(2) The three types of joint distribution matrices of the raw operating data are calculated, and the matrices are spliced into a convolutional neural network to predict motor energy consumption. The predicted motor energy consumption is then input into the XGBoost model by combining the influencing factors such as ambient temperature, geographic location information, route information, and speed. Finally, the CNN-XGBoost prediction model is constructed to improve the accuracy of the prediction of trip energy consumption, and the accurate prediction of trip energy consumption can be realized in all temperature intervals and time periods.

3. **Electric Bus Data Preprocessing.**

3.1. **Data analysis and preprocessing.**

3.1.1. *Introduction of data.* The bus data used in this study are accurate operating data of eight pure electric buses, spanning from January 2019 to December 2019. The frequency of sampling for raw data is 0.1Hz. Historical weather data were obtained from a meteorological website (https://rp5.ru/) and were joined together with the driving data based on timestamps. Bus route details were collected from the Gaode map (https://www.amap.com/).

3.1.2. *Data pre-processing.* Due to equipment communication signals, interference during data collection, and vehicle flameout, the raw data will have missing data, data anomalies, and invalid data, so it is necessary to carry out preliminary data preprocessing of the data. The number of times missing data is counted, and it is found that the percentage of missing data is tiny, so the value of the previous moment is directly used to fill in the missing data. The outliers in the data were eliminated by using the method of box-and-line diagram to guarantee the reliability of the data. In the process of data statistics, it is found that the data sampling frequency of a small portion of the segments will change to 1Hz. In order to facilitate the subsequent research in the data, pre-processing of the data will be re-sampled to 0.1Hz.

3.1.3. *Division of bus trips.* The object of this study is the trip energy consumption of electric bus, and the buses run on fixed routes during normal operating hours. Therefore, for the convenience of the study, the raw data are divided into long trip segments and short trip segments. The long trip is the trip between the two ends of the line, i.e., the trip from the starting station to the terminal station or the trip from the terminal station to the starting station.

Because the buses run under fixed routes on a daily basis, the trip is segmented based on the combination of latitude and longitude information of the terminal stations at both ends of the bus route as a sample of one bus trip. The specific steps of the long trip segmentation algorithm are as follows. Firstly, according to the latitude and longitude information in the original dataset, the distance between it and the two end points of the trip is calculated. Then, the bus trip is segmented according to the distance, traversing the dataset, and when the distance from the bus to the starting point is greater than 100m, and the vehicle speed is greater than 5km/h, the time frame at this point is recorded as the starting moment. When the distance from the bus to the endpoint is less than 100m, and the vehicle speed is greater than 5km/h, the time frame at this point is recorded as the end moment.

The specific steps of the short trip segmentation algorithm are as follows. Read the segmented long trip segment and traverse the distance between the latitude and longitude of each time point in the long trip time sequence segment and each bus stop. If the distance between the bus and the nearest station at this point is less than a threshold, add the moment as the bus arrival time to the list of arrival times at that station. The short trips after the division is completed are shown in Table 1, where S_n denotes the nth bus stop.

3.2. **Calculation of electric bus energy consumption.** The ECb (energy consumption of battery) of an electric bus during a trip can be obtained by summing the product of battery voltage and battery current at each moment, and the Equation (1) is expressed as:

$$EC_b = \sum_{i}^{N-1} U_{bi}I_{bi}\Delta t/1000/3600 \tag{1}$$

TABLE 1. Examples of short trip sample segmentation

Sample	S_1	S_2	$\cdots$	S_{n-1}	S_n
0	2019-12-1 08:05	2019-12-1 08:07	$\cdots$	2019-12-1 08:05	2019-12-1 08:07
1	2019-12-1 09:35	2019-12-1 09:37	$\cdots$	2019-12-1 09:35	2019-12-1 09:37
2	2019-12-1 11:02	2019-12-1 11:04	$\cdots$	2019-12-1 11:02	2019-12-1 11:04
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$

where U_{bi} and I_{bi} are the battery voltage and current at the moment i, respectively. N is the total number of time frames of the trip, Δt is the data sampling period. EC_b denotes the total battery expenditure of energy of the bus in one trip.

The sum of the voltage and current results of the motor at each moment is the ECm (Energy consumption of motor) in a trip, it can be written as Equation (2). The ECae (Energy consumption of auxiliary equipment) is the difference between battery energy consumption and motor energy consumption as shown in Equation (3):

$$EC_m = \sum_i^{N-1} U_{mi} I_{mi} \Delta t / 1000 / 3600 \tag{2}$$

$$EC_{ae} = EC_b - EC_m \tag{3}$$

where U_{mi} and I_{mi} are the motor voltage and current at moment i, respectively. EC_m denotes the motor expenditure of energy in a bus trip. EC_{ae} denotes the total auxiliary expenditure of energy of the bus trip. The total expenditure of energy of the battery, motor, and auxiliary equipment divided by the mileage is the corresponding energy consumption per kilometer, which is ECpb (Energy consumption per kilometer of battery), ECpm (Energy consumption per kilometer of motor), and ECpae (Energy consumption per kilometer of auxiliary equipment).

3.3. **Driving conditions clustering based on kmeans.** The K-means algorithm (K-means) is a popular clustering algorithm widely used in data analysis. It is an unsupervised learning algorithm that aids in grouping data points into distinct clusters based on their similarity. The primary stages of the algorithm are as follows: select the number of clusters k and randomly initialize k centers of mass, usually by randomly selecting k data points as the centers of mass. Allocate every data point to the cluster closest to its center of mass. Calculate the average of data points within each cluster and update the cluster's center of mass with this average. Continue repeating the steps until the center of mass stops moving significantly or the maximum number of iterations is reached.

The dataset used for clustering is 49,138 short travel segments with average speed, maximum speed and deceleration ratio as input features. In order to select the optimal number of clusters, this experiment calculates the contour coefficients under different amount of clusters to determine the amount of clusters, and finally selects three categories. The clustering centers of the three categories are calculated separately as shown in Table 2, and the clustering results can be classified into three driving conditions, namely congestion, general and smooth, according to the size of the values of each category.

3.4. **Energy consumption influencing factors analysis.**

TABLE 2. Clustering centers for driving conditions

Driving conditions	Average speed (km/h)	Maximum speed (km/h)	Deceleration ratio (%)
Congestion	10.13	28.80	25.90
General	17.63	38.28	29.42
Free-flowing	30.22	41.66	21.38

3.4.1. *Ambient temperature.* The relationship between ambient temperature and ECpb is shown in Figure 1. Based on the trends in the graph it can be assumed that the energy usage is relatively low in environmental temperatures from $5°C$ to $20°C$, and increases when the environmental temperature is less than $-5°C$ to $5°C$. When the environmental temperature is greater than $25°C$, the average ECbpk arrives at a maximum of 0.81 kWh/km. The maximum value of the difference in ECbpk in different ambient temperature ranges is 0.21 kWh/km.
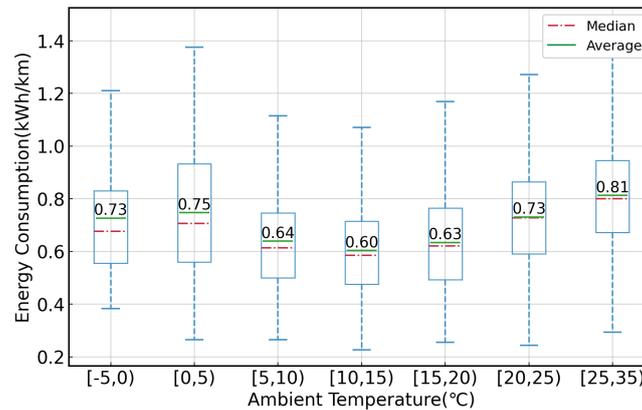

FIGURE 1. Energy consumption per kilometer for each temperature range

The distribution of battery, motor, auxiliary energy consumption and ambient temperature of electric bus in each month as shown in Figure 2. The figure demonstrates that can be inferred: auxiliary energy consumption in March to May and October to December is lower than the rest of the month, and negative correlation with the ambient temperature, auxiliary energy consumption peaked in August. Average motor energy consumption in a year fluctuates little by the influence of ambient temperature is small. The energy consumption of the battery is the sum of auxiliary and motor energy consumption, and the variation pattern of energy consumption is similar to that of auxiliary energy consumption.

3.4.2. *Driving conditions.* This study adopts the stratification method of energy consumption influencing factors to analyze the influence of driving conditions on the ECpb, and the structure of this method is shown in Figure 3. The methodology is founded on the idea of controlling variables, excludes the influence of other factors on energy usage as much as possible, and only studies the distribution of energy consumption under different driving conditions.

The detailed steps are as follows: the first layer is to exclude the effect of passenger load due to the large difference between the busload on weekdays and holidays; on weekdays, there is a significant morning and evening peak phenomenon of bus load, and the
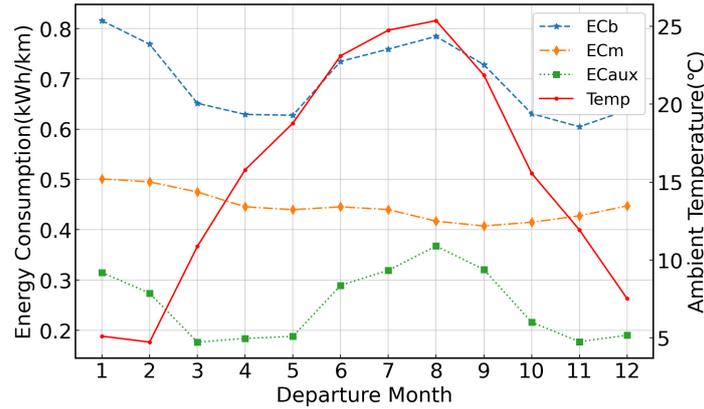
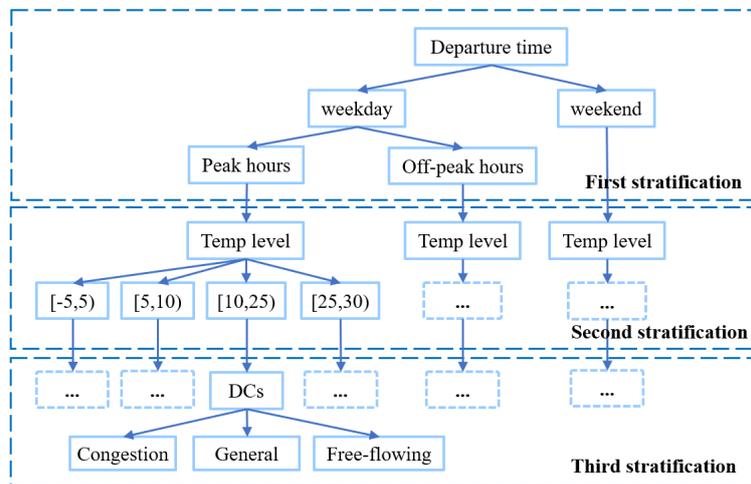FIGURE 2. Distribution of energy consumption and ambient temperature per month

FIGURE 3. Stratification methodology for energy consumption influencing factors

phenomenon is not obvious on holidays. Therefore, the sample was divided into weekday and weekend samples based on the travel times of short trips, and weekdays continued to be divided into peak and off-peak periods. The purpose of the second layer is to exclude the effect of temperature on energy consumption, and the samples are divided into each temperature interval according to the four temperature intervals specified. Finally, the third layer analyzes the effect of driving conditions on energy consumption by analyzing the distribution of ECpb under the three types of driving conditions in each leaf node separately.

Using this method, the impact of driving conditions on energy consumption is analyzed for the case of four temperature levels during weekday peak hours, and the results are shown in Figure 4. The figure demonstrates that can be inferred: the ECpb under congested driving conditions is the largest in all four temperature classes, which is greater than that under smooth driving conditions by 0.27 to 0.1 kWh/km. In the remaining cases, there is a maximum ECpb in congested driving conditions, followed by general driving conditions, and the minimum ECpb is in smooth driving conditions. The main reason is that when the ambient temperature is too heightened, to improve comfort, the

vehicle's air conditioning refrigeration system is turned on, increasing the air conditioning load power and increasing the demand for auxiliary energy consumption.
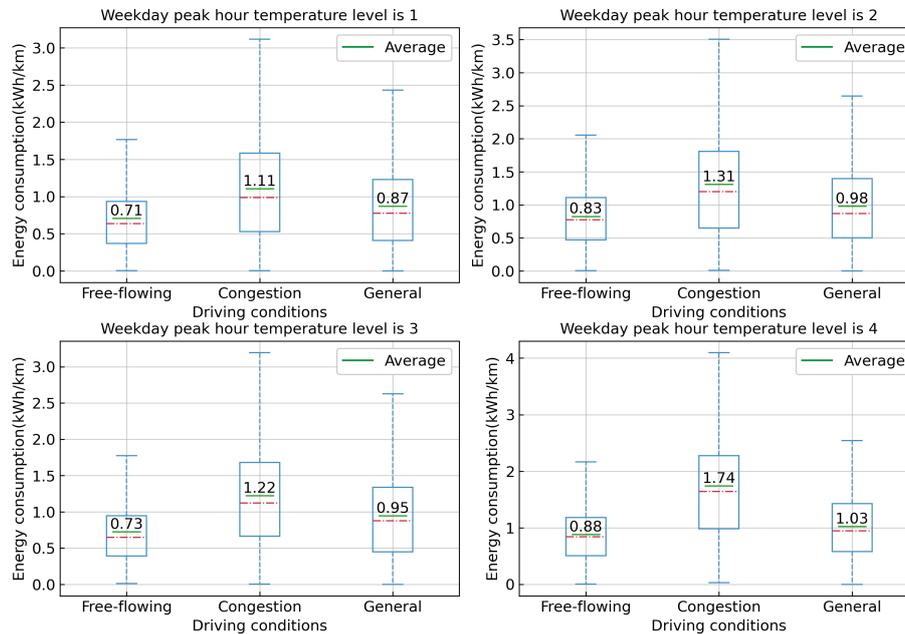

FIGURE 4. Distribution of energy consumption under different driving conditions

4. Energy Consumption Prediction Model.

4.1. **Methodology Framework.** The energy consumption forecasting framework for pure electric buses established in this research is demonstrated in Figure 5. This is divided into three significant steps: (1) Data processing and analysis, data preprocessing of the initial data and division of the long and short trips, and calculation of the energy consumption of each segment. (2) To build an energy consumption prediction model, first extract all the features combined with the trip energy consumption to form a complete regression model training sample. The features with high correlation are then eliminated by calculating the Pearson correlation coefficient. Finally, the samples are proportionally divided into the training and the testing dataset to train each model separately. (3) Model evaluation, evaluate the experimental energy consumption prediction models and select the optimal model.

4.2. **Feature construction and selection.** After the long trip division of the electric bus, seven features can be extracted: specific time, starting SOC, trip distance, starting longitude, latitude, ending longitude, latitude, and ambient temperature. For the departure time, not only the specific departure moment can be extracted, but also the specific month and days of the week can be extracted from it. For driving conditions, the following characteristics are included: average velocity, standard deviation of velocity, average acceleration, average deceleration, standard deviation of acceleration, acceleration and deceleration percentage, uniform speed velocity, and idling ratio.

Redundant features not only increase the computational complexity but also may reduce the ability to generalize the model, leading to the model overfitting the training data. Through feature screening, the most representative and relevant features are selected, thus reducing the overall dimension of the data and enhancing the productivity and precision of the model. In this study, the variance of the features in the sample and
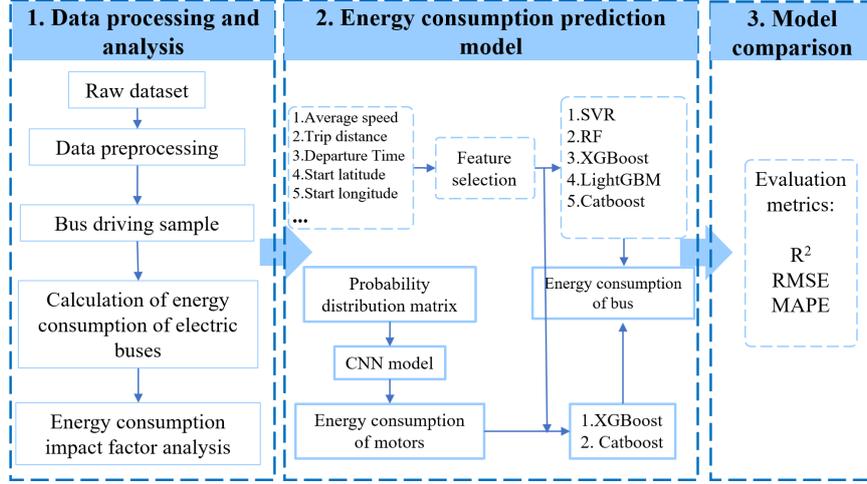
FIGURE 5. Energy consumption prediction model framework

the Pearson's correlation coefficient between the features were calculated separately for feature screening. The variance of each feature in the sample is calculated as shown in Equation (4), and the Pearson correlation coefficient between the features is calculated as shown in Equation (5):

$$var(x_i) = \frac{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}{n} \tag{4}$$

$$r = \frac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2 \sum\limits_{i=1}^{n}(y_i-\overline{y})^2}} \tag{5}$$

where x_i, y_i is the eigenvalue of sample i. $\overline{x}$, $\overline{y}$ is the average value of all sample features. n is the total number of samples in the dataset.

4.3. **XGBoost prediction algorithm.** Based on feature construction and screening, the Extreme Gradient Boost (XGBoost) algorithm is selected in this study to forecast trip energy consumption. XGBoost algorithm developed by Chen et al. belongs to the integrated learning algorithms, where the algorithm's result is determined by the joint decision of multiple interconnected base learners [29]. The algorithm performs well in classification and regression problems with the advantages of insensitivity to input data requirements, high computational efficiency, and the ability to effectively avoid overfitting. In the regression tree of this algorithm, the sum of the anticipated outcomes from K trees constitutes the forecast result, as given in Equation (6):

$$y_i = \sum_{k=1}^{k} f_k(x_i), f_k \in F \tag{6}$$

where x_i is the ith training sample, $f_k(x_i)$ denotes the score of the kth tree. Compared with the gradient tree boosting algorithm XGBoost enhances the objective function as shown in Equation (7). Where the canonical term $\Omega(f_k)$ depends on the base learner used and the canonical term is defined as shown in Equation (8):

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{k}^{K} \Omega(f_k) \tag{7}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j \tag{8}$$

where $\widehat{y}_i$ is the predicted value of energy consumption of data sample i, y_i is the true value of energy consumption of dataset sample i. The regularity term $\Omega(f_k)$ indicates the complexity of the model, T is the number of penalized leaves, and the complexity of the tree is positively correlated with T. ω_j denotes the score of each leaf node, γ and λ are the weighting coefficient. Expanding the loss function using the second order Taylor expansion, Equation (9) can be obtained finally:

$$L^{(t)} = \sum_{i=1}^{n} [l(y_i, \widehat{y}_i^{(t-1)}) + g_i f_t(x_i) + \tfrac{1}{2} h_i f_t{}^2(x_i)] + \sum_{k}^{K} \Omega(f_k) \tag{9}$$

where g_i is the first order derivative of the loss function, h_i is the second order derivative of the loss function, and $\widehat{y}_i^{(t-1)}$ is the predicted value of the ith sample at the $(t-1)st$ iteration. The simplified loss function is obtained by removing the constant term at step t as shown in Equation (10):

$$L^{(t)} \approx \sum_{j=1}^{T} \left((\sum_{i \in I_j} g_i)\omega_j + \tfrac{1}{2}(\sum_{i \in I_j} h_i + \lambda)\omega_j{}^2 \right) + \gamma T \tag{10}$$

where I_j denotes all data samples in leaf node j, $\sum g_i$ is the sum of g_i for all samples on a leaf node, $\sum h_i$ is the sum of h_i for all samples on a leaf node. ω_j is the jth leaf node value. Through Equation (10), the optimal weights of the tree structure can be calculated. Then the greedy algorithm is used to evaluate the loss before and after the node splitting to determine the optimal splitting, and finally the tree structure with the optimal objective function is obtained. XGBoost outperforms other integrated learning algorithms by adding the regularized objective function indicated above. In addition, XGBoost has quick computational speed and improved support for distributed computing systems [30].

4.4. **CNN-XGBoost prediction model.** In order to mine more features to enhance the prediction accuracy from the three types of time series data, namely, speed, accelerator pedal, and brake pedal, a two-step energy consumption prediction method based on CNN-XGBoost is proposed in this study. The joint distribution matrix is extracted from the speed, accelerator pedal, and brake pedal data and input to the CNN model to predict the ECb, which is then combined with the filtered features and input to XGboost to predict the trip energy consumption.

The basic framework for forecasting ECm based on a convolutional neural network (CNN) is shown in Figure 6. Step 1 is to calculate the computational joint distribution matrix and combine these three types of distribution matrices into a 3×10 matrix. Step 2 is to build the motor energy consumption predicted by the CNN, where the input layer of the CNN model is a $3 \times 10 \times 1$ two-dimensional vector, and the output is the motor energy consumption.

The motor energy consumption predicted by the CNN-based model and the features filtered by the Pearson features are jointly used as inputs to the XGBoost model to build the CNN-XGBoost-based trip energy consumption prediction model, and the specific framework of the model is shown in Figure 7. From the figure, it can be seen that the features after Pearson feature selection and the motor energy consumption predicted by the CNN model are a total of nine features input to the XGBoost model. The dataset for the experiment consists of 17463 traveling samples, which are divided into training and test sets in the ratio of 80% and 20%, where the samples for the training and test sets are 13970 and 3493, respectively. Hyperparameter optimization of the XGboost model is performed using the training set and applying the Bayesian algorithm. The
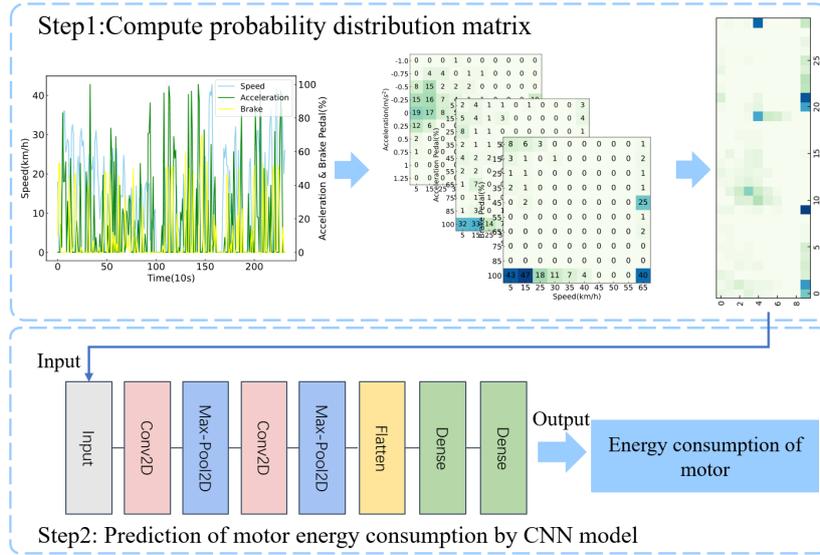
FIGURE 6. CNN framework for predicting motor energy consumption

hyperparameters are input into the model, and the model is trained using the training set. Finally, the test set is used to predict the energy consumption of the bus trip.
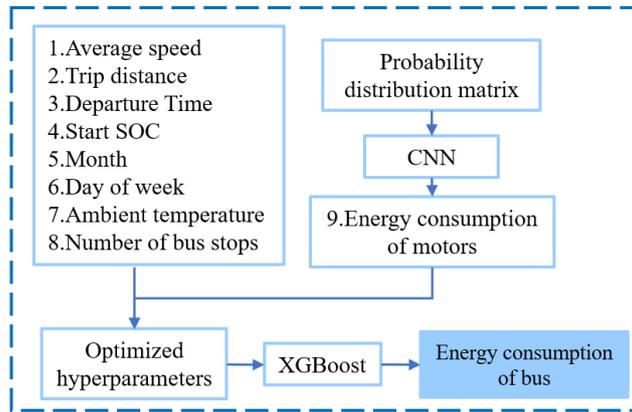

FIGURE 7. Basic framework of CNN-XGBoost modeling

5. Results and Discussion.

5.1. **Comparative analysis of models.** In this study, to validate the superiority of the models developed, commonly used machine learning models for regression prediction were selected for comparative analysis. The comparative models include the support vector regression (AVR), as well as the integrated learning algorithms Random Forest(RF), XGboost, LightGBM, and Catboost, in which the hyperparameters of AVR are searched using a lattice with $c = 3.88$, $epsilon = 0.01$. RF, XGboost, LightGBM, and Catboost use Bayesian hyperparameters for optimization search. The R-squared (R^2), Mean Absolute Percent Error ($MAPE$), and Root Mean Squared Error ($RMSE$) are calculated as in Equation (11-13), respectively.

$$R^2 = 1 - \frac{\sum_1^N (y_i - \widehat{y_i})^2}{\sum_1^N (y_i - \overline{y_i})^2} \tag{11}$$

$$MAPE = \frac{1}{N} \sum_i \frac{|\widehat{y_i} - y_i|}{y_i} \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_i |\widehat{y_i} - y_i|^2}{N}} \tag{13}$$

where N represents the total number of bus long trip samples in the dataset. y_i denotes the true value of the ith sample. $\widehat{y_i}$ denotes the predictive value of the ith sample. $\overline{y_i}$ denotes the average of all sample labels.

The prediction accuracy outcomes of different energy consumption prediction models are displayed in the Table 3. From the table, it can be concluded that the CNN-XGBoost-based model outperforms the other prediction models for all three evaluation metrics. So the CNN-XGBoost model was selected as the final prediction model.

TABLE 3. Model prediction performance comparison

Model	R^2	MAPE/%	RMSE/kWh
SVR	0.724	25.96	2.05
RF	0.907	11.83	1.19
LightGBM	0.916	11.57	1.14
CatBoost	0.916	11.52	1.13
XGBoost	0.918	11.50	1.12
CNN-XGBoost	0.923	10.45	1.08

5.2. **CNN-XGBoost prediction performance analysis.** The box plots of the relative errors of the model at different ambient temperatures are shown in Figure 8. The results show that the average relative error of prediction at different ambient temperatures remains stable, but the average relative error is higher in the range of $[-5, 0)$ at 12.31%. The box plots of the relative errors of the model predictions for each operating hour of the day are shown in Figure 9. It can be seen from the figure that for different departure times of the day, the average prediction error is higher at time periods 7, 14, and 18 with a maximum average relative error of 11.92%. This is because these three time periods, which are at the peak of the traffic and the peak of the passengers, have increased uncertainties, leading to an increase in the prediction error. In summary, the CNN-XGBoost based energy consumption prediction model maintains a stable prediction performance for different temperature intervals and departure times, and the average relative error increases only in the low-temperature and peak time periods.

6. **Conclusions.** In this paper, based on the historical real driving data of electric buses, on the basis of the analysis of energy consumption influencing factors, we comprehensively consider the energy consumption influencing factors such as ambient temperature, bus route information, and driving conditions, and innovatively build a CNN-XGBoost model to realize the two-step trip energy consumption prediction model. The model solves the problem of low accuracy of trip energy consumption prediction of pure electric buses in previous studies, and compared with other energy consumption prediction models, the CNN-XGBoost energy consumption prediction model established in this paper can effectively improve the model's prediction accuracy. The model has an average absolute percentage error of 10.45%, an R-squared of 0.923, and a root mean square error of 1.087
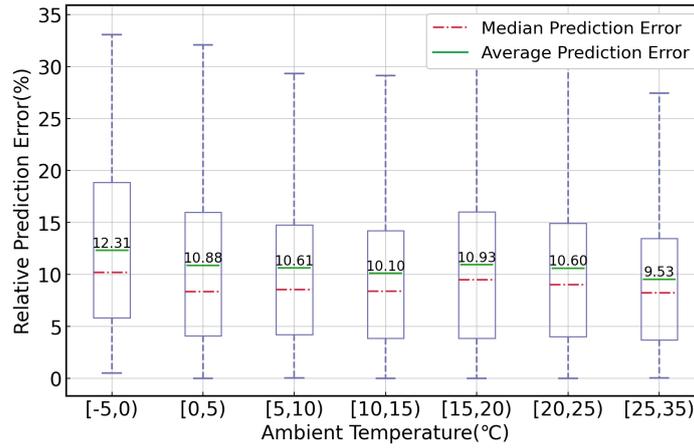
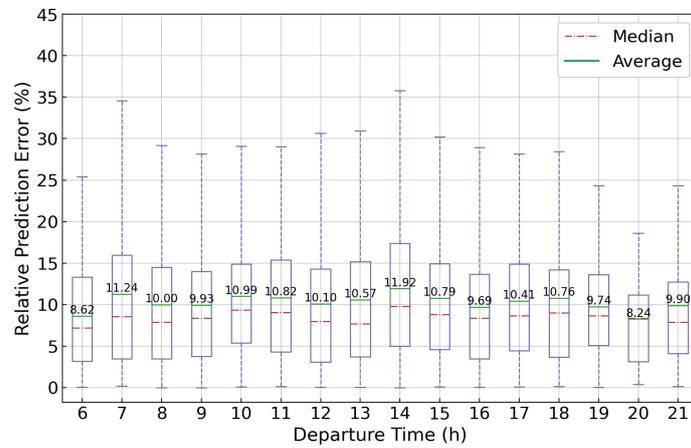FIGURE 8. Relative error distribution for each temperature interval

FIGURE 9. Relative error distribution by time interval

KWh in the test set and achieves accurate prediction of trip energy consumption in all temperature intervals and periods.

In practical application scenarios, the CNN-XGBoost built in this study can be applied to a bus system in a city, which can realize the prediction of the energy consumption of pure electric bus trips for multiple bus routes. In addition, the energy consumption predicted by the model can be used to optimize the bus charging planning and improve the estimation accuracy of the driving range. However, there are also some limitations to this study. The model constructed in this paper did not consider the real-time changes in passenger numbers and road gradients of buses, which can be improved in future research.

REFERENCES

[1] J. A. Manzolli and J. P. Trovão, "A review of electric bus vehicles research topics – methods and trends," *Renewable and Sustainable Energy Reviews*, vol. 159, p. 112211, 2022.

[2] F. Mao, Z. Li, and K. Zhang, "A comparison of carbon dioxide emissions between battery electric buses and conventional diesel buses," *Sustainability*, vol. 13, no. 9, p. 5170, 2021.

[3] R. M. Alexander Kunith and D. Goehlich, "Electrification of a city bus network—an optimization model for cost-effective placing of charging infrastructure and battery sizing of fast-charging electric bus systems," *International Journal of Sustainable Transportation*, vol. 11, no. 10, pp. 707–720, 2017.

[4] K. Liu, H. Gao, Y. Wang, T. Feng, and C. Li, "Robust charging strategies for electric bus fleets under energy consumption uncertainty," *Transportation Research Part D: Transport and Environment*, vol. 104, p. 103215, 2022.

[5] K. An, "Battery electric bus infrastructure planning under demand uncertainty," *Transportation Research Part C: Emerging Technologies*, vol. 111, pp. 572–587, 2020.

[6] H. Abdelaty and M. Mohamed, "A framework for beb energy prediction using low-resolution open-source data-driven model," *Transportation Research Part D: Transport and Environment*, vol. 103, p. 103170, 2022.

[7] A. Lajunen, K. Kivekaes, and F. Baldi, "Different approaches to improve energy consumption of battery electric buses," in *2018 IEEE Vehicle Power and Propulsion Conference VPPC*, 2018, pp. 1–6.

[8] R. Galvin, "Energy consumption effects of speed and acceleration in electric vehicles: Laboratory case studies and implications for drivers and policymakers," *Transportation Research Part D: Transport and Environment*, vol. 53, pp. 234–248, 2017.

[9] J. Zhang, Z. Wang, and P. Liu, "Energy consumption analysis and prediction of electric vehicles based on real-world driving data," *Applied Energy*, vol. 275, p. 115408, 2020.

[10] K.-C. Lin, C.-N. Lin, and J. J.-C. Ying, "Construction of analytical models for driving energy consumption of electric buses through machine learning," *Applied Sciences*, vol. 10, p. 6088, 2020.

[11] H. Basma, C. Mansour, M. Haddad, M. Nemer, and P. Stabat, "Comprehensive energy modeling methodology for battery electric buses," *Energy*, vol. 207, p. 118241, 2020.

[12] Z. Yi, J. Smart, and M. Shirk, "Energy impact evaluation for eco-routing and charging of autonomous electric vehicle fleet: Ambient temperature consideration," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 344–363, 2018.

[13] X. Hao, H. Wang, Z. Lin, and M. Ouyang, "Seasonal effects on electric vehicle energy consumption and driving range: A case study on personal, taxi, and ridesharing vehicles," *Journal of Cleaner Production*, vol. 249, p. 119403, 2020.

[14] H. Wang, X. Zhang, and M. Ouyang, "Energy consumption of electric vehicles based on real-world driving patterns: A case study of beijing," *Applied Energy*, vol. 157, pp. 710–719, 2015.

[15] N. A. El-Taweel, A. Zidan, and H. E. Z. Farag, "Novel electric bus energy consumption model based on probabilistic synthetic speed profile integrated with hvac," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1517–1531, 2021.

[16] M. Gallet, T. Massier, and T. Hamacher, "Estimation of the energy demand of electric buses based on real-world data for large-scale public transport networks," *Applied Energy*, vol. 230, pp. 344–356, 2018.

[17] P. A. Szilassy and D. Foldes, "Consumption estimation method for battery-electric buses using general line characteristics and temperature," *Energy*, vol. 261, p. 125080, 2022.

[18] T.-Y. Wu, X. Guo, L. Yang, Q. Meng, and C.-M. Chen, "A lightweight authenticated key agreement protocol using fog nodes in social internet of vehicles," *Mobile Information Systems*, vol. 2021, pp. 1–14, 2021.

[19] T. Pamuła and W. Pamuła, "Estimation of the energy consumption of battery electric buses for public transport networks using real-world data and deep learning," *Energies*, vol. 13, p. 2340, 2020.

[20] Y. Gao, S. Guo, J. Ren, Z. Zhao, A. Ehsan, and Y. Zheng, "An electric bus power consumption model and optimization of charging scheduling concerning multi-external factors," *Energies*, vol. 11, p. 2060, 2018.

[21] C. Fiori, M. Montanino, S. Nielsen, M. Seredynski, and F. Viti, "Microscopic energy consumption modelling of electric buses: model development, calibration, and validation," *Transportation Research Part D: Transport and Environment*, vol. 98, p. 102978, 2021.

[22] J. Ji, Y. Bie, Z. Zeng, and L. Wang, "Trip energy consumption estimation for electric buses," *Communications in Transportation Research*, vol. 2, p. 100069, 2022.

[23] H. Abdelaty and M. Mohamed, "A prediction model for battery electric bus energy consumption in transit," *Energies*, vol. 14, no. 10, p. 2824, 2021.

[24] Y. Chen, Y. Zhang, and R. Sun, "Data-driven estimation of energy consumption for electric bus under real-world driving conditions," *Transportation Research Part D: Transport and Environment*, vol. 98, p. 102969, 2021.

[25] S. Nan, R. Tu, T. Li, J. Sun, and H. Chen, "From driving behavior to energy consumption: A novel method to predict the energy consumption of electric bus," *Energy*, vol. 261, p. 125188, 2022.

[26] P. Li, Y. Zhang, Y. Zhang, K. Zhang, and M. Jiang, "The effects of dynamic traffic conditions, route characteristics and environmental conditions on trip-based electricity consumption prediction of electric bus," *Energy*, vol. 218, p. 119437, 2021.

[27] W. Qin, L. Wang, Y. Liu, and C. Xu, "Energy consumption estimation of the electric bus based on grey wolf optimization algorithm and support vector machine regression," *Sustainability*, vol. 13, p. 4689, 2021.

[28] X. Li, T. Wang, J. Li, and Y. Tian, "Energy consumption estimation for electric buses based on a physical and data-driven fusion model," *Energies*, vol. 15, p. 4160, 2022.

[29] N. Yan, Y.-B. Yao, and Z.-D. Jia, "Online battery health diagnosis for electric vehicles based on dtw-xgboost," *Energy Reports*, vol. 8, pp. 121–128, 2022.

[30] X. Deng, A. Ye, and J. Zhong, "Bagging–xgboost algorithm based extreme weather identification and short-term load forecasting model," *Energy Reports*, vol. 8, pp. 8661–8674, 2022.