Applied Energy 282 (2021) 116175

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Data-driven framework for large-scale prediction of charging energy in electric vehicles

Yang Zhao [a,b], Zhenpo Wang [a,*], Zuo-Jun Max Shen [b,*], Fengchun Sun [a]

[a] National Engineering Laboratory for Electric Vehicles, Beijing Institute of Technology, Beijing 100081, China
[b] Department of Civil and Environmental Engineering and Department of Industrial Engineering and Operations Research, University of California Berkeley, Berkeley, CA, USA

HIGHLIGHTS

- A novel framework for large-scale EV charging energy predictions is introduced.
- The MAPE retains at 2.5–3.8% with a testing/training ratio varying from 0.1 to 1000.
- MICs and PCCs are combined for feature analyses of charging energy predictions.
- Multiple data sources are coupled by linking the timestamps and location data.

ARTICLE INFO

Keywords:
Charging energy
Large-scale prediction
Machine learning
Electric vehicle

ABSTRACT

Large-scale and high-precision predictions of the charging energy required for electric vehicles (EVs) are essential to ensure the safety of EVs and provide reliable inputs for grid-load calculations. However, the complex and dynamic operating conditions of EVs make it challenging to accurately predict the charging energy under real-world conditions, especially for large-scale EV utilization. In this study, a novel data-driven framework for large-scale charging energy predictions is developed by individually controlling the strongly linear and weakly nonlinear contributions. The proposed framework concurrently addresses the overfitting of nonlinear networks using a low proportion of training data as well as the poorly descriptive ability of linear networks under complex environments. For each charging session, the charging energy predictions appropriately account for important factors such as the variations in the state of charge (SOC) of the battery, ambient temperatures, charging rates, and total driving distances. The results suggest that, compared with existing prediction models (such as the random forest, xgboost, and neural network), the proposed framework persists with evidently higher accuracy and stability over a wide range of the ratio between the number of EVs used for testing and training; its mean absolute percentage error (MAPE) is maintained at 2.5–3.8% when the ratio ranges from 0.1 to 1000. The proposed models can be further utilized for cloud-based battery diagnoses and large-scale forecasting of the energy demands of EVs.

1. Introduction

Electrified transportation, considered as an effective solution to atmospheric pollution, has experienced rapid scaling up in recent years [1]. The global stock of light-duty electric vehicles (EVs) surpassed 7.5 million in 2019 and is anticipated to reach 140 million by 2030, according to the New Policies Scenario [2]. This widespread deployment of EVs and the advanced technologies of Internet of Things have promoted the development of numerous cloud-based services for EVs [3,4]. Multiple parties, including governments [5] and manufacturers [6], have established integrated data centers for plug-in or hybrid EVs. In this context, developments through research on large-scale EV applications are essential and are expected to be highly popular in the near future.

EV charging has been considered as an impactful factor that affects

* Corresponding authors at: National Engineering Laboratory for Electric Vehicles, Beijing Institute of Technology, No. 5 South Zhongguancun Street, Haidian District, Beijing 100081, China. (Z. Wang). Department of Civil and Environmental Engineering and the Department of Industrial Engineering and Operations Research, University of California at Berkeley, Berkeley, California 94720, USA. (Z.-J.M. Shen).
E-mail addresses: wangzhenpo@bit.edu.cn (Z. Wang), maxshen@berkeley.edu (Z.-J.M. Shen).

https://doi.org/10.1016/j.apenergy.2020.116175
Received 17 August 2020; Received in revised form 29 October 2020; Accepted 30 October 2020
Available online 19 November 2020

the adoption of EVs and the demand and economic costs of urban power [7]. Compared to the expeditious refueling of gasoline/diesel vehicles, the recharging speed of EVs has needed to be improved to convince more customers to purchase EVs [8]. In the meantime, since EV charging stations are inadequately constructed worldwide, the relationships between EV activities and the number, location, and scale of related recharging infrastructure have become popular topics [9,10]. A study [11] indicated that the global average number of chargers for EVs was 153 chargers per 1000 plug-in EVs, while in China and the Netherlands, the numbers were 217 and 239, respectively. To better support the design of charging stations, many efforts [12,13] have modeled and optimized the rated power of charging facilities, the number of parking slots, and renewable energy utilization. On the other hand, the impact of EV charging on the power grid has become increasingly nonnegligible as the scale of EV applications increases rapidly. This has promoted research on charging demand forecasting [14], vehicle-to-grid technologies [15], and energy management of EVs [16]. In [17], an advanced energy management system that can automatically select the charging time and patterns was proposed to achieve a total cost reduction of over 50%. A study [18] showed that peak power demands can be shed by up to 47% if certain recommendations to change charging times of EVs were implemented.

The prediction of EV charging energy is critical for providing a methodological basis for EV-related energy supply analyses and monitoring the safety of EVs [19,20]. Charging energy predictions that rely on data-driven methods are more adaptive and cost-effective than traditional physical models. By associating with cloud-based technologies, data-driven models can be modified or upgraded freely. Numerous studies [21,22] have focused on methods to leverage the power of big data analyses and data-driven methods through research on EVs. However, predictions on the battery energy in EVs suffer from two difficulties. First, the charge/discharge battery energy can be altered due to varying real-world operating conditions, and the complexity of these variations would be exacerbated in large-scale EV applications. Second, for large-scale predictions, the number of EVs that need to be predicted is typically significantly greater than that used for training. Existing machine learning-based regression models face difficulty in yielding high-precision results when the ratio of testing/training data sizes varies greatly. Owing to these challenges, the prediction of the battery energy state in EVs has been receiving extensive attention. Several previous studies have predicted the remaining battery energy by using an adaptive estimator [23] and the online approach [24]. Similarly, other analyses attempted to utilize advanced control algorithms or machine learning models to estimate the state of energy of the battery [25,26]. Such efforts either estimated battery energy states via experiments and small-scale tests or analyzed the macroscopic energy demand of charging stations for EVs. Very few of these studies have focused on large-scale predictions of real-world EV charging energy. This deficit has hindered the research on large-scale electricity demands of urban EVs and the planning of charging infrastructures.

This study aims at addressing these deficits by developing a novel high-precision framework for the large-scale prediction of EV charging energy. First, we combine three datasets including EV operating data (such as SOCs, charging energy, and locations), climatic data, and vehicle feature data (such as vehicle types). In this case, a portion of EV operating data (such as SOCs and total driving distances) and climatic data are employed as feature data in prediction modeling, while charging energy data are used as label data; vehicle feature data are used for vehicle classification and result calibration. Thereafter, a data-driven prediction framework is established by individually controlling linear and nonlinear contributions. Subsequently, real-world charging profiles of light-duty EVs are utilized to examine the accuracy and stability of the prediction results for different seasons. Different prediction models are applied to EVs, and the performances of these models are compared from the perspective of big data. It is found that the proposed charging energy prediction model can achieve better accuracy and stability when utilized for real-world vehicle operation and for EV groups of different sizes. The remainder of this work is summarized as follows. The real-world datasets employed in this work are described in Section 2. The proposed methodological framework and prediction models are detailed in Section 3. The prediction results and performance evaluations are detailed in Section 4. Conclusions and future works are discussed in Section 5.

2. Big data platform and real-world EV data

To incorporate real-world EV charging profiles in the analyses, we utilize the datasets and cloud computing ability of the National Monitoring and Management Center for New Energy Vehicles, which is China's national big data platform for EVs. The framework of the large-scale EV data processing used in this work is illustrated in Fig. 1A; multiple modules, including data storage, distributed computing ability, interfaces for statistics and mathematical algorithms, and data visualization tools, are utilized collaboratively for the large-scale predictions. The content of EV data primarily includes general vehicle states (such as the velocity and total driving distance), subsystem data (such as the total voltage of the battery system and the motor power), location data, and other user-defined data.

The starting and ending battery SOC during real-world charging sessions are generally irregular, because it is practically impossible for EVs to completely deplete their batteries during daily usage. By identifying the switching points between various EV states (such as charging, driving, and parking), fragments of the charging sessions are extracted from long-term EV operating data. Fig. 1B displays the variations in the SOC during the charging sessions of a light-duty EV with a rated driving range of 252 km. The disordered bars demonstrate the irregularity in EV charging profiles during real-world operation. Moreover, owing to the strong correlation between battery capacity and energy, the charging energy of EVs also exhibits irregularity (Fig. 1C and Fig. 3A). In addition, the data in ref. [27,28] also indicate that the electrochemical processes during EV battery charging are affected by several factors such as the ambient temperature and charging rate. Thus, to improve the precision of charging energy prediction, additional attributes are included with the original charging profiles by linking multiple data types, such as regional climatic data, general vehicle states, spatiotemporal data, and charging states. The time frame of the EV data used in this study ranges from 2017 to 2019. A combined EV data sample is presented in Table 1. Vehicle specification datasets are acquired from the open databases of EV manufacturers and the Ministry of Industry and Information Technology of China. The climate data of different regions worldwide are collected from the open datasets of Weather Underground (https://www.wunderground.com/) and Reliable Prognosis (https://rp5.ru/).

3. Methods

3.1. Data-driven framework for large-scale EV charging energy predictions

This section presents the novel data-driven framework used for retaining a high accuracy during large-scale EV charging energy predictions. The structure of this framework is depicted in Fig. 2; it consists of four modules including input data, correlation analyses, prediction models, and outputs. The detailed mathematical illustrations of correlation analyses and prediction models are given in Section 3.3 and 3.4, respectively. First, four features are incorporated into the input data: SOC variations, ambient temperatures, total driving distances (reflecting the aging states of EV batteries), and time per SOC change (representing charging rates). Second, correlation analyses and feature filters are employed to classify input features according to the strengths of their relationships. Third, prediction models utilize these training data with feature classifications to obtain final prediction outputs. Different features that affect battery charging processes contribute differently to the charging energy [29,30]. For example, during a charging session, the

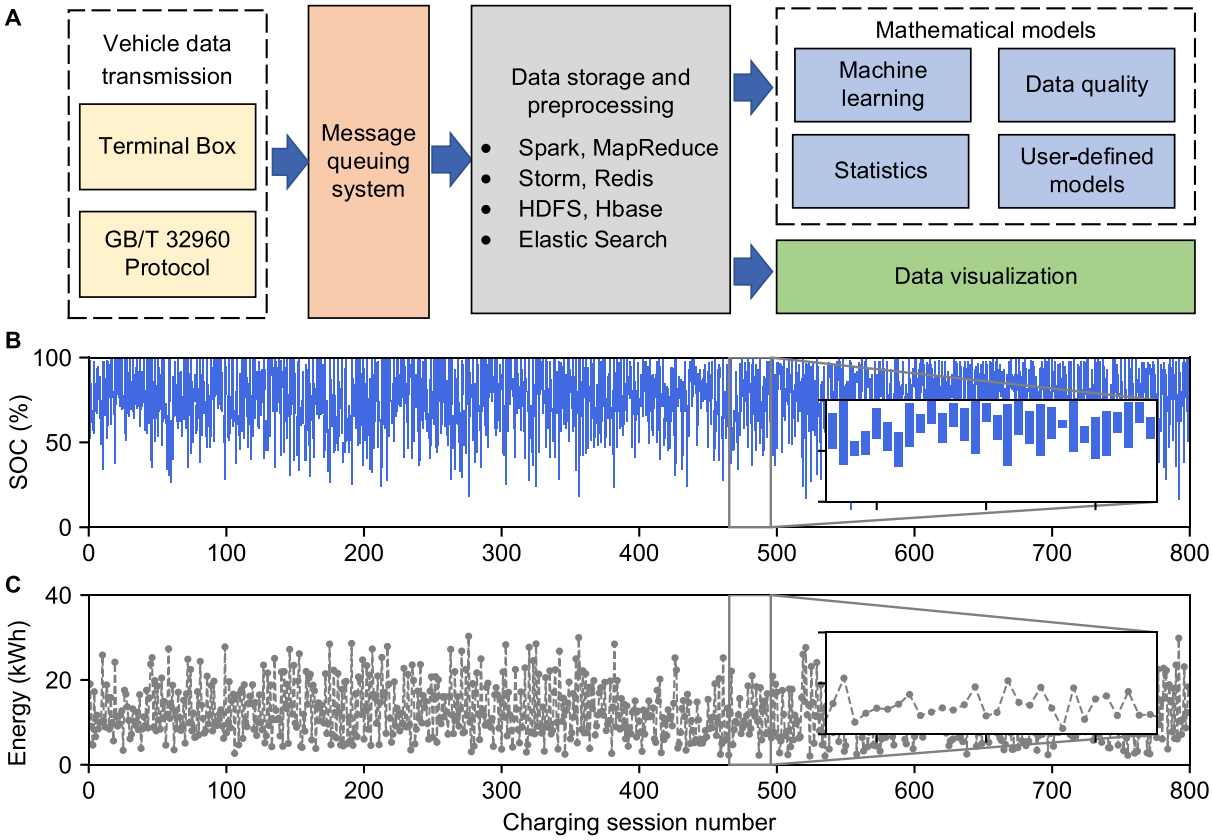

Fig. 1. Large-scale data processing and EV data. (A) Big data processing for large-scale EV applications. (B) Real-world charging profiles of a light-duty EV. For each bar, the bottom represents the starting SOC during a charging session, while the top represents the ending SOC. (C) Charging energy corresponding to SOC variations. Each point represents the amount of charged energy during a charging session.

Table 1
Combined data sample of an EV.

Start time	End time	Start SOC	End SOC	Vehicle state	Charging energy (k·Wh)	Total driving distance (km)	Temperature (°C)	…	Region
2019-02-24 23:29:07	2019-02-25 00:46:45	34	99	Charging	23.6	37,509	20.7	…	Guangzhou
2019-02-25 07:32:24	2019-02-25 08:33:44	46	99	Charging	20.2	37,691	23.0	…	Guangzhou
2019-02-27 23:50:57	2019-02-28 01:53:25	10	99	Charging	32.6	37,864	26.9	…	Guangzhou
2019-02-28 07:10:39	2019-02-28 07:58:59	19	88	Charging	26.0	38,024	28.1	…	Guangzhou
2019-02-28 10:13:02	2019-02-28 10:49:52	54	95	Charging	15.8	38,098	29.2	…	Guangzhou

amount of charge entering the battery has a more direct impact on the charging energy than the ambient temperature. These unbalanced contributions amplify the difficulties in maintaining the prediction accuracy, especially when the testing/training ratios are high. Typically, for large-scale EV predictions, the number of prediction targets notably exceeds the training samples. To address these problems, a dual-layer prediction model, coupled with linear and nonlinear correlation analyses, has been utilized to individually control the contributions of different features.

The linear and nonlinear correlations between these features and the charging energy are investigated by using the Pearson correlation coefficient (PCC) and the max information coefficient (MIC). To compare the correlation results across different features, the absolute PCCs and the MICs between charging energy and different features are shown in Fig. 3A and B, respectively. It is evident that the relationship between charging energy and SOC variations is stronger than that between charging energy and the other features (the absolute value of PCC and the MIC reach 0.98 and 0.85, respectively). According to the classifications in Table 2, the relationship between charging energy and SOC variations is strongly linear. When the parameters of the regression models are determined via a training process, this strong relationship with SOC variations can significantly curtail the description ability of the regression models for the other features.

To address this problem, a linear model was first employed to learn the contributions of SOC variations. Thereafter, contributions from the linear model were eliminated, and a tree-based model was utilized to learn the residual nonlinear components. Finally, the outputs of these two prediction models were combined to obtain the final output. An iterative process of learning and elimination was employed to achieve better separation between the linear and nonlinear contributions. The details of this prediction modeling are illustrated in Section 3.4. As shown in Fig. 3B, based on the MICs before and after the elimination, the

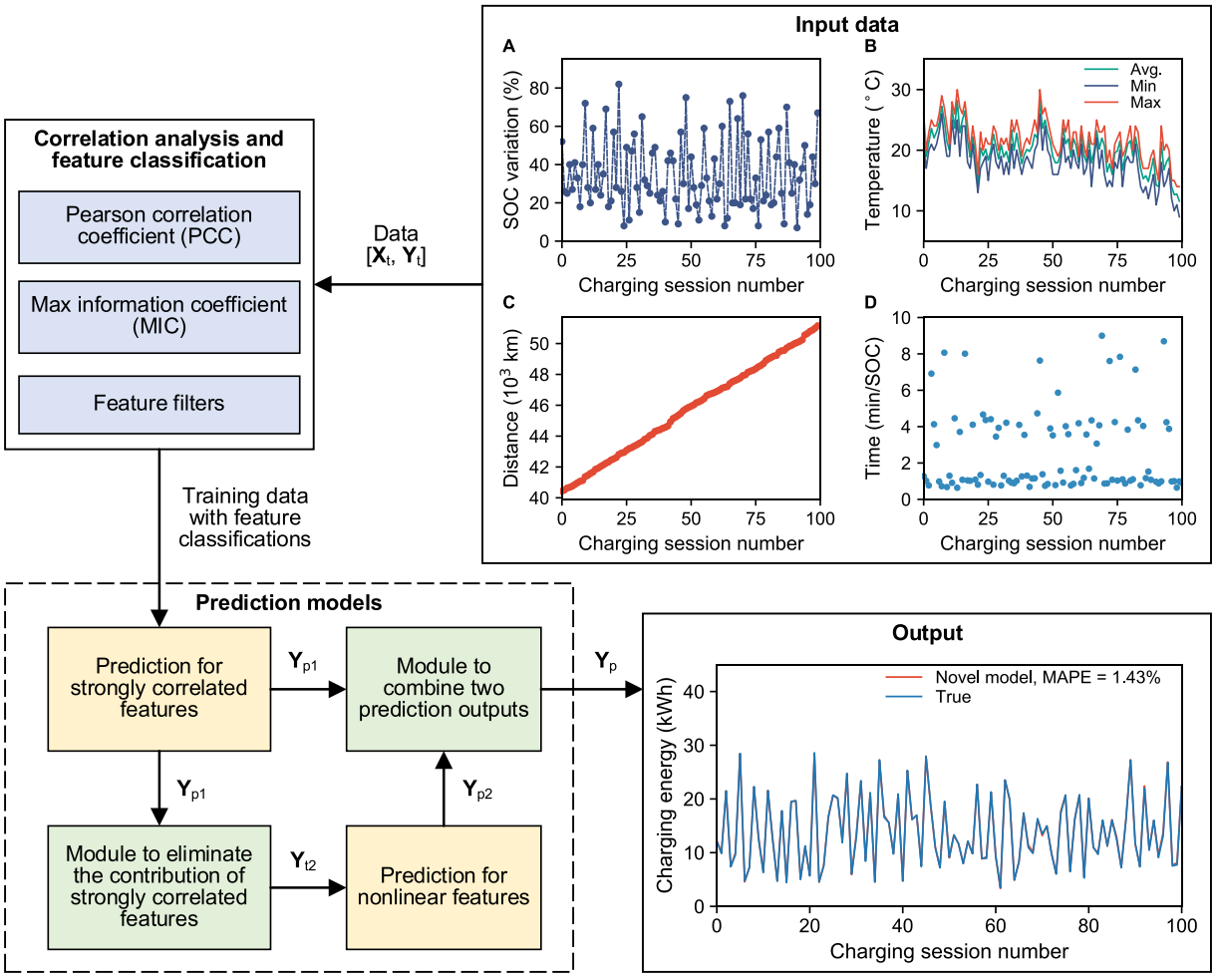

Fig. 2. Large-scale prediction framework.

strength of the relationship between charging energy and SOC variations decreases significantly after the elimination, whereas the strength of the relationship between charging energy other features increases. As a stronger or lower-noise relationship can lead to better learning of the regression model, this observation further indicates the importance of the framework.

3.2. Framework to combine features with different continuity characteristics

A continuity analysis of the different features has been also considered during the prediction modeling. Among the four targeted features, three features—SOC variations, ambient temperatures, and total driving distances—can increase or decrease freely. Nevertheless, the charging rates of EVs are generally predefined; thus, charging rates can only be selected under several fixed modes. Fig. 3C presents the distribution of the charging rates of a light-duty EV and three classifications of these charging rates are described in Fig. 3D. In this study, a framework for combining the prediction and classification models is developed to improve the performance of charging energy predictions (Fig. 4).

First, a classification model was utilized to divide the target datasets with respect to different charging rates. The classification model is established as follows:

$$\begin{cases} \phi \leftarrow (\mathbf{X}_t, \mathbf{C}_t) \\ \mathbf{C}_t = f_{clu}(\mathbf{X}_t) \end{cases} \tag{1}$$

where ϕ is a classifier built by using the decision tree model [31]; $\mathbf{X}_t$ is the feature data matrix that utilized to train the classification model; $\mathbf{C}_t$ is the corresponding label matrix that can be obtained by manual labeling or automatic clustering; f_{clu} is a clustering process by using the DBSCAN model [32]. The classification of target data is described as

$$\mathbf{X}_p \xrightarrow{\phi} \{\mathbf{X}_{p,1}, \cdots, \mathbf{X}_{p,n}\}, \tag{2}$$

where $\mathbf{X}_p$ are the testing data to be classified; $\mathbf{X}_{p,i}$ are the testing data classified into category i; n is the total number of classifications. Subsequently, by training models with corresponding datasets, prediction models with the same number of classifications are obtained. That is

$$m = \{m_1, m_2, \cdots, m_n\}, \tag{3}$$

where m_i is the prediction model in category i and multiple models can be used here, such as the random forest regression, neural network regression, and the proposed model; m is the set of prediction models. Finally, prediction outputs ($\mathbf{Y}_p$) were generated by combining the results from the prediction models across different categories (Eq. (4)). By individually learning the charging processes with different charging rates, this framework can better leverage features with different continuity characteristics. The results of the performance comparison are illustrated in Fig. 6.

$$\{m_1(\mathbf{X}_{p,1}), m_2(\mathbf{X}_{p,2}), \cdots, m_n(\mathbf{X}_{p,n})\} \rightarrow \mathbf{Y}_p, \tag{4}$$

3.3. Correlation calculation

To quantify the strength of the relationships between EV charging

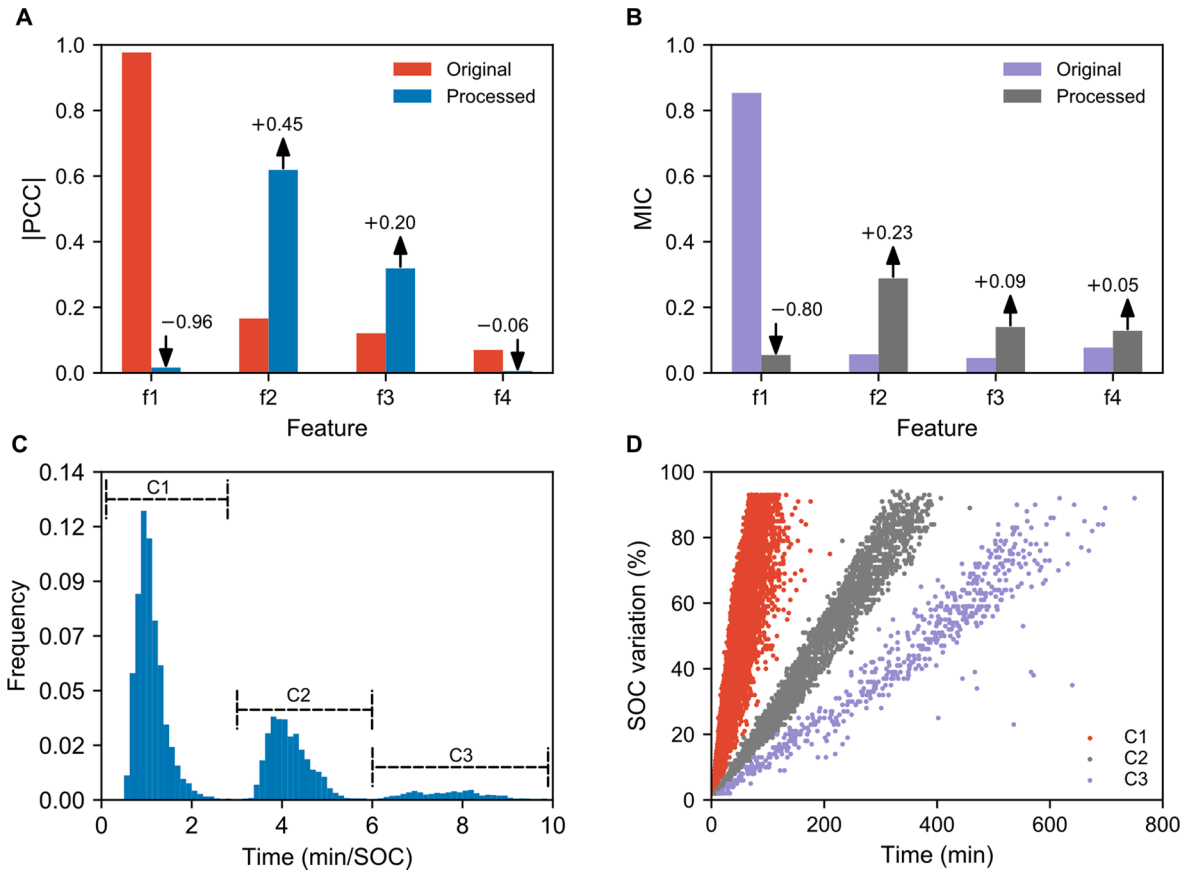

Fig. 3. Correlation and feature analyses. Four features (f1, f2, f3, and f4) represent SOC variations, ambient temperatures, total driving distances, and the time per SOC change, respectively. C1, C2, and C3 are the classification results of the EV charging rates.

Table 2
Classification of relationships.

Class	|PCC|	MIC	Relationship
C1	High	High	Strongly linear
C2	High	Low	Noisy linear
C3	Low	High	Strongly nonlinear
C4	Low	Low	Independent

energy and the other features, linear and nonlinear correlations are investigated based on the PCC and MIC. The MIC, which was introduced by Reshef et al. in 2011 [33], is a maximal information-based measure that describes the strength of the linear or nonlinear relationship between two variables. This method is based on acquiring normalized mutual information, which is calculated by using a data-dependent binning scheme [34]. To guarantee a fair comparison, the output values are normalized between 0 and 1. By using this method, the dependencies of charging energy and the other features can be examined, despite their linear or nonlinear relationships. The equation of a PCC is given in Eq. (5); the equations of a MIC are given in Eq. (6) and Eq. (7).

$$C_{\mathrm{pcc}}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{5}$$

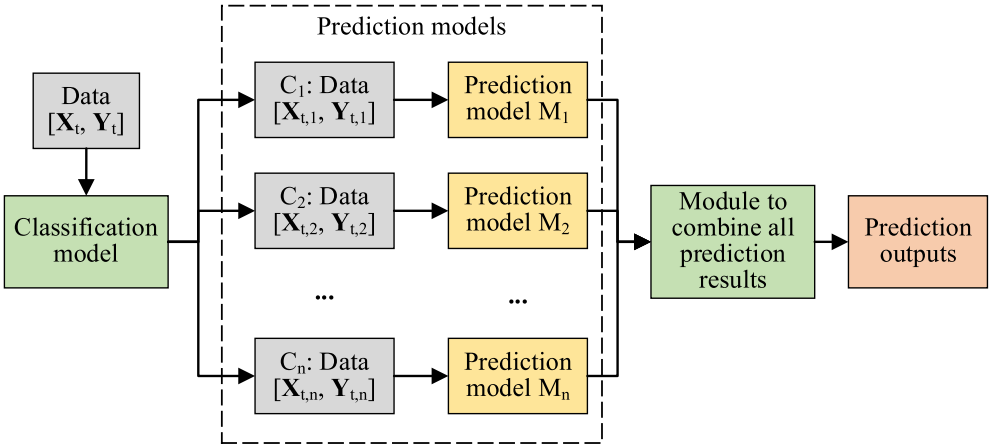

Fig. 4. Framework combining the regression and classification models.

where n is the sample size; $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively.

$$C_{\mathrm{mic}}(x,y)=\frac{max\{I_{\mathrm{G}}(x;y)\}}{log_2 min\{n_x,n_y\}} \tag{6}$$

$$I_{\mathrm{G}}(x;y)=\sum_{\tilde{x},\tilde{y}}\hat{p}(\tilde{x},\tilde{y})log_2\frac{\hat{p}(\tilde{x},\tilde{y})}{\hat{p}(\tilde{x})\hat{p}(\tilde{y})} \tag{7}$$

where I_{G} is the mutual information of the binning scheme G (n_{x}−by−n_{y} grid), and $\hat{p}(\tilde{x},\tilde{y})$ is the fraction of data points in the bin $(\tilde{x},\tilde{y})$. The binning scheme is chosen under the constraint of $n_x n_y < N^{0.6}$ where N is the number of total data points.

Combinations of PCCs and MICs are used to categorize the relationships, as presented in Table 2. A high MIC indicates that the relationship between two variables is strong, despite their linear or nonlinear correlations. In this case, if the absolute PCC of this relationship is also high, the strong relationship can be determined as a strong linear relationship (corresponding to class C1 in Table 2). If the absolute PCC is small and the MIC is high, this strong relationship can be classified as a strongly nonlinear one. By contrast, a low MIC indicates a weak association between two variables.

3.4. Prediction modeling

Herein, the mathematical illustration of the prediction modeling is presented. The numerical scales of different input features are different; for example, in this study, the total driving distances of the EVs vary from 0 to 200,000 km, and the SOC varies from 0% to 100%. Therefore, linear and tree-based models are utilized to describe the linear and nonlinear contributions. The constructed function of charging energy, $\delta_{\mathrm{c}}(\mathbf{x}_{\mathrm{t}};\mathbf{w})$, is defined as

$$\delta_{\mathrm{c}}(\mathbf{x}_{\mathrm{t}};\mathbf{w})=\zeta\left(\mathbf{x}_{\mathrm{t1}};\widehat{\mathbf{w}}_1\right)\xi\left(\mathbf{x}_{\mathrm{t2}};\widehat{\mathbf{w}}_2\right), \tag{8}$$

where $\zeta\left(\mathbf{x}_{\mathrm{t1}};\widehat{\mathbf{w}}_1\right)$ is a linear model, and $\xi\left(\mathbf{x}_{\mathrm{t2}};\widehat{\mathbf{w}}_2\right)$ is a tree-based model; $\mathbf{x}_{\mathrm{t1}}$ and $\mathbf{x}_{\mathrm{t2}}$ are the feature matrices with selected features, which are subsets of the original feature matrix $\mathbf{x}_{\mathrm{t}}$. $\widehat{\mathbf{w}}_1$ and $\widehat{\mathbf{w}}_2$ are the coefficient vectors, which are calculated as

$$\begin{bmatrix}\widehat{\mathbf{w}}_1^{(i)}\\ \widehat{\mathbf{w}}_2^{(i)}\end{bmatrix}=\begin{bmatrix}\underset{\mathbf{w}_1}{\mathrm{argmin}}\left\|\zeta^{(i)}(\mathbf{x}_{\mathrm{t1}};\mathbf{w}_1)-\mathbf{y}_1^{(i-1)}\right\|_2^2\\ \underset{\mathbf{w}_2}{\mathrm{argmin}}L_{\mathrm{t}}\left(\xi^{(i)}(\mathbf{x}_{\mathrm{t2}};\mathbf{w}_2),\mathbf{y}_2^{(i)}\right)\end{bmatrix}, \tag{9}$$

where L_{t} is the loss function of the tree-based model, and $\mathbf{y}_1^{(i)}$ and $\mathbf{y}_2^{(i)}$ are the label vectors generated in the i-th iteration, which are

$$\begin{bmatrix}\mathbf{y}_2^{(i)}\\ \mathbf{y}_1^{(i)}\end{bmatrix}=\mathbf{y}_{\mathrm{t}}\begin{bmatrix}1/\zeta^{(i)}\left(\mathbf{x}_{\mathrm{t1}};\widehat{\mathbf{w}}_1^{(i)}\right)\\ 1/\xi^{(i)}\left(\mathbf{x}_{\mathrm{t2}};\widehat{\mathbf{w}}_2^{(i)}\right)\end{bmatrix}, \tag{10}$$

where the initial vector of $\mathbf{y}_1^{(i)}$ is the original charging energy vector $\mathbf{y}_{\mathrm{t}}$. The number of iterations i is less than 3 in this study. It should be noted that $\mathbf{y}_1^{(i)}$ is the extracted linear contribution, and $\mathbf{y}_2^{(i)}$ denotes that the influence of the strongly correlated feature is eliminated.

3.5. Validation models

To validate the proposed model, five existing high-performance models are employed in this work. They are the linear model (Linear), proportion model (PM), neural network model (NN), random forest model (RF), and xgboost model (XGB). Note that the proportion model is a linear regression model only requiring the input of SOC variations. Since these five models are rather mature, the mathematical illustrations of them can be found in [35] (Linear), [36] (NN), [37] (RF), and [38] (XGB). For linear and proportion models, the target value is expected to be a linear combination of the features. For this reason, these two models have relatively simple structures and low computational costs. The neural network, random forest, and xgboost are machine-learning-based models capable of describing nonlinear relationships. The structures of them are relatively complex compared to linear models. In addition, by using big data calculation frameworks (such as Spark), the calculation efficiency of these models can be prominently elevated [39].

4. Results and discussion

4.1. Large-scale predictions for EV charging energy

In this section, we examine the performance of the proposed framework for the large-scale prediction of real-world EV charging energy. The number of EVs utilized for the assessments exceeded 14,800 and that of the charging sessions was approximately 4.7 million. To better evaluate the prediction results from a practical perspective, relative errors and mean absolute percentage errors (MAPEs) were considered. Accordingly, we first assess the accuracy and stability of the prediction of EV charging energy under different seasons. Subsequently, we compared the large-scale prediction performance of the proposed model with that of several existing high-performance prediction models.

Fig. 5 depicts the relative error distributions for the prediction results under different time scales (annual, spring and autumn, summer, and winter). The MAPEs and the standard deviations (σ) of the four models, i.e., the proposed model, linear model, random forest model, and proportion model, are compared under each time frame. It should be noted that the proportion model is a linear regression model only requiring the input of SOC variations. From a holistic perspective, under all the time frames (Fig. 5A, B, C, and D), the proposed model achieves the lowest relative errors and standard deviations, as compared with the other models; the MAPE and σ of the proposed model are lower than those of the other models by 14.7–39.4% and 12.6–32.8%, respectively. During the summer and winter (see Fig. 5C and D), the centers of the relative error distributions for the proportion model are evidently different from those of the other models. These differences could be attributed to its limited consideration of ambient temperatures. Moreover, all the prediction models suffer from precision losses during the winter; highest values for the MAPEs and standard deviations were observed during winter. Although the low temperatures during winter increase the difficulty of accurately predicting the charging energy, the MAPE of the proposed model is 16.0–39.4% lower than those of the other models.

For the large-scale prediction of EV charging energy, it is necessary to maintain high model performances to improve the average prediction accuracy, despite the significant variations in the ratio of the number of EVs used for testing and training (referred to as the testing/training ratio). In this context, real-world operating data of several light-duty EVs classified as the same vehicle type are utilized. Fig. 6A and B depict the accuracy of various prediction models with a testing/training ratio ranging from 0.1 to 1000. It can be seen that variations in the results of these models are of three types, as indicated by the different line types in Fig. 6. The first type comprises linear models, i.e., the proportion model and the linear regression model; the relative errors in the results of these models remain fairly stable (the MAPEs are in the range of 4.4–5.0%) regardless of the varying ratio. The second type consists of several widely utilized nonlinear models, i.e., the xgboost, random forest, and neural network models. These models exhibit low relative errors when the testing/training ratio is low (the MAPEs are approximately 3% when the testing/training ratio varies from 0.1 to 1). However, their performance notably decreases when the testing/training ratio is high (the MAPEs are in the range of 6.2–10.6% when the testing/training

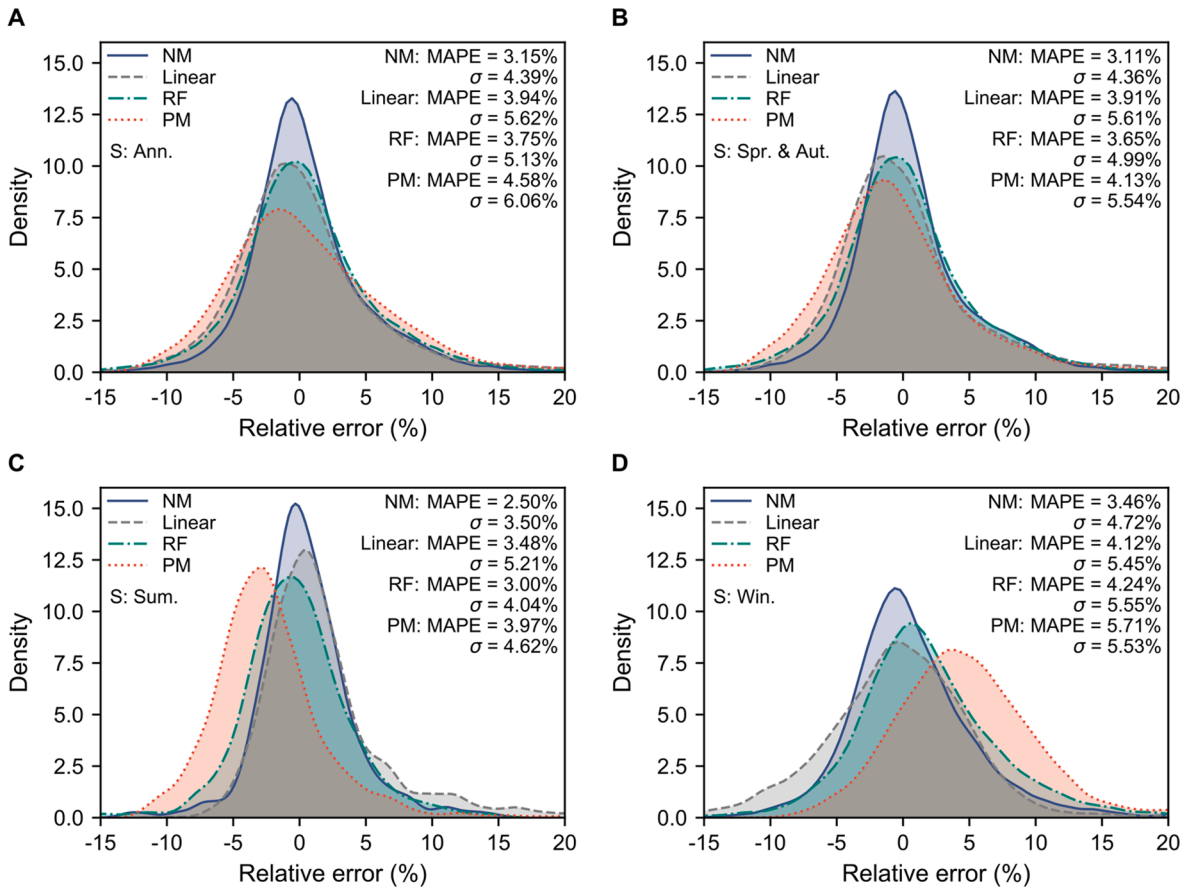

Fig. 5. Predictions during different seasons. (A), (B), (C), and (D) are the relative error distributions for the EV charging energy prediction results of the entire year, spring and autumn, summer, and winter, respectively. The MAPEs and standard deviations (σ) are used to reflect the accuracy and stability of the prediction results.

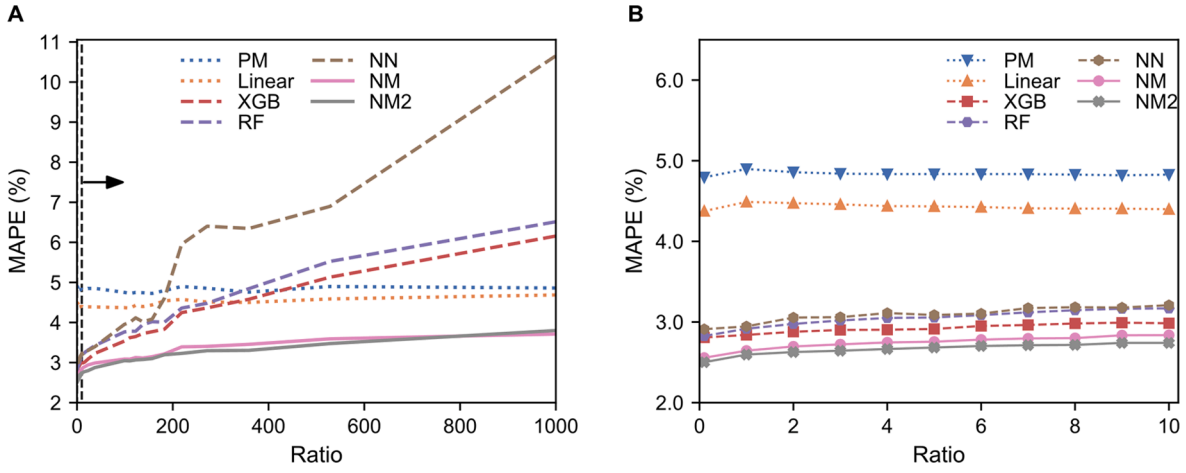

Fig. 6. Performance comparison from the perspective of big data. (A) and (B) are the MAPEs of the large-scale EV charging energy predictions when the testing/training ratio varies from 0.1–1000 and 0.1–10, respectively.

ratio is 1000). The third type includes two novel models wherein the proposed data-driven framework is employed. The difference between these two models is that one model (NM) does not employ the continuity analysis (classifications for charging rates are described in Section 3.2), whereas the other model (NM2) does. The MAPEs of these two models remain low (2.5–3.8%) when the testing/training ratio varies from 0.1 to 1000. In this case, the highest and lowest MAPEs for model NM2 are lower than those of the xgboost model by 38.3% and 9.7% and those of the linear model by 18.9% and 41.9%, respectively. Overall, the relative errors of these two proposed novel models are significantly lower than those of the other models for any testing/training ratio between 0.1 and 1000; the lowest MAPEs were observed for model NM2. The recommendations for models in different scenarios are shown in Table 3. It can be seen that linear and nonlinear models, respectively, are recommended when the testing/training ratio is high and low; the novel model is recommended when both high and low testing/training ratios prevail.

4.2. Comparison of performance on replacing core algorithms

The proposed prediction framework consists of linear and nonlinear predictions. In this case, linear models with different regularization (such as linear, lasso, and ridge regression models) and tree-based models with different ensemble methods (such as random forest, gradient boosting, and xgboost models) can be applied in this data-driven framework. Fig. 7 presents the highest and lowest MAPEs of the nine combinations of different core algorithms with testing/training ratios ranging from 0.1 to 1000. Overall, the highest and lowest MAPEs of these combinations are obtained when the testing/training ratio is equal to the highest and lowest values, respectively.

The lowest and highest values are obtained when the proportion of training data is the highest and lowest, respectively. The combination of the linear and random forest regression (L0-RF) models achieves the lowest prediction errors at both the highest and lowest MAPEs with the testing/training ratio varying from 0.1 to 1000. Interestingly, the models with the same nonlinear algorithms yielded the same lowest values, for a testing/training ratio of 0.1. A reasonable explanation is that a high proportion of training data enables a higher descriptive ability of the nonlinear algorithms, which decreases the effect of linear contributions. In addition, although the accuracy of these combinations is different, their highest and lowest MAPEs are below 2.8% and 4.2%, respectively; therefore, they outperform existing, commonly used models such as neural network and xgboost models. This can be elucidated by a comparison of the results in Figs. 6 and 7.

4.3. Model discussion

In this section, modeling adaptability, modeling limitations, and prospects of real-world scaling up are discussed. For adaptability, as illustrated previously, the proposed prediction model is established based on a data-driven method that individually controls linear and nonlinear contributions. This foundation makes the proposed model particularly adaptable in two aspects; (1) the proposed model can be compatible with varied vehicle types because the input data only focuses on SOC variations, time length, and environmental factors, and (2) the proposed prediction model maintains high accuracy over a wide range of testing/training ratios, which enables it to be adapted to EV groups with different sizes. These two features of adaptability make the proposed model advantageous for large-scale predictions.

There are limitations in modeling. There is a fact that errors in datasets could jeopardize the accuracy of the correlation analysis [40]. Moreover, the influence of data errors on correlation analyses and parameter optimization would be exacerbated in large-scale datasets [41]. As shown in Fig. 2, the correlation calculations are incorporated into the prediction modeling. The challenges to ensure the quality and balance of targeted EV data would increase accordingly. In this context, both a high-performance fault diagnosis model and a data quality control model are required before implementing the proposed prediction model. Another shortcoming is that the prediction model may not obtain accurate charging energy of an EV when the SOC variation is extremely small (such as 1–3%). There are many reasons for this, such as unstable charging power at the beginning, sensor errors, etc.

The proposed prediction model can be extended to multiple future applications. Firstly, the prediction of EV charging energy is an indispensable input for energy/power demand calculations [42]. Secondly, by combining energy/power demands and spatiotemporal information, the proposed model can be utilized for assessing impacts on the grid [43] and infrastructure planning [10]. Moreover, the charging energy of EVs is also a bridge between EV operation and the impacts on the economy and environment. In this case, the proposed model can be employed to analyze greenhouse gas reductions [44], monetary benefits [45], and relevant policies [46]. In addition, the rising scale of EV applications signifies a need for more novel models that can be used in the big data context [14]. The proposed model can be utilized in city-level or nationwide analyses of EV charging energy, which opens avenues for future large-scale energy management of EVs.

Table 3
Model recommendations.

Model	High testing/training ratio	Low testing/training ratio
Proportion	+	−−−
Linear	++	−−
Xgboost	−−	++
Random forest	−−	+
Neural network (BP)	−−−	+
Novel model	+++	+++

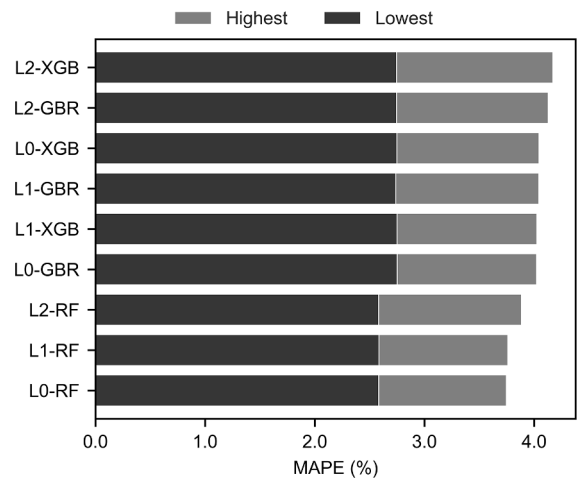

Fig. 7. Comparison of model performance on replacing core algorithms. The testing/training ratio varies from 0.1 to 1000; combinations of six models are employed: the linear model (L0), lasso model (L1), ridge model (L2), xgboost model (XGB), random forest model (RF), and gradient boosting model (GBR).

5. Conclusions

Existing models face difficulties in maintaining high accuracies during the large-scale prediction of EV charging energy; this study addresses this issue through the development of a novel data-driven prediction framework. The primary contributions of the proposed framework are that (1) it improves the adaptability of the charging energy prediction model when facing EV groups with varying sizes, and that (2) it improves the prediction accuracy in complex real-world operating conditions. The way to achieve these goals is that we propose a novel data-driven method that individually controls linear and nonlinear contributions in prediction modeling. The results indicate that the MAPEs of the proposed model vary from 2.5% to 3.8% when the testing/training ratio ranges from 0.1 to 1000. To validate the performance of the proposed prediction model, substantial real-world EV data and five existing high-performance prediction models (PM, Linear, XGB, RF, and NN) are employed. The comparison results illustrate that even in the case of a high testing/training ratio, the MAPEs of the proposed framework are lower than those of these existing models by at least 18.9%. As the charging energy of an EV correlates EV utilization and the power supply, the models and framework presented herein can be widely utilized for future research on EV energy demands, charging planning, impacts on urban power grids, and policymaking. Moreover, by combining data on electricity generation, the proposed models can also be extended for assessing emission reduction benefits, impacts on air pollution, and health benefits of urban EVs.

CRediT authorship contribution statement

Yang Zhao: Conceptualization, Methodology, Software, Writing - original draft. **Zhenpo Wang:** Resources, Data curation. **Zuo-Jun Max Shen:** Supervision. **Fengchun Sun:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (2019YFB1600800).

References

[1] Liang X, Zhang S, Wu Y, Xing J, He X, Zhang KM, et al. Air quality and health benefits from fleet electrification in China. Nat Sustainability 2019;2:962–71.
[2] Global EV Outlook 2020. (International Energy Agency, 2020).
[3] Chen YW, Chang JM. Fair demand response with electric vehicles for the cloud based energy management service. IEEE Trans Smart Grid 2016;9:458–68.
[4] Atif Y, Ding J, Jeusfeld MA. Internet of things approach to cloud-based smart car parking. Procedia Comput Sci 2016;98:193–8.
[5] U.S. Department of Energy's Office of Energy Efficiency & Renewable Energy. Alternative Fuels Data Center, https://afdc.energy.gov; [accessed 23 May 2020].
[6] Albert Ahdoot. How Big Data Drives Tesla. https://www.colocationamerica.com/blog/how-big-data-drives-tesla; [accessed 23 May 2020].
[7] Muratori M. Impact of uncoordinated plug-in electric vehicle charging on residential power demand. Nat Energy 2018;3:193–201.
[8] Collin R, Miao Y, Yokochi A, Enjeti P, von Jouanne A. Advanced electric vehicle fast-charging technologies. Energies 2019;12:1839.
[9] Mehrjerdi H, Hemmati R. Stochastic model for electric vehicle charging station integrated with wind energy. Sustain Energy Technol Assess 2020;37:100577.
[10] Micari S, Polimeni A, Napoli G, Andaloro L, Antonucci V. Electric vehicle charging infrastructure planning in a road network. Renew Sustain Energy Rev 2017;80:98–108.
[11] Hardman S, Jenn A, Tal G, Axsen J, Beard G, Daina N, et al. A review of consumer preferences of and interactions with electric vehicle charging infrastructure. Transport Res Part D: Trans Environ 2018;62:508–23.
[12] Mehrjerdi H, Hemmati R. Electric vehicle charging station with multilevel charging infrastructure and hybrid solar-battery-diesel generation incorporating comfort of drivers. J Storage Mater 2019;26:100924.
[13] Domínguez-Navarro JA, Dufo-López R, Yusta-Loyo JM, Artal-Sevil JS, Bernal-Agustín JL. Design of an electric vehicle fast-charging station with integration of renewable energy and storage systems. Int J Electr Power Energy Syst 2019;105:46–58.
[14] Arias MB, Bae S. Electric vehicle charging demand forecasting model based on big data technologies. Appl Energy 2016;183:327–39.
[15] Robledo CB, Oldenbroek V, Abbruzzese F, van Wijk AJ. Integrating a hydrogen fuel cell electric vehicle with vehicle-to-grid technology, photovoltaic power and a residential building. Appl Energy 2018;215:615–29.
[16] Hannan MA, Hoque MM, Hussain A, Yusof Y, Ker PJ. State-of-the-art and energy management system of lithium-ion batteries in electric vehicle applications: Issues and recommendations. IEEE Access 2018;6:19362–78.
[17] Wu Y, Ravey A, Chrenko D, Miraoui A. Demand side energy management of EV charging stations by approximate dynamic programming. Energy Convers Manage 2019;196:878–90.
[18] Xu Y, Çolak S, Kara EC, Moura SJ, González MC. Planning for electric vehicle needs by coupling charging profiles with urban mobility. Nat Energy 2018;3:484–93.
[19] Li W, Zhu J, Xia Y, Gorji MB, Wierzbicki T. Data-driven safety envelope of lithium-ion batteries for electric vehicles. Joule 2019;3:2703–15.
[20] Luo Y, Zhu T, Wan S, Zhang S, Li K. Optimal charging scheduling for large-scale EV (electric vehicle) deployment based on the interaction of the smart-grid and intelligent-transport systems. Energy 2016;97:359–68.
[21] Li B, Kisacikoglu MC, Liu C, Singh N, Erol-Kantarci M. Big data analytics for electric vehicle integration in green smart cities. IEEE Commun Mag 2017;55:19–25.
[22] Zhao Y, Liu P, Wang Z, Zhang L, Hong J. Fault and defect diagnosis of battery for electric vehicles based on big data analysis methods. Appl Energy 2017;207:354–62.
[23] Liu G, Ouyang M, Lu L, Li J, Hua J. A highly accurate predictive-adaptive method for lithium-ion battery remaining discharge energy prediction in electric vehicle applications. Appl Energy 2015;149:297–314.
[24] Dong G, Chen Z, Wei J, Zhang C, Wang P. An online model-based method for state of energy estimation of lithium-ion batteries using dual filters. J Power Sources 2016;301:277–86.
[25] Zhang Y, Xiong R, He H, Shen W. Lithium-ion battery pack state of charge and state of energy estimation algorithms using a hardware-in-the-loop validation. IEEE Trans Power Electron 2016;32:4421–31.
[26] Dong G, Zhang X, Zhang C, Chen Z. A method for state of energy estimation of lithium-ion batteries based on neural network model. Energy 2015;90:879–88.
[27] Zhang C, Jiang J, Gao Y, Zhang W, Liu Q, Hu X. Charging optimization in lithium-ion batteries based on temperature rise and charge time. Appl Energy 2017;194:569–77.
[28] Ma S, Jiang M, Tao P, Song C, Wu J, Wang J, et al. Temperature effect and thermal impact in lithium-ion batteries: A review. Progr Nat Sci: Mater Int 2018;28:653–66.
[29] Pandžić H, Bobanac V. An accurate charging model of battery energy storage. IEEE Trans Power Syst 2018;34:1416–26.
[30] Lindgren J, Lund PD. Effect of extreme temperatures on battery charging and performance of electric vehicles. J Power Sources 2016;328:37–45.
[31] Ranka S, Singh V. CLOUDS: A decision tree classifier for large datasets. In: Proceedings of the 4th Knowledge Discovery and Data Mining Conference; 1998; 2:8.
[32] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data Knowl Eng 2007;60:208–21.
[33] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science 2011;334:1518–24.
[34] Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proc Natl Acad Sci 2014;111:3354–9.
[35] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2012.
[36] Specht DF. A general regression neural network. IEEE Trans Neural Networks 1991;2:568–76.
[37] Liaw A, Wiener M. Classification and regression by randomForest. R news 2002;2:18–22.
[38] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016: 785–794.
[39] Han Z, Zhang Y. Spark: A big data processing platform based on memory computing. In: 2015 Seventh International Symposium on Parallel Architectures, Algorithms and Programming (PAAP) IEEE; 2015.
[40] Lai CS, Tao Y, Xu F, Ng WW, Jia Y, Yuan H, et al. A robust correlation analysis framework for imbalanced and dichotomous data with uncertainty. Inf Sci 2019;470:58–77.
[41] Wang K, Xu C, Zhang Y, Guo S, Zomaya AY. Robust big data analytics for electricity price forecasting in the smart grid. IEEE Trans Big Data 2017;5:34–45.
[42] Kisacikoglu MC, Erden F, Erdogan N. Distributed control of PEV charging based on energy demand forecast. IEEE Trans Ind Inf 2017;14:332–41.
[43] Deilami S, Masoum AS, Moses PS, Masoum MA. Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile. IEEE Trans Smart Grid 2011;2:456–67.
[44] Foley A, Tyther B, Calnan P, Gallachóir BÓ. Impacts of electric vehicle charging under electricity market operations. Appl Energy 2013;101:93–102.
[45] Xiang Y, Liu J, Li R, Li F, Gu C, Tang S. Economic planning of electric vehicle charging stations considering traffic constraints and load profile templates. Appl Energy 2016;178:647–59.
[46] Ji Z, Huang X. Plug-in electric vehicle charging infrastructure deployment of China towards 2020: Policies, methodologies, and challenges. Renew Sustain Energy Rev 2018;90:710–27.