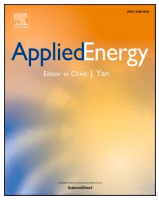
Applied Energy 339 (2023) 120954

Contents lists available at ScienceDirect

Applied Energy

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



Prognostics of battery capacity based on charging data and data-driven methods for on-road vehicles

Zhongwei Deng [a], Le Xu [b,*], Hongao Liu [c], Xiaosong Hu [c], Zhixuan Duan [d], Yu Xu [d]

[a] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] School of Sustainability, Stanford University, Stanford, CA 94305, USA
[c] College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, China
[d] DAO Smart Energy Inc., Hefei 230088, China

HIGHLIGHTS

- A labeled capacity acquisition method is proposed for on-road electric vehicles.
- A general feature engineering method is constructed to obtain optimal feature set.
- Two residual models are used to compensate prediction errors of battery capacity.
- The method is verified by the data of 20 electric vehicles running about 29 months.

ARTICLE INFO

Keywords:
Lithium-ion battery
Electric vehicles
Capacity prediction
Feature extraction
Seqence-to-sequence method
Gaussian process regression

ABSTRACT

The large-scale application of lithium-ion batteries makes it urgent to accurately predict their capacity degradation so as to achieve timely maintenance and second-life utilization. For on-road electric vehicles (EVs), due to limitation of battery management system in measurement and computing power, it is still a tricky challenge to accurately predict the capacity of battery pack. To this end, a battery capacity prognostic method based on charging data and data-driven algorithms is proposed in this paper. First, battery capacity is calculated based on a variant of Ampere integral formula, and statistical values of the capacity during a month are regarded as labeled capacity to reduce errors. Then, statistical characteristics of battery charging data are extracted, and correlation analysis and feature selection are conducted to determine optimal feature sets. Moreover, a sequence-to-sequence (Seq2Seq) model is employed to predict future capacity trajectory, and two residual models based on Gaussian process regression (GPR) are proposed to compensate the prediction error caused by local capacity change. Finally, the data of 20 EVs operating about 29 months are used to verify the proposed methods. By using the first 3 months data as input, the remaining capacity sequence can be accurately predicted with error lower than 1.6%.

1. Introduction

Lithium-ion batteries have been widely used in electric vehicles (EVs) to reduce dependence on fossil fuels and achieve the vision of carbon neutrality in transportation sector [1]. In the next decades, it is expected that the application scale of EVs will skyrocket to replace the dominance of traditional cars [2]. For limited lithium resources, it is indispensable to realize efficient use and recycling of lithium batteries to ensure the sustainable development of EVs [3]. Moreover, as a key and the most expensive part of EVs, it is of significance to achieve timely maintenance, residual value assessment and second-life utilization of battery system. Among them, battery degradation prognostic is one of the key and generic techniques [4,5], which needs to be resolved urgently.

For lithium-ion batteries, their performance inevitably declines over time during use or storage due to the natural property of electrochemical devices. The intuitive manifestation of battery degradation is the decrease in capacity and increase in internal resistance. The battery internal resistance can be obtained by several mature methods, such as hybrid pulse power characterization (HPPC) test [6], and parameter

* Corresponding author.
E-mail address: lexu1209@stanford.edu (L. Xu).

https://doi.org/10.1016/j.apenergy.2023.120954
Received 6 January 2023; Received in revised form 21 February 2023; Accepted 5 March 2023
Available online 20 March 2023

Table 1
Items of battery charging data.

Items	Unit	Resolution
Time	yyyy-mm-dd hh:mm:ss	–
Current	A	
Pack voltage	V	0.1 V
SOC	–	0.1
Maximum cell voltage	V	0.001
Minimum cell voltage	V	0.001
Maximum cell temperature	℃	1℃
Minimum cell temperature	℃	1℃

identification of equivalent circuit models (ECMs) [7]. However, the calculation of battery capacity depends on a complete charging and discharging process, which is heavily affected by temperature and current rate. For real applications, more attention is paid to the battery capacity degradation, because the capacity directly determines how much energy the battery can store and release. In recent years, numerous studies [8] have been carried out to predict battery capacity degradation, and can be roughly divided into model-based and data-driven methods.

The model-based methods need to establish a mathematical model to describe the degradation of battery capacity, typically including empirical models and electrochemical models. The empirical models can be acquired by fitting the capacity loss with exponential or polynomial functions [9], and then extrapolating the battery capacity based on calendar time, cycle number or mileage. Usually, a large number of orthogonal experiments are required to consider the effects of different influencing factors. For specific applications, the model parameters can be updated by using Kalman filter or particle filter [10]. Due to the lack of physical meanings, the empirical models established by using experimental data are difficult to apply to practical scenarios. In contrast, electrochemical model with side reactions can describe the internal degradation processes of battery at micro level, and typical degradation mechanisms include the growth of solid electrolyte interface (SEI) [11] and lithium plating [12]. After clarifying the degradation mechanism of the battery under study, the future capacity degradation of batteries can be derived by simulation [13]. In the literature, extensive work has been devoted to simplify the electrochemical models [14] and identify model parameters [15]. However, it is difficult to figure out and quantify the complicated degradation processes due to the mutual coupling of side reactions inside battery. In addition, with the aging of the battery, the parameters variation of the electrochemical models will cause a large error in capacity prediction.

In recent years, with the generation and accumulation of a large amount of battery data and the penetration of artificial intelligence technology, data-driven methods have been widely used in battery capacity prognostics [16]. Without any prior knowledge of battery dynamics, a nonlinear model can be established to map the relationship between battery features and degradation state (battery capacity or remaining useful life). For data-driven methods, the key is to develop advanced algorithms and feature extraction methods for specific applications. Both lightweight and deep learning machine learning algorithms have been successfully applied in battery degradation prognostics. Usually, when features are well extracted, the lightweight algorithms are employed to build the prediction model, such as multiple linear regression [5], support vector regression (SVR) [17], relevance vector machine [18], and Gaussian process regression (GPR) [19]. In contrast, with the ability of automatic feature extraction, deep learning algorithms can directly use the raw time-series data as input, typical algorithms including long short-term memory network (LSTM) [20] and deep convolutional neural network (DCNN) [21]. For feature engineering, increment capacity (IC) [22] and differential voltage (DV) [23] curves have been applied to analyze the degradation mechanisms of battery, and the characteristics of the curves such as peak, valley and area values can be used as features to predict battery capacity. To obtain IC/DV curves, the battery usually needs to be charged or discharged at a small current rate, and lots of data noise could be caused by differentiation operation. Further studies [24,25] indicate that effective features can also be extracted from the charge or discharge capacity sequence without the differentiation operation. In addition to extracting features from battery measurement data, Xu et al. [26] obtain physics-based features through identifying the parameters of electrochemical model, and a hybrid feature is formed by multiplying the data-based and physics-based features. Then the sequence-to-sequence (Seq2Seq) model combined with battery degradation classification is employed to predict future capacity trajectory. However, this study just focusses on battery cell under laboratory test. To make feature extraction more general, Zhang et al. [27] extract the statistic properties of battery data histograms as features to learn battery aging and develop global models and individualized models to achieve online prediction of battery aging trajectory and lifetime. Their method has been validated under three large data sets, including real-world fleet data of over 7000 plug-in hybrid EVs. For battery capacity prognostics based on data-driven methods, labeled capacity need to be acquired to train the prediction models. For practical applications, many researches [28,29] obtain the labeled capacity by a variant of Ampere integral formula which calculates the capacity as the ratio of accumulated capacity (Q) to state of charge (SOC) interval, but the inevitable estimation error of SOC could cause a large error in battery capacity. Although a large number of battery capacity prediction models have been established with outstanding performance for specific application scenarios, there are a few studies suitable for battery system of on-road vehicles.

To fill this research gap, a capacity prognostic method of battery pack for on-road vehicles is proposed in this study. First, battery capacity is calculated based on the variant of Ampere integral formula, and the statistical values of calculated capacity during a month are regarded as labeled capacity to reduce errors. Then, the statistical characteristics of battery charging data are extracted, and correlation analysis and feature selection is conducted to determine optimal feature sets. Moreover, the Seq2Seq model is employed to predict the future capacity trajectory, and two residual models are proposed to compensate the prediction error caused by local capacity change. Finally, the data set of 20 EVs operating about 29 months are used to verify the proposed methods. The main contributions of this paper include,

(1) A labeled capacity acquisition method is proposed for on-road vehicles. For battery capacity calculated by $Q/\triangle$SOC, the statistical mean or median values during a month are used to reduce the effect of SOC error and data noise.
(2) A general feature engineering method is established to obtain optimal feature sets from battery charging data. The statistical characteristics (mean, sum, and standard deviation values) of battery charging data are extracted, and then selected by two correlaion analysis methods.
(3) The Seq2Seq model combined with GPR-based residual model is employed to obtain accurate battery capacity prediction results. The Seq2Seq model can effectively capture the whole degradation trend of battery capacity, while the GPR model can compensate the local capacity change.

The remaining parts of this paper are arranged as follows: Section 2 introduces the battery field data and labeled capacity acquisition method; the feature engineering is described in Section 3; the capacity prediction based on the Seq2Seq and GPR models are illustrated in Section 4; the results of the battery capacity prediction are presented in Section5; Section 6 concludes this paper.

2. Battery field data

The charging data of battery packs for 20 commercial EVs are employed in this study, and the time span of vehicle operation is all over

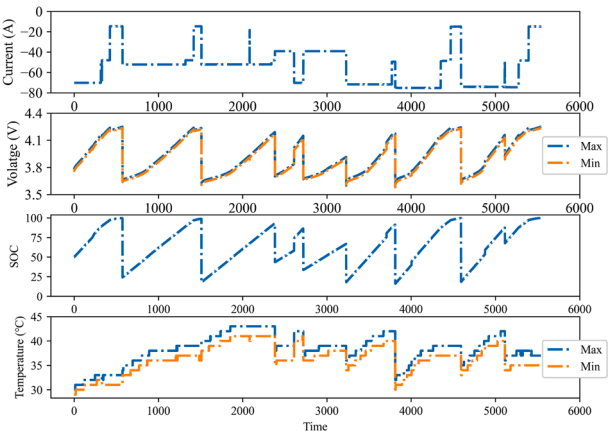


Fig. 1. The charging curves of a vehicle.

two years (around 29 months). These vehicles have the same battery system and are numbered as #1, #2, …, #20 in this study. The data are collected though charging devices, which receive the battery charging data via control area network (CAN) communication during the charging process. The recoding frequency of the charging data is 8 s. Table 1 lists the main items of battery charging data related to battery health evaluation, and the resolution of these data is also revealed. Given the limitation of data transmission, the resolution of real vehicle data is lower than that of laboratory testing data.

The charging curves of a vehicle are illustrated in Fig. 1, in which several charging processes within a certain time are spliced together, and the maximum and minimum cell voltages and temperatures are given. It can be seen that the vehicle uses a multi-stage constant current charging strategy, but the specific current value at each stage is adjusted according to battery temperature.

2.1. Labeled capacity acquisition

To predict the capacity degradation trajectory of battery pack, a large number of the labeled capacity need to be calculated in advance. Based on the available battery data, a variant of Ampere integral formula is used to calculate battery capacity,

$$C_a = \frac{-\int_{t_1}^{t_2} \Delta t I(t)}{SOC_{t_2} - SOC_{t_1}} \tag{1}$$

where Δt is the fixed sampling interval, I is the battery current with negative value for charging process, and t_1, t_2 are the start and end charging time. It is noted that Eq. (1) depends on high-precision estimation of battery SOC to acquire accurate battery capacity. Besides, a large SOC interval is also required to avoid calculation error. However, due to the lack of test data and key characteristics of batteries, it is much difficult to utilize advanced state estimation method to obtain accurate SOC with negligible error. To tackle this problem, the statistical mean values or median values of the calculated capacity among a period of time is employed to obtain the labeled capacity, which can effectively exclude the outliers caused by SOC error or data noise.

Fig. 2 presents the calculated capacity of battery pack for two EVs, in which the raw values and the statistical values are compared. As shown in Fig. 2(a), because of the SOC error, short SOC interval, data noise, etc., the calculated capacity based on Eq. (1) exist lots of outliers, and a large number of points fluctuate in the same month. According to statistics, on average, >90 capacity points can be obtained every month. For vehicle application, there is no need to pay attention to the battery capacity of each charging process, which is also hard to obtain

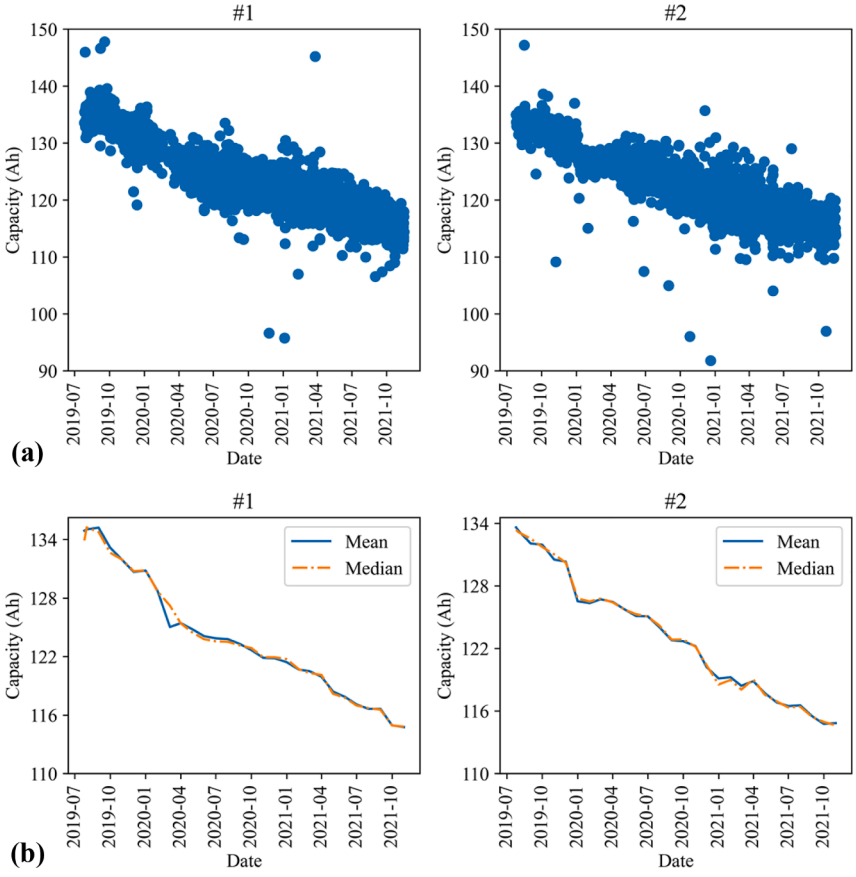


Fig. 2. The calculated capacity of battery pack for two EVs.

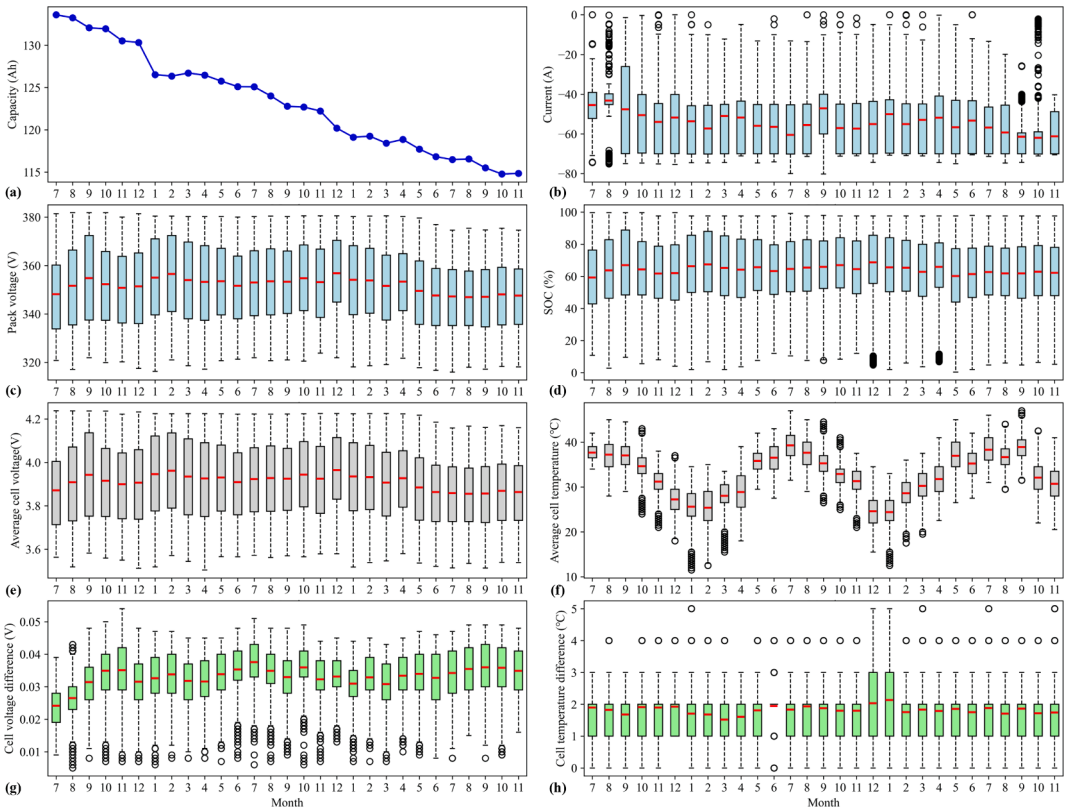


Fig. 3. The statistical characteristics of charging data of battery pack (take vehicle #2 as an example). On each box, the central mark indicates the **mean** value, and the bottom and top edges of the box represent the 25th and 75th percentiles, respectively. The whiskers extend at both ends to the most extreme data points, while the outliers are drawn separately using the '○' symbol. (a) battery capacity; (b) battery current; (c) pack voltage; (d) SOC; (e) average cell voltage; (f) average cell temperature; (g) cell voltage difference; (h) cell temperature difference.

accurately. As it is reasonable to monitor the capacity change monthly, the statistical mean and median values of the calculated capacity during each month are derived, as shown in Fig. 2(b). Battery capacity curves with obvious degradation trend are acquired through this operation. The mean values are almost equal to the median values, indicating the calculated capacity points during a month are symmetrically distributed. Both of them can be used to effectively represent the degradation state of the battery system. Besides, it can be observed that there are several local capacity recovery processes for these vehicle field data, which could be caused by different factors, such as, long-time rest and temperature change.

2.2. Charging data evolution

To determine the effect factors of battery degradation, the statistical characteristics of battery charging data, including battery current (I), pack voltage (V_{pack}), SOC, average cell voltage (V_{ave}), average cell temperature (T_{ave}), cell voltage difference (V_d), and cell temperature difference (T_d) are investigated, as shown in Fig. 3. The average cell voltage or temperature is the mean value of maximum and minimum cell voltage or temperature, while the difference of cell voltage or temperature is obtained by subtracting the maximum value from the minimum value. The voltage and temperature differences are utilized to reflect the effect of inconsistency between battery cells on battery pack capacity. It is noted that the box plots are drawn based on the data of each month. According to Fig. 3, as the battery ages, there is no obvious changing characteristics in battery current, voltage and SOC, so it is difficult to decide how these data affect battery aging. In contrast, battery temperature has an obvious effect on battery capacity. It can be found that battery temperature varies periodically with the calendar time. In the months of low temperatures, such as January, the battery capacity is also relatively smaller. Then, with the increase of temperature, the battery capacity is able to recover partially.

3. Features engineering

In order to establish a data-driven model to accurately predict the capacity degradation of battery pack, features highly related to battery capacity need to extracted from the large-scale field data in advance. In this section, the statistical characteristics of battery charging data are extracted, correlation analysis is conducted to get rid of ineffective features and a feature selection procedure is applied to obtain optimal feature sets.

3.1. Correlation analysis

Firstly, the statistical characteristics of all battery charging data are calculated over a month, including ***mean***, ***sum***, and ***standard deviation*** values. The commonly used two metrics is employed to evaluate the correlation between battery capacity and the features, namely Pearson correlation coefficient (PCC) [30] and gray relational grade (GRG) [31,32]. The former provides the direction and strength of the linear correlation, while the latter provides a quantitative measurement of system evolution and is considered to be suitable for dynamic processes. Generally, the higher the correlation between the input features and the target of a data-driven model, the better the accuracy of the model. The PCC is calculated as [30],

$$\rho_{x_i} = \frac{\sum(x_i - \bar{x}_i)(z - \bar{z})}{\sqrt{\sum (x_i - \bar{x}_i)^2 \sum (z - \bar{z})^2}} \tag{2}$$

where x_i is the i^{th} feature, z is the battery pack capacity, $\bar{x}_i$ and $\bar{z}$ are the

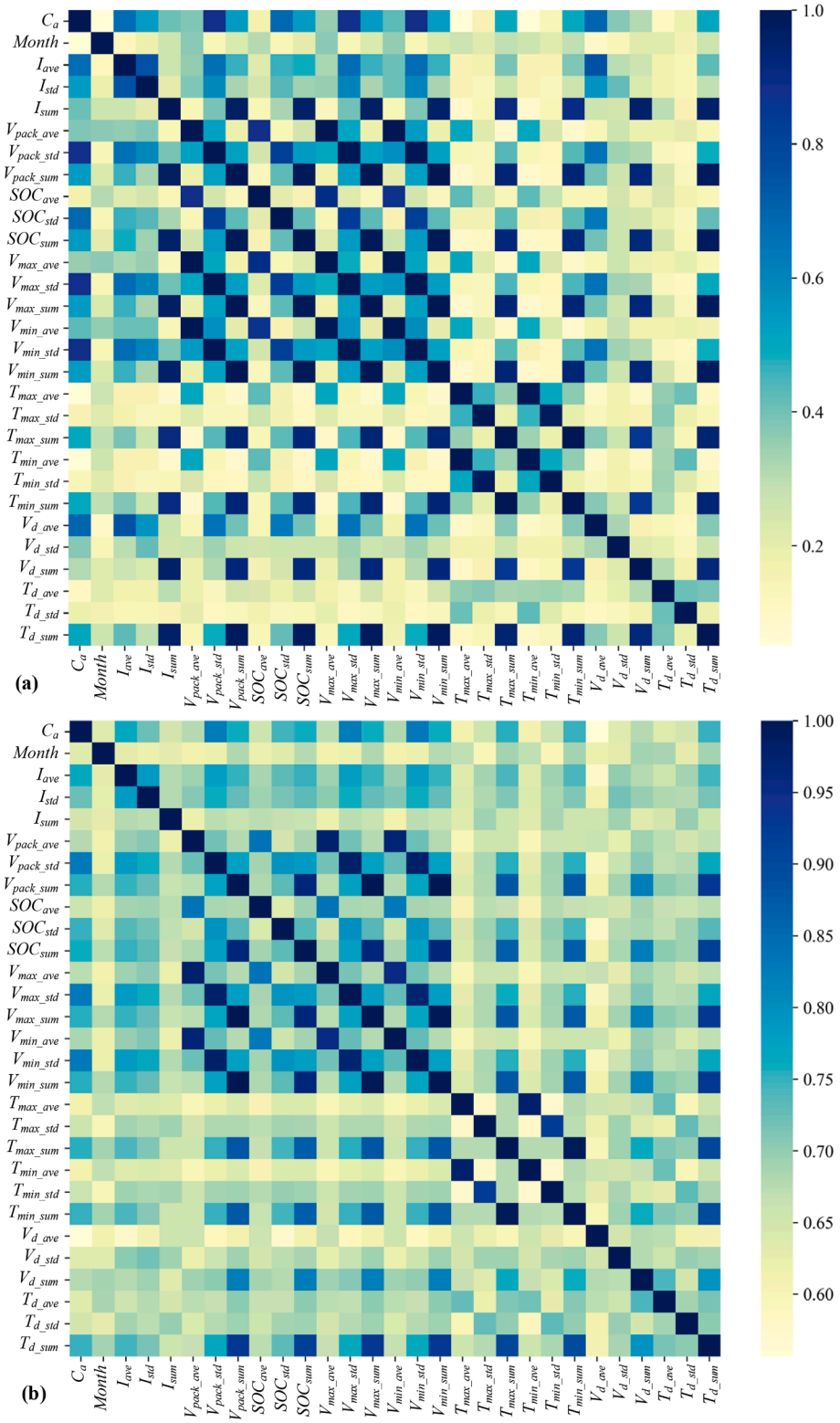


Fig. 4. Correlation analysis results between battery capacity and the features. The suffix "*_ave*", "*_sum*", and "*_std*" denote average value, sum value and standard deviation value of the variable, respectively. (a) Pearson correlation analysis; (b) gray relation analysis.

average values of the specific feature sequence and capacity sequence, respectively. In contrast, the GRG is obtained by,

$$r_i = \frac{1}{n}\sum_{k=1}^{n}\xi_i(k) \tag{3}$$

where $\xi_i(k)$ is the gray relation coefficient of the i^{th} feature at time k, and can be expressed as [31],

$$\xi_i(k) = \frac{\min\limits_i\min\limits_k|z(k)-x_i(k)| + \rho\min\limits_i\max\limits_k|z(k)-x_i(k)|}{|z(k)-x_i(k)| + \rho\min\limits_i\max\limits_k|z(k)-x_i(k)|} \tag{4}$$

Table 2
The selected feature sets.

PCC	I_{ave}	I_{std}	V_{pack_sum}	V_{pack_std}	SOC_{std}	T_{max_sum}	V_{d_ave}	T_{d_sum}
GRG	I_{ave}	I_{std}	V_{pack_sum}	V_{pack_std}	SOC_{std}	T_{max_sum}	V_{d_ave}	

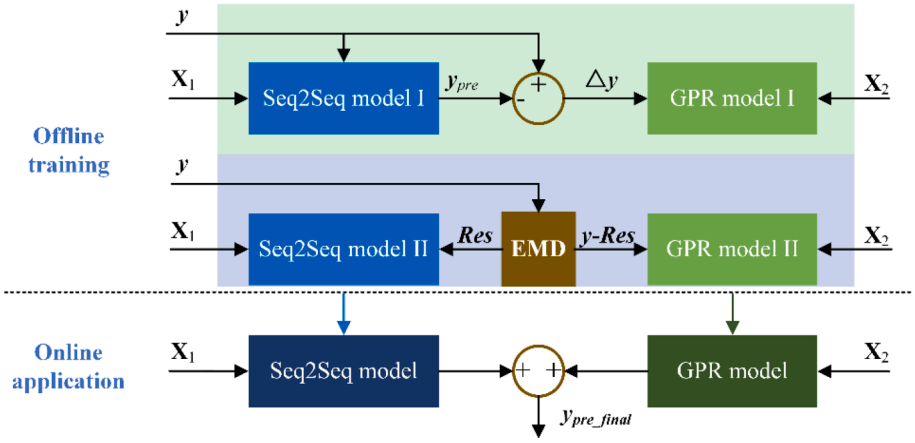


Fig. 5. The framework of battery capacity degradation prediction. y is the capacity sequence, X_1 is the feature input of Seq2Seq model, y_{pre} is the predicted capacity sequence of Seq2Seq model, $\triangle y$ is the predicted capacity error of Seq2Seq model, *Res* is the residual of capacity sequence obtained by empirical mode decomposition (EMD), X_2 is the feature input of GPR model, and y_{pre_final} is the finally predicted capacity sequence.

where ρ is the distinguishing coefficient and is usually set to 0.5.

By using the charging data of each vehicle, the PCC and GRG of different features can be calculated based on the above formulas. To obtain accurate correlation analysis results, the average values of the PCC and GRG for 20 vehicles are evaluated, as illustrated in Fig. 4. In addition to the charging data, the time denoted by month is also used as a potential feature, but its correlation to battery capacity is relatively weak. It can be observed that the determined features with high correlation to battery capacity are almost consistent for the two methods. Besides, some features have a high autocorrelation to other features, which means redundant information exist in them.

***Procedure* 1: Feature selection**

1. Calculate the correlation coefficient of all features for 20 vehicles
2. Get the average correlation coefficient (ρ_i) for each feature
3. Get feature f_i with $\rho_i >= 0.5$ or $r_i >= 0.8$, and rank from large to small
4. Construct F = [f_1, f_2, …, f_n]
5. For $i = 1, \ldots, n$
   For $j = i + 1, \ldots, n$
      If $\rho_{ij} > 0.9$ or $r_{ij} > 0.9$
         Delete f_j from F
      End
   End
End
6. Output F

3.2. Features selection

In order to select optimal feature sets from these potential features, a procedure is formulated, as described in Procedure 1. The main principle of feature selection is to obtain a feature set with high correlation to battery capacity and low autocorrelation between different features. In this study, a feature with correlation to battery capacity over a threshold (0.5 for the PCC and 0.8 for the GRG) is considered to be highly correlated, while if two features have a correlation coefficient over 0.9, they are regraded to be highly autocorrelated, one of them need to be abandoned to reduce redundancy. These threshold values are set according to the literature [28,33] and the results of correlation analysis. By using the PCC and GRG as the criteria, two optimal feature sets can be determined by the procedure, as listed in Table 2. Except for the parameter V_{d_ave} for the PCC, the selected features obtained based on the two methods are the same. Since the cell voltage difference is a critical factor that affects the capacity of battery pack, it is valuable to include the parameter V_{d_ave} in the feature set. Therefore, the feature set determined by the PCC is employed to establish battery capacity prediction model in the later study.

4. Battery capactiy degradation prediction

The proposed framework of battery capacity degradation prediction is shown in Fig. 5, including offline training and online application processes. The Seq2Seq models based on LSTM network is employed to predict the future capacity degradation, and two residual models based on GPR are used to compensate the prediction error caused by the change of month and temperature.

4.1. Seq2Seq method

The capacity trajectory is a time series, sampled by month in this paper. The future capacity degradation trajectory is forecasted by the early capacity sequence and features data, which is a typical sequence-to-sequence prediction problem. Therefore, the Seq2Seq deep learning network is employed to solve this problem, which has been widely used to solve sequence-to-sequence prediction problems, such as natural language processing. The structure of Seq2Seq network generally consists of an encoder, a decoder and a semantic vector connecting them [33]. The encoder reads the input sequence and compresses it into a fixed-length vector, namely the semantic vector, and then the semantic vector is decoded by the decoder to generate the final output sequence. The encoder and decoder are usually achieved by the RNN to learn time series characteristics. In this study, the long short-term memory (LSTM) is used to construct the encoder and decoder, which can avoid the problem of gradient vanishing or explosion in the training process of the standard RNN.

For the capacity trajectory predication based on the Seq2Seq model, the input and output sequences are formed as,

Table 3
The input and output of the models.

Model	Input	Output
Seq2Seq I	$\mathbf{X}_1$	y
Seq2Seq II	$\mathbf{X}_1$	Res
GPR I	$\mathbf{X}_2$	$\triangle y$
GPR II	$\mathbf{X}_2$	y-Res

$$D = [input|output] = \left[\begin{array}{cccc|c} c_1 & f_{11} & \cdots & f_{m1} & c_{n+1} \\ c_2 & f_{12} & \cdots & f_{m2} & c_{n+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_n & f_{1n} & \cdots & f_{mn} & c_{n+p} \end{array}\right] \quad (5)$$

where c is the capacity sequence, f_i is the i^{th} feature, n is the length of input sequence, m is the number of features (equal to 8 in this study), and p is the length of prediction sequence.

4.2. GPR-based residual models

For vehicle application, the battery capacity is heavily affected by calendar time and temperature, which could cause local recovery or sharp decrease in the capacity degradation curves. As illustrated in Fig. 3, the battery capacity decreases fast at the months of low temperature and recover steadily with the increase of temperature, and the average temperature of battery cell varies cyclically with calendar time. For the Seq2Seq model, it can effectively represent the entire degradation trend of battery capacity, but it is hard to capture the local change caused by calendar time and temperature. To enhance the prediction

Table 4
The statistical error of battery capacity estimation.

	Method	MAE	RMSE
Training	SVR	0.96%	1.36%
	GPR	0.97%	1.28%
Test	SVR	1.16%	1.55%
	GPR	1.07%	1.41%

Table 5
The statistical errors of battery capacity prediction with different feature sets.

Feature set	Process	MAE_*ave*	RMSE_*ave*
F1 = [I_{ave}, I_{std}, V_{pack_sum}, V_{pack_std}, SOC_{std}, T_{max_sum}, V_{d_ave}, T_{d_sum}]	Training	1.18%	1.59%
	Testing	1.34%	1.66%
F2 = [I_{ave}, I_{std}, V_{pack_sum}, V_{pack_std}, SOC_{std}, SOC_{sum}, V_{max_std}, V_{max_sum}, V_{min_std}, V_{min_sum}, T_{max_sum}, V_{d_ave}, T_{d_sum}]	Training	1.20%	1.63%
	Testing	1.32%	1.66%
F3 = [I_{ave}, V_{pack_std}, SOC_{std}, T_{max_sum}, V_{d_ave}, T_{d_sum}]	Training	1.25%	1.75%
	Testing	1.43%	1.87%

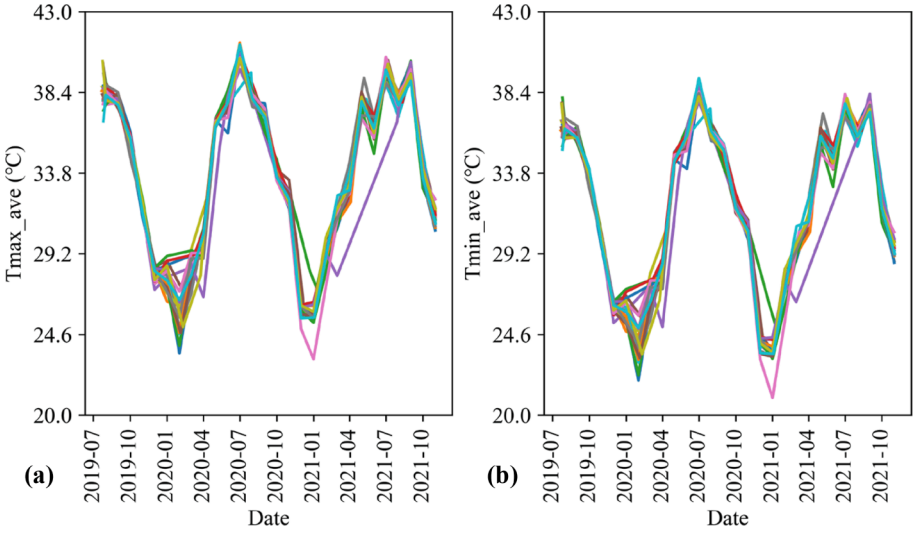


Fig. 6. The battery temperature evolution of different vehicles. Each color curve represents a car. (a) maximum cell temperature; (b) minimum cell temperature.

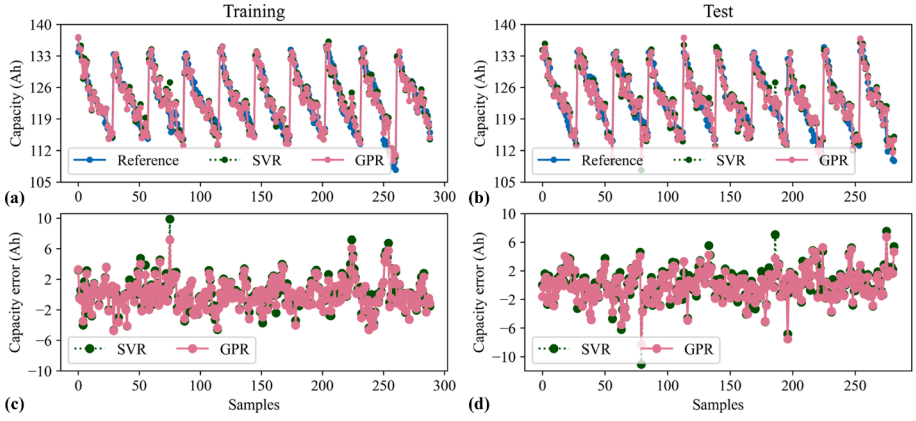


Fig. 7. Battery capacity estimation results. (a) training results; (b) test results; (c) training errors; (d) test errors.

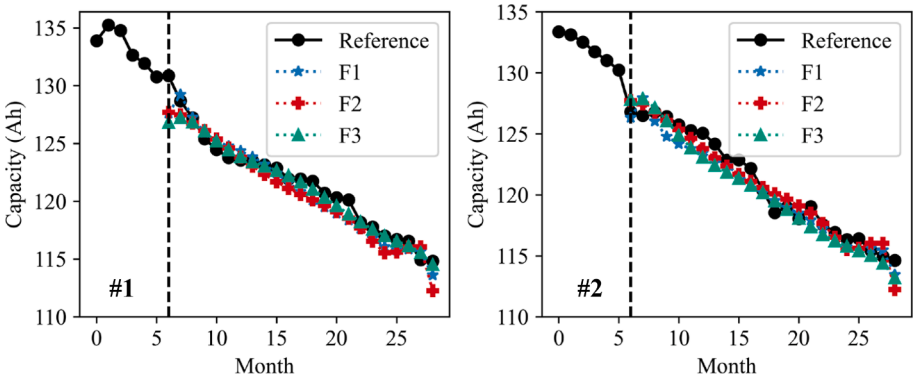


Fig. 8. The results of battery capacity prediction based on different feature sets.

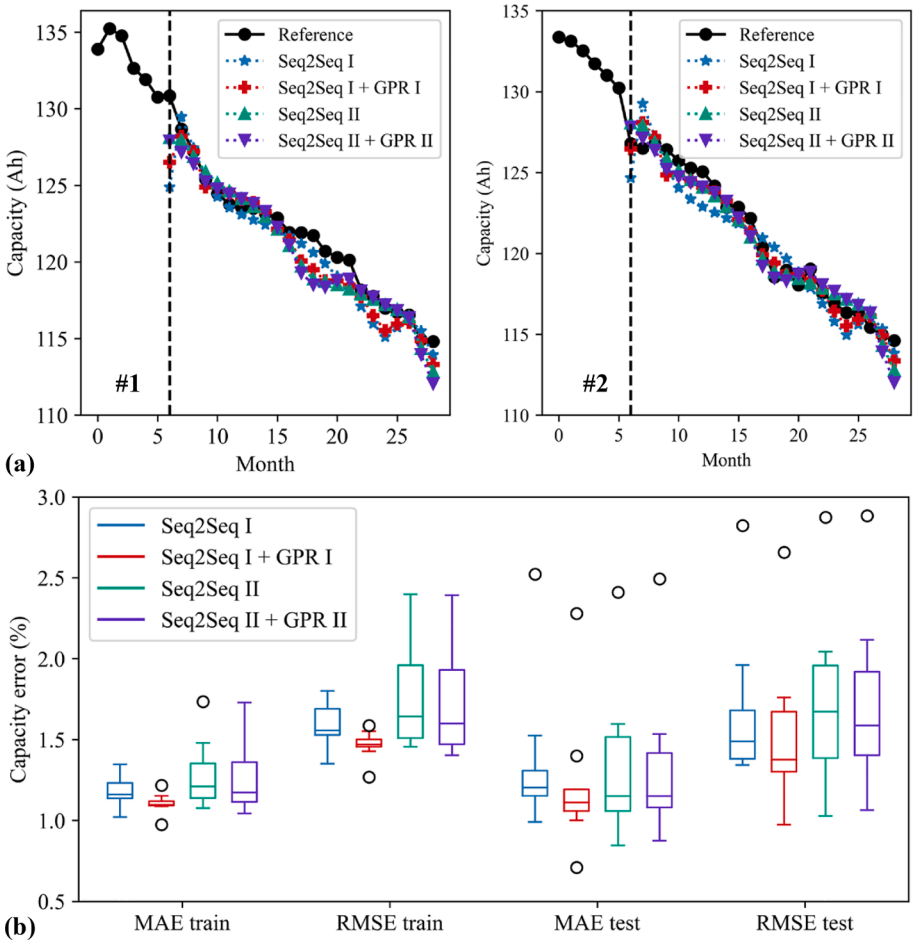


Fig. 9. The results of battery capacity prediction for different models. (a) capacity prediction results; (b) prediction error distribution.

accuracy of battery capacity trajectory, two residual models are established based on the GPR to compensate the prediction error.

4.2.1. The principle of GPR

The GPR is known as a machine learning method with non-parametric modeling and uncertainty evaluation [34]. For a typical regression problem, the observations usually contain Gaussian white noises and can be formulated as,

$$y_i = f(x_i) + N\left(0, \sigma_n^2\right) \tag{6}$$

where $\boldsymbol{f}(\boldsymbol{x}) = [f(x_1), f(x_2), \ldots, f(x_n)]$ is a Gaussian process, x_i is the i^{th} feature, and σ_n^2 is the noise covariance. $\boldsymbol{f}(\boldsymbol{x})$ can be described as $\boldsymbol{f}(\boldsymbol{x}) \sim N$ (0, **K**), where $K_{ij} = k\ (x_i, x_j)$ is the kernel function, which measures the distance between points x_i and x_j and mainly determines the estimation accuracy of regression model. The most widely used kernel function is the squared exponential function, and can be expressed as,

$$k\left(x_i, x_j\right) = \sigma_f^2 \exp\left(\frac{-\left\|x_i - x_j\right\|^2}{2l^2}\right) \tag{7}$$

where σ_f controls the amplitude of the kernel function and l determines the importance of each input feature. In the training process, the GPR only needs to optimize the hyperparameters [σ_f, l, σ_n], which can be achieved by maximizing the marginal likelihood [32,35]. Thanks to the

Table 6
The statistical errors of battery capacity prediction for different models.

	Method	MAE_*ave*	RMSE_*ave*
Training	Seq2Seq I	1.18%	1.59%
	Seq2Seq I + GPR I	1.10%	1.47%
	Seq2Seq II	1.28%	1.75%
	Seq2Seq II + GPR II	1.26%	1.72%
	Iterative GPR	1.01%	1.33%
	Iterative LSTM	1.05%	1.43%
Testing	Seq2Seq I	1.34%	1.66%
	Seq2Seq I + GPR I	1.21%	1.52%
	Seq2Seq II	1.32%	1.72%
	Seq2Seq II + GPR II	1.31%	1.71%
	Iterative GPR	2.34%	3.34%
	Iterative LSTM	2.56%	3.23%

advantages in non-parametric modeling and probabilistic prediction, the GPR is an ideal choice to approximate nonlinear processes with unknown functions, such as the residual model in this study.

4.2.2. Residual models

In order to compensate the prediction error of the Seq2Seq model, two methods are used to establish the GPR-based residual models, as illustrated in Fig. 5. One is to model the prediction error of the Seq2Seq model directly. In contrast, another one first uses the empirical mode decomposition (EMD) to divide the raw capacity sequence into two parts, namely a residual function and intrinsic mode functions (IMFs), then the residual which can represent the whole degradation trend of capacity [36,37] is simulated by the Seq2Seq model, while the remaining part (IMFs) is fitted by the GPR. Table 3 lists the input and output of the above models, in which Seq2Seq I uses the raw capacity sequence (y) to train the model directly, Seq2Seq II uses the residual of capacity sequence (*Res*) obtained by the EMD, GPR I uses the prediction error ($\triangle y$) of the Seq2Seq I as output, and GPR II regards *y-Res* as target.

To consider the effect of time and temperature on the local prediction error of battery capacity, the input of the residual model ($\mathbf{X}_2$) consists of three related features, namely month, T_{max_ave}, and T_{min_ave}. For the GPR model, it is a single point estimation problem, which means that the estimation of the future target needs to know the corresponding future features. For the month feature, the future value can be known naturally. For the temperature features, the values of tested vehicles can be assumed to be equal to the average values of other vehicles. Since these vehicles are used as commercial taxis in the same city, they have a similar operating condition, which ensures that their battery temperature evolution is almost the same. The reasonability of this assumption can be further verified by Fig. 6. It can be found that the battery temperature of different vehicles has a similar evolution trend, which is obviously determined by month. Therefore, it would not cause a large error by setting the temperature features of the tested vehicle equal to the average values of the other training vehicles.

5. Results and discussion

In order to verify the prediction accuracy of the proposed methods, two metrics, namely mean absolute error (MAE) and root mean square error (RMSE) are chosen to quantify the prediction error, and are calculated as,

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

where y_i is the reference value of battery pack capacity, $\hat{y}_i$ is the predicted value and n is the number of samples. The pyGPs library [38] is imported to achieve the GPR model in this study, and the TensorFlow Keras is implemented to construct the Seq2Seq model. The parameters setting of the Seq2Seq model refers to reference [39]. The model first has a masking layer to trim zero paddings, so that it only learns from actual data. Both the encoder and decoder of the model consist of two bidirectional LSTM layers, and each layer has 64 nodes. As the prediction sequences have different lengths for different vehicles, the training data are padded with zero before being forwarded to the model. During the training process of the Seq2Seq model, the "Adam" optimizer is used to train the network, the learning rate is set to 1e-4, and the number of epochs is set to 300 to obtain a fully trained model.

5.1. Verification of capacity estimation

To verify the effectiveness of the selected feature set, battery capacity estimation models are constructed by utilizing mature machine learning methods. The selected feature set is served as the input, and the labeled battery capacity is the output. High-precision estimation can be easily achieved if the selected features contain complete battery aging information. In this study, two commonly used methods (the SVR and GPR) are employed to construct the capacity estimation model. The performance and accuracy of these two methods have been validated by the literature [17,40].

For the studied 20 vehicles, the samples produced by 10 vehicles are utilized to conduct model training, while the samples produced by the remaining 10 vehicles are used to test the model. The capacity estimation results of the SVR and GPR models are illustrated in Fig. 7, including both the training and the test results, and the statistical errors of capacity estimation are listed in Table 4. It can be observed that the two models are well trained with the MAE lower than 1% and present high accuracy for tested vehicles, both the MAEs and RMSEs of the two models for the tested vehicle do not exceed 1.2% and 1.6%, respectively. Besides, the GPR model has a slightly higher accuracy than the SVR model. These results verify that the selected feature set can sufficiently represent battery capacity degradation.

5.2. Verification of capacity prediction

The prediction accuracy of the Seq2Seq model is evaluated with the data set of 20 EVs using ten-fold cross validation. The complete data set is divided into ten mutually exclusive folds, and for each trial, the data of 90% vehicles (18 vehicles) are used for training and the data of remaining 10% vehicles (2 vehicles) are used for testing. For each vehicle, the data of first 6 months is assumed to be known, and the remaining capacity sequence about 23 months is to be predicted. To verify the effectiveness of the feature selection procedure, the predicted results based on different feature sets are compared in this section. Three feature sets as listed in Table 5 are used as the input of the Seq2Seq model. F1 is the feature set determined by the proposed feature selection procedure. F2 consists of all features with PCC $>$ 0.5, and F3 is obtained by removing I_{std} and V_{pack_std} from F1. Compared with F1, F2 has 5 more redundant features, while F3 has 2 fewer useful features. The results of the first trial based on different feature sets are shown in Fig. 8, in which the predicted results of two vehicles are denoted by #1 and #2, respectively. Besides, the average value of MAEs (MAE_*ave*) and the average value of RMSEs (RMSE_*ave*) for all trials are calculated and are listed in Table 5.

It can be observed that all the three groups of features can obtain capacity prediction results with relatively high accuracy. The highest accuracy is obtained by using F2. Compared with the results of using F1, using F2 feature only reduces prediction error by 0.02% MAE_*ave* for testing, but at the cost of adding 5 more features. For F3, due to the lack of important features, its accuracy is obviously lower than that of F1. Therefore, considering the accuracy and model complexity, the feature set F1 is the best choice.

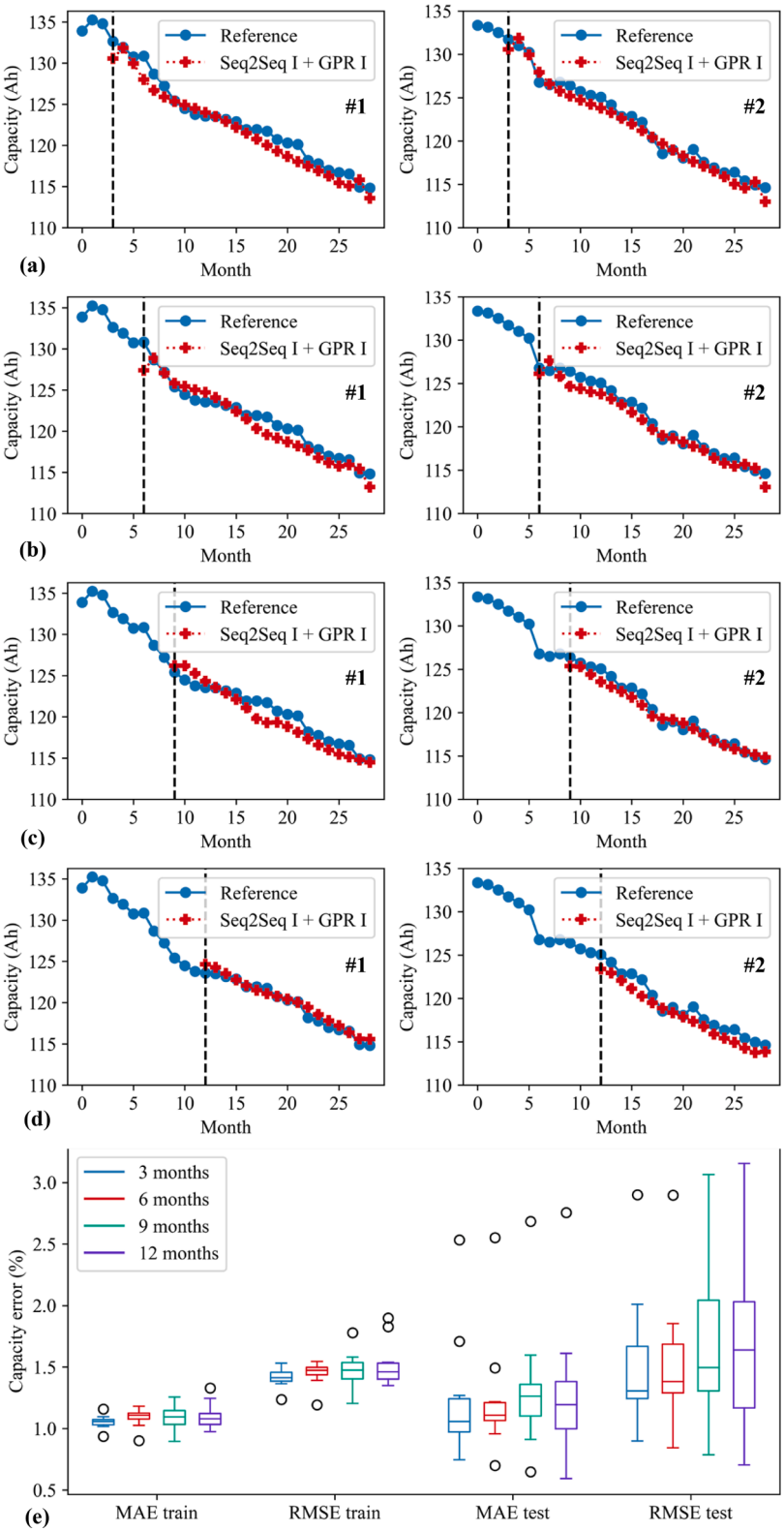


Fig. 10. The results of battery capacity prediction under different start points. (a) first 3 months data as input; (b) first 6 months data as input; (c) first 9 months data as input; (d) first 12 months data as input; (e) the distribution of prediction errors.

5.3. *Results of different prediction methods*

The results of battery capacity prediction based on different methods are shown in Fig. 9, in which the results of the first trial are presented in Fig. 9(a) and the distribution of prediction errors of all trails is illustrated in Fig. 9(b). Besides, the MAE_*ave* and the RMSE_*ave* for all trials are calculated and are listed in Table 6. According to the results, all the four models can predict the battery capacity trajectory with acceptable accuracy based on the early known data. The GPR can get the confidence interval of the estimated residual, but compared with the

capacity, the magnitude of its residual is much smaller. The confidence interval of the final prediction curve is much narrower, so it is not plotted in Fig. 9.

In detail, the combination of Seq2Seq I + GPR I has the best performance, and its MAE_*ave* and RMSE_*ave* in test process are equal to 1.21% and 1.52%, respectively. Compared with the Seq2Seq I, the Seq2Seq II uses the residual of capacity sequence through the EMD as target, but no obvious improvement is observed by this operation in this study. The GPR I can effectively compensate the prediction errors of the Seq2Seq I model, and the prediction errors can be further reduced by around 10%. In contrast, the compensation ability of the GPR II is not evident in this case.

For battery capacity sequence prediction, there are two types of commonly used methods, namely iterative prediction methods and sequence-to-sequence prediction methods [4,26]. For iterative prediction methods, the historical data of the previous n-steps are used as the training input, and the predicted m-steps ($m < n$) output is added to the input to conduce next prediction. Therefore, the future degradation trajectory is obtained after several times iteration.

To compare the performance of the sequence-to-sequence prediction and iterative prediction, the prediction results based on two iterative prediction methods are listed in Table 6. For iterative prediction, the training process has a higher accuracy due to the shorter prediction step. However, during the testing process, the prediction error of each step will go to the next prediction, resulting in a lower overall prediction accuracy. It can be observed that the prediction accuracy of the iterative prediction methods under testing process is obviously lower than that of the Seq2Seq methods.

5.4. Results of different prediction start points

In this section, the performance of the proposed method under different prediction start points is investigated. The early data of battery for the first 3 months, 6 months, 9 months, and 12 months are assumed to be known, respectively. For the data set of 20 EVs, the ten-fold cross validation is also used to evaluate the prediction accuracy, so each case has two vehicles to be predicted. The prediction results of the Seq2Seq I + GPR I model are illustrated in Fig. 10, where the predicted capacity curves under different start points are shown in Fig. 10(a)-(d) and the distribution of prediction errors is presented in Fig. 10(e). It can be found that the future capacity trajectory can be predicted accurately for all the four cases. As the decrease of the amount of prior data, no obvious drop in prediction accuracy is observed, which validates the proposed method has strong robustness. When only first 3 months of data are known, the battery capacity of the remaining 23 months is predicted, and the obtained MAE_ave and RMSE_ave are equal to 1.24% and 1.53%, respectively.

6. Conclusion

For on-road EVs, it is of significance to forecast the capacity degradation trajectory of battery system, which contributes to timely maintenance, residual value assessment and second-life utilization. In this paper, a battery capacity prognostic method based on charging data and data-driven algorithms is proposed to solve this problem.

First, battery capacity is calculated based on the collected battery charging data and the variant of Ampere-integral formula, and the statistical median values of the calculated capacity during a month is regarded as the labeled capacity, which can effectively reduce the effect of SOC errors and data noise. Battery capacity sequences with obvious evolution trend can be obtained via this method. Then, the statistical characteristics (*mean*, *sum* and *standard deviation* values) of battery charging data, including battery current, pack voltage, SOC, maximum and minimum cell voltages, maximum and minimum cell temperatures, cell voltage difference and cell temperature difference are investigated, and their correlations to battery capacity are analyzed through the PCC and GRG. A feature selection procedure based on the threshold is enacted to get rid of ineffective features, and an optimal feature set with 8 features is determined. Moreover, the Seq2Seq model is employed to conduct the future capacity sequence prediction by using the extracted features as input, and two residual models based on the GPR are also proposed to compensate the prediction errors. Finally, the data set of 20 EVs operating about 29 months are used to verify the proposed methods. The experimental results indicate that the Seq2Seq I + GPR I owns the best performance, given the first 3 months of data as input, the battery capacity of the remaining 23 months can be predicted with the MAE and RMSE equal to 1.24% and 1.53%, respectively.

The performance of the proposed method needs to be further verified on other vehicles with different types of batteries, such as lithium iron phosphate, and with different usage scenario, such as private cars.

CRediT authorship contribution statement

Zhongwei Deng: Conceptualization, Investigation, Methodology, Software, Writing – original draft. **Le Xu:** Conceptualization, Investigation, Methodology, Validation, Writing – original draft. **Hongao Liu:** Conceptualization, Investigation, Methodology. **Xiaosong Hu:** Conceptualization, Investigation. **Zhixuan Duan:** Investigation, Validation. **Yu Xu:** Investigation, Methodology, Validation.

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.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 52102420), National Key Research and Development Program of China (Grant No. 2022YFE0102700), and China Postdoctoral Science Foundation (Grant No. 2021M693725). The authors gratefully acknowledge the great help of the fund.

Data availability

The data set of 20 EVs used in this study with sensitive information removing are available online at https://github.com/TengMichael/battery-charging-data-of-on-road-electric-vehicles.

References

[1] Taalbi J, Nielsen H. The role of energy infrastructure in shaping early adoption of electric and gasoline cars. Nat Energy 2021;6(10):970–6.
[2] Lebrouhi BE, Khattari Y, Lamrani B, Maaroufi M, Zeraouli Y, Kousksou T. Key challenges for a large-scale development of battery electric vehicles: A comprehensive review. J Energy Storage 2021;44:103273.
[3] Harper G, Sommerville R, Kendrick E, Driscoll L, Slater P, Stolkin R, et al. Recycling lithium-ion batteries from electric vehicles. Nature 2019;575(7781):75–86.
[4] Hu X, Xu L, Lin X, Pecht M. Battery Lifetime Prognostics. Joule 2020;4(2):310–46.
[5] Severson KA, Attia PM, Jin N, Perkins N, Jiang B, Yang Z, et al. Data-driven prediction of battery cycle life before capacity degradation. Nat Energy 2019;4(5):383–91.
[6] Freedom C. Battery test manual for power-assist hybrid electric vehicles. INEEL, October. 2003.
[7] Deng Z, Yang L, Cai Y, Deng H, Sun L. Online available capacity prediction and state of charge estimation based on advanced data-driven algorithms for lithium iron phosphate battery. Energy 2016;112:469–80.
[8] Sulzer V, Mohtat P, Aitio A, Lee S, Yeh YT, Steinbacher F, et al. The challenge and opportunity of battery lifetime prediction from field data. Joule 2021;5(8):1934–55.

[9] Wang D, Yang F, Tsui KL, Zhou Q, Bae SJ. Remaining useful life prediction of lithium-ion batteries based on spherical cubature particle filter. IEEE Trans Instrum Meas 2016;65(6):1282–91.
[10] Li S, Fang H, Shi B. Remaining useful life estimation of Lithium-ion battery based on interacting multiple model particle filter and support vector regression. Reliab Eng Syst Saf 2021;210:107542.
[11] Randall AV, Perkins RD, Zhang X, Plett GL. Controls oriented reduced order modeling of solid-electrolyte interphase layer growth. J Power Sources 2012;209:282–8.
[12] Han X, Lu L, Zheng Y, Feng X, Li Z, Li J, et al. A review on the key issues of the lithium ion battery degradation among the whole life cycle. eTransportation 2019;1:100005.
[13] Reniers JM, Mulder G, Howey DA. Review and performance comparison of mechanical-chemical degradation models for lithium-ion batteries. J Electrochem Soc 2019;166(14):A3189–200.
[14] Deng Z, Hu X, Lin X, Xu L, Li J, Guo W. A reduced-order electrochemical model for all-solid-state batteries. IEEE Trans Transp Electrif 2021;7(2):464–73.
[15] Li C, Cui N, Wang C, Zhang C. Reduced-order electrochemical model for lithium-ion battery with domain decomposition and polynomial approximation methods. Energy 2021;221:119662.
[16] Li Yi, Liu K, Foley AM, Zülke A, Berecibar M, Nanini-Maury E, et al. Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review. Renewable Sustainable Energy Rev 2019;113:109254.
[17] Zhao Q, Qin X, Zhao H, Feng W. A novel prediction method based on the support vector regression for the remaining useful life of lithium-ion batteries. Microelectron Reliab 2018;85:99–108.
[18] Jiang B, Dai H, Wei X, Jiang Z. Multi-kernel relevance vector machine with parameter optimization for cycling aging prediction of lithium-ion batteries. IEEE J Emerg Sel Top Power Electron 2021:1.
[19] Deng Z, Hu X, Li P, Lin X, Bian X. Data-driven battery state of health estimation based on random partial charging data. IEEE Trans Power Electron 2022;37(5):5021–31.
[20] Deng Z, Lin X, Cai J, Hu X. Battery health estimation with degradation pattern recognition and transfer learning. J Power Sources 2022;525:231027.
[21] Tian J, Xiong R, Shen W, Lu J, Yang X-G. Deep neural network battery charging curve prediction using 30 points collected in 10 min. Joule 2021;5(6):1521–34.
[22] Tang X, Wang Y, Liu Qi, Gao F. Reconstruction of the incremental capacity trajectories from current-varying profiles for lithium-ion batteries. iScience 2021;24(10):103103.
[23] Zheng L, Zhu J, Lu D-D-C, Wang G, He T. Incremental capacity analysis and differential voltage analysis based state of charge and capacity estimation for lithium-ion batteries. Energy 2018;150:759–69.
[24] Saxena S, Ward L, Kubal J, Lu W, Babinec S, Paulson N. A convolutional neural network model for battery capacity fade curve prediction using early life data. J Power Sources 2022;542:231736.
[25] Lu J, Xiong R, Tian J, Wang C, Hsu C-W, Tsou N-T, et al. Battery degradation prediction against uncertain future conditions with recurrent neural network enabled deep learning. Energy Storage Mater 2022;50:139–51.
[26] Xu L, Deng Z, Xie Y, Lin X, Hu X. A novel hybrid physics-based and data-driven approach for degradation trajectory prediction in Li-ion batteries. IEEE Trans Transp Electrif 2022:1.
[27] Zhang Y, Wik T, Bergström J, Pecht M, Zou C. A machine learning-based framework for online prediction of battery ageing trajectory and lifetime using histogram data. J Power Sources 2022;526:231110.
[28] Meng J, Cai L, Stroe D-I, Luo G, Sui X, Teodorescu R. Lithium-ion battery state-of-health estimation in electric vehicle using optimized partial charging voltage profiles. Energy 2019;185:1054–62.
[29] Farmann A, Waag W, Marongiu A, Sauer DU. Critical review of on-board capacity estimation techniques for lithium-ion batteries in electric and hybrid electric vehicles. J Power Sources 2015;281:114–30.
[30] Zhou Y, Huang M, Chen Y, Tao Y. A novel health indicator for on-line lithium-ion batteries remaining useful life prediction. J Power Sources 2016;321:1–10.
[31] Tosun N. Determination of optimum parameters for multi-performance characteristics in drilling by using grey relational analysis. Int J Adv Manuf Technol 2006;28(5):450–5.
[32] Yang D, Zhang X, Pan R, Wang Y, Chen Z. A novel Gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve. J Power Sources 2018;384:387–95.
[33] Tang T, Yuan H. An indirect remaining useful life prognosis for Li-ion batteries based on health indicator and novel artificial neural network. J Energy Storage 2022;52:104701.
[34] Williams CK, Rasmussen CE. Gaussian processes for machine learning. MA: MIT press Cambridge; 2006.
[35] Deng Z, Hu X, Lin X, Che Y, Xu Le, Guo W. Data-driven state of charge estimation for lithium-ion battery packs based on Gaussian process regression. Energy 2020;205:118000.
[36] Li K, Wang Y, Chen Z. A comparative study of battery state-of-health estimation based on empirical mode decomposition and neural network. J Energy Storage 2022;54:105333.
[37] Cheng G, Wang X, He Y. Remaining useful life and state of health prediction for lithium batteries based on empirical mode decomposition and a long and short memory neural network. Energy 2021;232:121022.
[38] Neumann M, Huang S, Marthaler DE, Kersting K. pyGPs: a Python library for Gaussian process regression and classification. J Mach Learn Res 2015;16(1):2611–6.
[39] Li W, Zhang H, van Vlijmen B, Dechent P, Sauer DU. Forecasting battery capacity and power degradation with multi-task learning. Energy Storage Mater 2022;53:453–66.
[40] Deng Z, Hu X, Lin X, Xu L, Che Y, Hu L. General discharge voltage information enabled health evaluation for lithium-ion batteries. IEEE/ASME Trans Mechatron 2021;26(3):1295–306.