AIMS Energy, 13(2): 290–308.
DOI: 10.3934/energy.2025012
Received: 14 January 2025
Revised: 13 March 2025
Accepted: 21 March 2025
Published: 26 March 2025

https://www.aimspress.com/journal/energy

Research article

State of health estimation of lithium-ion batteries based on data-driven methods with a selected charging voltage interval

Junguang Sun[1], Xiaodong Zhang[1], Wenrui Cao[2], Lili Bo[3], Changhai Liu[2] and Bin Wang[2,*]

[1] State Key Laboratory of Environmental Adaptability for Industrial Products, China National Electric Apparatus Research Institute Co., Ltd, Guangzhou, 510663, China
[2] State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, 710049, China
[3] China aviation lithium battery Co., Ltd, Luoyang, 471003, China

* **Correspondence:** Email: b.wang@xjtu.edu.cn. Tel: +8615891448179.

Abstract: Accurate state of health (SOH) estimation of lithium-ion batteries is of great importance for achieving efficient energy management in the overall battery energy storage system. Traditional data-driven methods for the SOH estimation of lithium-ion batteries usually require an enormous amount of data from the whole charging phase, which leads to poor performance in both computational efficiency and computational cost. To address this issue, this paper proposes the SOH estimation methods for lithium-ion batteries based on the limited data from a selected charging voltage interval. First, this study uses incremental capacity curves and Pearson correlation analysis to select an optimal and limited charging voltage interval that is the most relevant to lithium-ion battery degradation. Then, the SOH estimation based on two typical data-driven methods, including random forest regression (RFR) and support vector regression (SVR), would be implemented with the selected charging voltage interval. Results show that both the RFR and the SVR methods can achieve excellent accuracy, while each has its own irreplaceable advantages. However, compared with other voltage intervals using the two data-driven methods, the corresponding SOH estimation with the selected charging voltage interval shows the best performance. Hence, the data-driven methods based on the selected charging voltage interval have significant potential and advantages in the field of lithium-ion battery SOH estimation.

Keywords: lithium-ion battery; state of health; data-driven methods; random forest regression; support vector regression

1. Introduction

Lithium-ion batteries (LIBs), known for their high energy density, low self-discharge rate, and long lifespan, have been widely used in electric vehicles, renewable energy storage systems, and portable electronic devices [1,2]. However, with the long-term operation, the overall performance of LIBs would gradually degrade [3]. Therefore, accurate estimation of the state of health (SOH) for LIBs is very crucial for ensuring efficient energy management and safety of the overall battery energy storage system [4,5]. Generally, the battery SOH is influenced by these factors, such as discharge rate, operating temperature, and charging-discharging cycles [6]. The common definition of the SOH is the ratio of the current maximum usable capacity to the maximum usable capacity of a new battery [7,8]. Currently, features from the constant current-constant voltage (CC-CV) charging protocols are widely used for the SOH estimation of LIBs due to their simplicity and ease of acquisition [9,10]. However, users usually do not fully discharge and then fully charge the battery regularly in accordance with the CC-CV protocol in practical applications, posing a significant challenge for the SOH estimation based on these data-driven methods with complete data from the entire charging phase. Therefore, the SOH estimation based on partial or optimal charging data becomes an urgent requirement in practical applications.

Existing SOH estimation methods can be roughly divided into three categories: 1) empirical or semi-empirical models [11,12]; 2) physics-based models [13–15]; and 3) data-driven methods [16–18]. Although empirical models are simple, they lack a clear physical significance for LIBs. Physics-based models are accurate. However, using physics-based models to estimate the battery SOH would be very complicated for users and engineers. Nowadays, data-driven methods have gained increasing attention in the battery SOH estimation field due to their model-free nature [19–23]. Wen et al. [19] successfully employed an IC curve feature based on a BP neural network to achieve high estimation precision for the lithium battery SOH prediction across various temperature conditions. Ren et al. [20] proposed a novel SOH estimation method for the lithium-ion battery pack based on cross-generative adversarial networks. With a hybrid extreme learning machine based on an adaptive boosting algorithm, this SOH estimation method could achieve accurate SOH estimation with incomplete data. Li et al. [21] extracted health indicators from public datasets and eliminated redundant features through Pearson correlation coefficients and principal component analysis. In addition, Liu et al. [22] proposed a capture method for battery degradation by considering the tested voltage, temperature, and current data, which could achieve accurate SOH estimation even with minimal training data. Moreover, some studies explored data and model joint-driven methods for more precise battery state estimation based on unscented Kalman filter or deep learning methods [18,23]. Although these research efforts have yielded satisfactory accuracy, they should consider the feasibility of data acquisition and processing in practical applications. From the engineering standpoint, obtaining relevant data from smaller datasets is very crucial. Therefore, this work employs a small amount of data based on an optimal charging voltage interval and efficient machine-learning methods to achieve the SOH estimation for the lithium battery.

As we know, the charging curve of an LIB would present regular changes in accordance with its performance degradation [24,25]. Therefore, using the data-driven methods based on features from charging curves of LIBs to estimate their SOH has become a new trend. Generally, these data-driven methods for the battery SOH estimation should consider the changes in the entire charging phase, which leads to poor performance in both computational efficiency and computational cost [26].

To improve these data-driven methods based on the entire charging interval, Xiong et al. [27] proposed a feature selection method based on the correlation coefficient and the ReliefF algorithms. This method could significantly improve the accuracy of the battery SOH estimation. The corresponding results demonstrated that the high-precision SOH estimation could be achieved based on the selected features. Li et al. [28] analyzed the charging time during the constant current and the constant voltage stages, as well as the proportion of the charging time of the constant voltage stage. Then, a correlated feature for the battery SOH estimation could be determined. This study indicated that both the constant current and the constant voltage charging phases could be considered for the feature extraction of the battery SOH estimation. Lin et al. [26] used data-driven algorithms with the features extracted from four partial voltage segments to achieve an accurate estimation of the battery SOH. However, the above studies did not conduct detailed research on the selection of charging voltage intervals. The previous selection methods were only based on engineering experience.

According to previous studies, the incremental capacity (IC) curve was commonly used for the feature extraction of the battery SOH estimation [29]. The IC curve can effectively reflect the capacity and voltage changes during the charging operation, which can be used for identifying characteristic peaks related to phase transitions during the transmission processes of lithium ions into the battery. Previous studies indicated that the charging voltage interval selection could be performed based on the peak positions of the IC curve [27,30]. Bian et al. [31] proposed a battery SOH estimation method combining an open-circuit voltage (OCV) model and the IC analysis (ICA). According to the OCV model, interference-free IC curves could be obtained, enabling the extraction of a series of morphological features. Li et al. [32] also proposed a method combining the ICA and the grey relational analysis for the battery SOH estimation. In addition, Lin et al. [33] used the internal resistance and the peak/valley points of the IC curve as features. On this basis, accurate SOH estimation could be achieved based on a machine learning algorithm. However, these studies did not propose a clear method for the optimal voltage interval selection. In summary, existing references generally lack systematic exploration for the battery SOH estimation based on the limited or optimal charging interval selection. The selection of a partial charging voltage interval was only based on engineering experience, which lacks detailed correlation analyses between the selected charging voltage interval and the SOH of LIBs. To achieve high performance in both computational efficiency and computational cost, this paper employs two data-driven methods using the ICA and Pearson correlation analysis to select the optimal charging voltage interval for the SOH estimation of LIBs.

The main contributions of this paper are as follows:

(1) The proposed SOH estimation method combines the ICA and Pearson correlation analysis, which can identify and select the optimal charging voltage interval. On this basis, the most relevant health features for the battery SOH estimation can be extracted from the selected charging voltage interval.

(2) With the most relevant health features from the selected charging voltage interval, it is demonstrated that high accuracy in battery SOH estimation can be achieved with the data-driven methods even with a small amount of training data, which can effectively improve computational efficiency and reduce computational cost.

(3) Two typical data-driven methods, including the RFR and the SVR for the SOH estimation, are analyzed. Results show that both the RFR and the SVR methods have their own irreplaceable advantages: the SVR has stronger generalization ability, whereas the RFR possesses better overall fitting capability.

This paper includes four sections after the introduction. Section 2 presents the selection and

detailed analyses for the optimal charging voltage interval. Section 3 introduces two data-driven methods for the SOH estimation. Results and discussions are presented in Section 4. Finally, conclusions are given in Section 5.

2. Selection of the optimal charging voltage interval

2.1. Aging experiment

To obtain the charging data for the SOH estimation, three 18650 LIBs are used for the experimental test. The detailed parameters are shown in Table 1. Every battery would undergo CC-CV charging to reach the upper limited voltage at first. Then, these batteries would undergo CC discharging to reach the low-limited voltage. Figure 1 shows the capacity degradation curves of the three 18650 LIBs. In the experiment, Cell1 is charged at 1 C and discharged at 0.5 C, respectively. After 509 charge-discharge cycles, its SOH reaches 91.8%. Meanwhile, Cell2 is charged and discharged with 1 C. Its SOH is 81.03% after 720 cycles. In addition, Cell3 is charged at 1 C and discharged at 3 C, respectively. Its SOH is equal to 80.69% after 1181 cycles. All charge and discharge tests for the three 18650 LIBs are conducted at an ambient temperature of 25 °C. It should be noted that it needs 1 hour for the rest time between CC-CV charging and CC discharging.

It is worth noting that the tested lithium battery is a type of modified high-power lithium battery designed for unmanned aerial vehicle (UAV) applications. The high-power lithium battery would show better performance when it operates with a suitable high-power output compared with the low-power operation. Therefore, under high-power discharge conditions, the high-power lithium battery might perform at a lower degradation rate compared with low-discharging conditions. For more details, please refer to our previous work in [34].

Table 1. Properties of tested LIBs.

Positive electrode material	Nickel Manganese Cobalt Oxid
Positive electrode material	Graphite
Nominal voltage	3.7 V
Nominal capacity	2.4 Ah
Operating voltage range	2.5~4.2 V

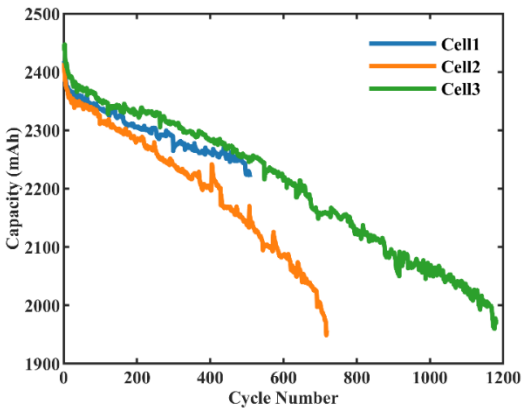

Figure 1. Capacity degradation of the three 18650 batteries.

2.2. Selection for the optimal charging voltage interval

Figure 2 presents the capacity-voltage (C-V) curves based on the experimental data. It can be seen that the envelope area of the C-V curve decreases along with the increasing number of operation cycles. It is evident that the C-V curves contain rich information about the aging or capacity degradation of LIBs. In fact, the voltage points in the regions with the most significant changes on the C-V curves can serve as external behavior characteristics of the battery aging [35].

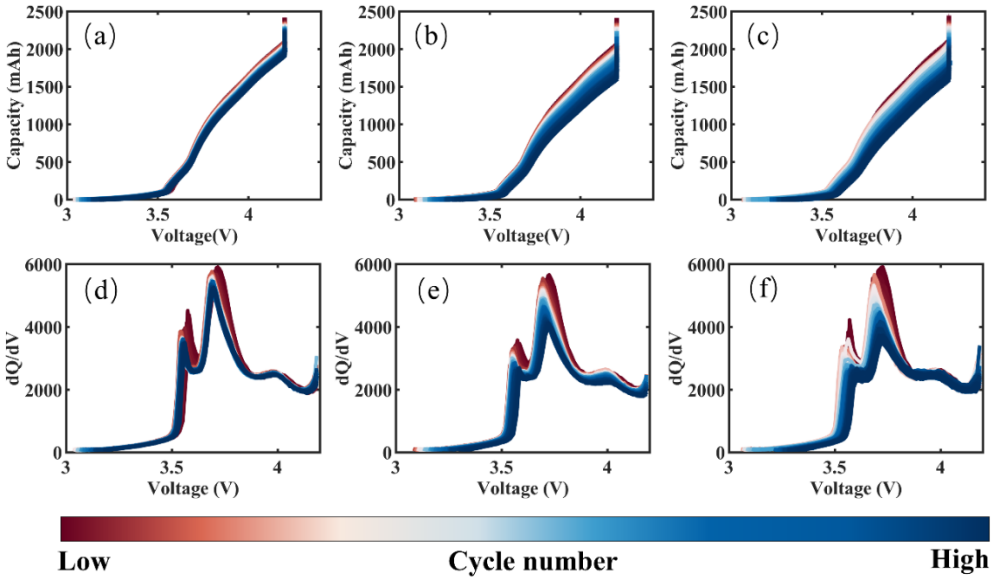

Figure 2. C-V curves and aging features of the three 18650 batteries. (a) The C-V curve of Cell1. (b) The C-V curve of Cell2. (c) The C-V curve of Cell3. (d) The IC curve of Cell1. (e) The IC curve of Cell2. (f) The IC curve of Cell3.

By differentiating the C-V curves of the three 18650 LIBs, specific voltage points can be observed. The corresponding definition for the IC can be expressed as:

$$\mathrm{IC} = \frac{\mathrm{d}Q}{\mathrm{d}V} \tag{1}$$

where Q is the charging capacity and V is the charging voltage.

Furthermore, an appropriate voltage step needs to be designed, as the voltage difference appears in the denominator in numerical differentiation. If the voltage step is too large, the health features from the IC curve would not be obvious. Conversely, if the voltage step is too small, it may result in a sudden jump in the calculation value. After multiple calculations and corrections, a voltage sequence V_C with a voltage step of 0.015 V is constructed, as shown in Eq (2). The corresponding capacity values for this voltage sequence are determined by using smooth spline interpolation based on the C-V curves.

$$V_C = [2.5, 2.515, 2.530, \ldots, 4.20] \tag{2}$$

Based on Eqs (1) and (2), the formula for the IC curve can be rewritten as follows:

$$IC = \frac{\Delta Q}{\Delta V} = \frac{Q_{k+1} - Q_k}{0.015} \tag{3}$$

Figure 2 (d–f) shows the changes of the IC curves on the whole charging voltage interval of different operation cycles. To select the optimal charging voltage interval, Figure 3 illustrates the enlarged and detailed IC curves with different SOH of Cell2. It can be seen that the first peak is around 3.52 V, the second peak is around 3.73 V, and the third peak is around 3.95 V, respectively. Meanwhile, the first valley is around 3.65 V, and the second valley is around 4.02 V, respectively. In addition, the peak and valley points of Cell1 and Cell3 are similar to those of Cell2.

It has been demonstrated that the peak and valley points on the IC curves are highly correlated with the kinetics of electrochemical reactions and the phase transitions within the electrode materials [36]. For these peak points, the rate of change in battery capacity in accordance with the voltage reaches its maximum value, indicating the fastest progress of chemical reactions. Similarly, the valley points also have significant physical meanings. For these valley points, the rate of change in capacity in accordance with the voltage reaches its minimum value, reflecting the changes in the activity of the electrode materials during the charge-discharge process. Therefore, this study employs the Pearson correlation analysis to investigate the degree of correlation between these peak/valley points and the battery aging. The formula for the Pearson correlation analysis can be represented as follows:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{4}$$

where X_i is the i^{th} feature, Y_i is the i^{th} SOH, $\bar{X}$ and $\bar{Y}$ are average values of the specific feature sequence and the battery SOH sequence, respectively.

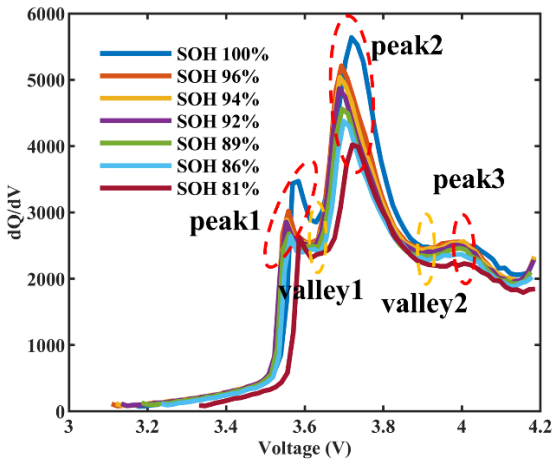

Figure 3. Detailed IC curves of Cell2.

The Pearson correlation coefficients between these peak/valley points and the battery SOH are presented in Table 2. Obviously, the correlation analysis results are all negative values, indicating that during the capacity degradation process, the peak and valley points exhibit the overall leftward shift trend, which is consistent with the characteristics illustrated in Figure 2 (d–f). Hence, these peak and valley points provide suitable locations for the calibration of charging voltage intervals.

Table 2. Pearson correlation analysis between selected points and the battery SOH.

	peak 1	peak 2	peak 3	valley 1	valley 2
Cell1	–0.9760	–0.9762	–0.9465	–0.9710	–0.9800
Cell2	–0.9868	–0.9877	–0.9770	–0.9906	–0.9868
Cell3	–0.9960	–0.9959	–0.9500	–0.9913	–0.9948

As shown in Table 3, ten voltage intervals can be divided based on the five points. Additionally, to achieve comprehensive correlation analyses of different voltage intervals and the battery SOH, the voltage interval including the entire constant current charging phase should be included in this study. With the defined voltage intervals, the relationship among the charging time, charging capacity, and the SOH is investigated based on Pearson correlation analysis. It should be noticed that the charging voltage intervals before 3.52 V and after 4.02 V are not suitable for the battery SOH estimation, as existing research has confirmed that the data at the beginning and end of the charging process cannot effectively reflect the battery aging characteristics [27]. Moreover, it can be observed from Figure 2 that the aging trends in the two charging voltage intervals are not obvious. Therefore, the maximum charging voltage interval is designed as 3.52~4.02 V.

Table 3. Calibration of voltage intervals.

Voltage interval	Voltage range
S1	3.52~3.65 V
S2	3.52~3.73 V
S3	3.52~3.95 V
S4	3.52~4.02 V
S5	3.65~3.73 V
S6	3.65~3.95 V
S7	3.65~4.02 V
S8	3.73~3.95 V
S9	3.73~4.02 V
S10	3.95~4.02 V

As shown in Table 4, the Pearson correlation of features from different charging voltage intervals of the three 18650 LIBs reveals that the S3 segment has the highest correlation with the battery SOH compared with other voltage intervals. It is noteworthy that the S11 segment (i.e., the entire constant current charging phase) shows less correlation than the S3 segment, which indicates the presence of information redundancy during the whole charging phase. Meanwhile, the correlation between the S4 segment and battery aging would also be very high and at a level second only to that of the S3 segment. Figure 4 presents the correlation analyses among the normalized charging time, normalized charging capacity, and battery SOH for different charging voltage intervals. It can be observed that the trend in the S3 segment closely resembles the battery aging trend. Similarly, the use of ΔQ and Δt as features to reflect the battery aging trend can refer to the literature [37]. This provides a more intuitive demonstration of the effectiveness of the aging characteristics based on the selected S3 segment proposed in this study. In these figures, ΔQ and Δt represent the two battery aging features (i.e., the charging capacity and time) extracted from different voltage intervals, respectively. Finally, this study utilizes charging capacity and time extracted from the S3 segment as features for the SOH estimation.

Table 4. Pearson correlation analysis between intervals and SOH.

Voltage interval	Cell1	Cell2	Cell3
S1	0.6308	0.9395	0.9282
S2	0.8323	0.9753	0.9657
S3	0.9906	0.9988	0.9922
S4	0.9883	0.9986	0.9919
S5	0.9237	0.9893	0.9807
S6	0.9583	0.9917	0.9821
S7	0.9731	0.9951	0.9838
S8	0.7430	0.9951	0.9838
S9	0.8089	0.9631	0.8157
S10	0.5981	0.9609	0.8483
S11	0.9533	0.9960	0.9854

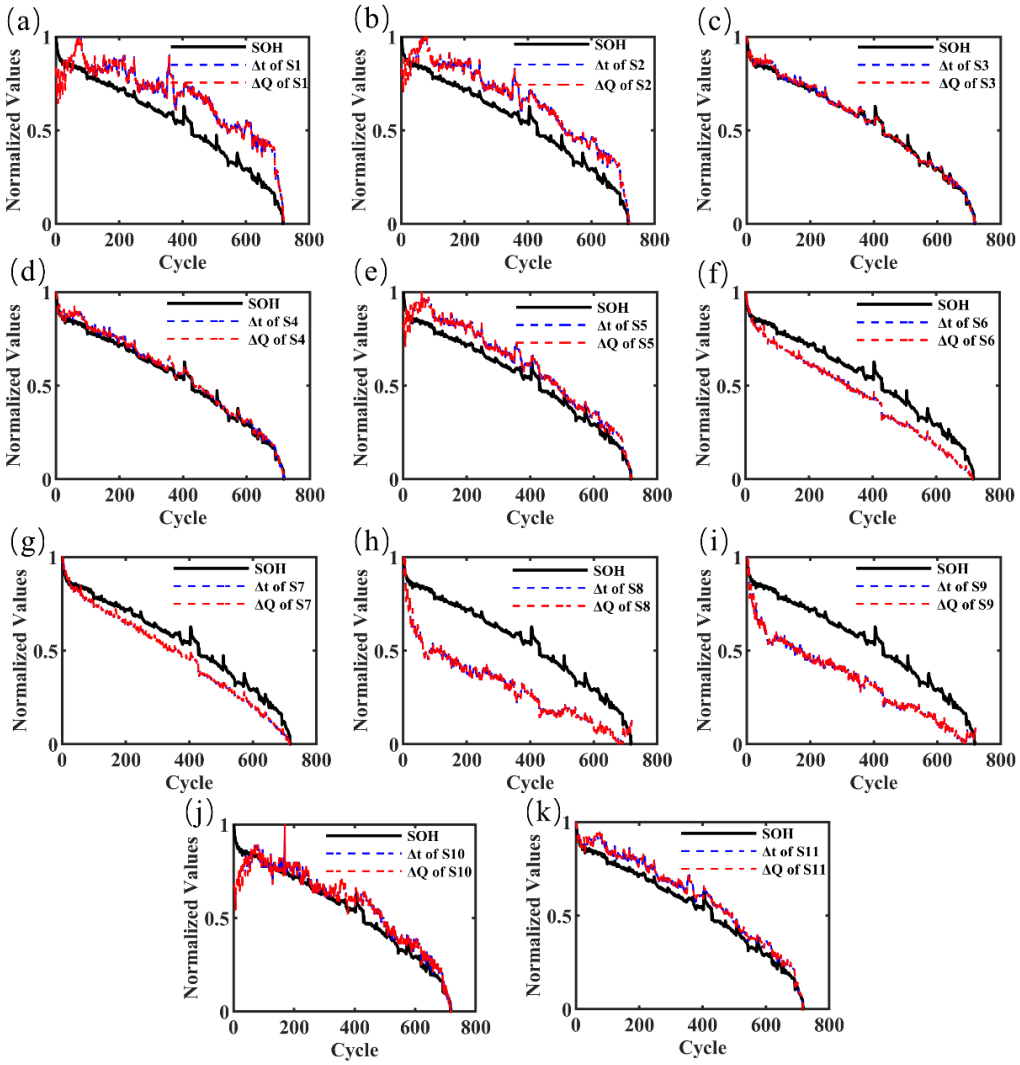

Figure 4. Correlation analyses among the SOH and the two features extracted from different voltage intervals. (a) S1, (b) S2, (c) S3, (d) S4, (e) S5, (f) S6, (g) S7, (h) S8, (i) S9, (j) S10, (k) S11.

3. Data-driven methods

Nowadays, using data-driven methods for the battery SOH estimation has become a research hotspot [31–33]. These data-driven methods would employ machine learning, deep learning, or statistical techniques to extract features. Machine learning methods can be categorized into supervised learning, unsupervised learning, and semi-supervised learning [38,39]. Deep learning methods mainly include artificial neural networks (ANN) [26], recurrent neural networks (RNN) [40], and their variants, such as long short-term memory (LSTM) [41]. Statistical methods include regression analysis and Bayesian methods, which utilize prior knowledge and data for inference and analysis [42]. For the aforementioned methods, machine learning algorithms RFR and SVR offer significant advantages in computational efficiency and model interpretability, enabling them to be implemented easily. By contrast, traditional neural networks, particularly deep neural networks, typically require substantial training time and high-performance computing resource to deal with large-scale datasets and complex nonlinear calculation. The corresponding algorithmic steps involved in the training process, such as backpropagation and gradient descent, are often time-consuming, limiting their feasibility for real-time or near real-time applications. Furthermore, due to their numerous parameters and intricate nonlinear transformations, neural networks are often regarded as black-box models. Although advancements in visualization techniques and interpretability algorithms have been made in recent years, neural networks still face significant challenges in directly understanding and interpreting the decision-making processes of models.

The RFR makes predictions through an ensemble of decision trees, resulting in satisfactory model performance and a reduced risk of overfitting, while the SVR exhibits strong generalization ability, achieving good fitting accuracy even with a small training set. Therefore, this study would employ the SVR and the RFR methods for the battery SOH estimation based on the selected charging voltage interval with an 80% training and 20% test dataset split. A type of $n \times 7$ input data matrix is constructed to feed the RFR and the SVR models, as shown in Eq (5).

$$Input = \begin{bmatrix} peak_1^{(1)}, peak_2^{(1)}, peak_3^{(1)}, valley_1^{(1)}, valley_2^{(1)}, \Delta Q^{(1)}, \Delta t^{(1)} \\ peak_1^{(2)}, peak_2^{(2)}, peak_3^{(2)}, valley_1^{(2)}, valley_2^{(2)}, \Delta Q^{(2)}, \Delta t^{(2)} \\ \vdots \\ peak_1^{(n)}, peak_2^{(n)}, peak_3^{(n)}, valley_1^{(n)}, valley_2^{(n)}, \Delta Q^{(n)}, \Delta t^{(n)} \end{bmatrix} \tag{5}$$

where $\Delta Q^{(n)}$ and $\Delta t^{(n)}$ represent the capacity and charging time during the n-th cycle within the selected charging voltage interval, respectively, the value of n is determined by 80% of the length of the input sequence, the term of *peak* and *valley* represents different points, while the superscript indicates different operation cycles.

3.1. Random forest regression

The RFR is a powerful ensemble learning method used for predicting continuous outcomes. It builds upon the foundations of decision trees and integrated techniques to enhance the predictive performance and robustness. The core of random forest is decision trees, which partition the feature space into regions with homogeneous target values. A decision tree T is used to partition the data based on input features and minimize a loss function. Formally, for a dataset, the objective is to find a tree

structure with the minimum solution: $\mathcal{D}=\{\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$. The decision trees divide the data to minimize prediction errors. In other words, it aims to minimize the sum of squared differences between the actual values y_i and the predicted values $\hat{y}_j$ in each leaf node R_j:

$$\min_{T} \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in R_j} (y_i - \hat{y}_j)^2 \tag{6}$$

where J is the number of terminal nodes (leaves), R_j represents the region corresponding to the j-th leaf, $\hat{y}_j$ is the predicted value in region R_j, i.e., the mean y_i of in R_j.

Single decision trees are prone to overfitting and high variance. Ensemble methods mitigate this by aggregating multiple models to improve generalization. Bagging is one of the ways to achieve ensemble methods, which involves 1) generating B bootstrap samples from the original dataset; 2) training a base learner (e.g., decision tree) on every bootstrap sample; and 3) aggregating the predictions from all base learners. For regression, the bagged predictor $\hat{f}_{\text{bag}}(\boldsymbol{x})$ is:

$$\hat{f}_{\text{bag}}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(\boldsymbol{x}) \tag{7}$$

where $\hat{f}_b(\boldsymbol{x})$ is the prediction from the b-th base learner.

Random forests extend bagging by introducing additional randomness in the model, particularly in feature selection. The random forest comprises B decision trees that are trained based on their bootstrap samples. Meanwhile, every decision tree uses a random subset of features when conducting splits. Its key steps include: 1) For each tree b, generating a bootstrap sample $\mathcal{D}_b$ from $\mathcal{D}$; 2) selecting a random subset of m features from the total p features ($m < p$) and determining the best split among these at each node in tree b; 3) growing each tree to its maximum depth without pruning. For a new input $\boldsymbol{x}$, the RFR prediction $\hat{y}(x)$ is the average of predictions from all trees:

$$\hat{y}(\boldsymbol{x}) = \frac{1}{B} \sum_{i=1}^{B} \hat{y}_b(\boldsymbol{x}) \tag{8}$$

where $\hat{y}_b(\boldsymbol{x})$ is the prediction from the b-th tree.

3.2. Support vector regression

The SVR extends the concepts of the support vector machine, which is originally designed for classification tasks, to handle regression problems. The core idea of the SVR is to model the relationship between input features and continuous target variables. It is achieved based on a function that deviates from the actual target values, which is lower than a predefined margin ε for all training data points. The regression function $f(\boldsymbol{x})$ of the SVR is typically modeled as follows:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x}) + b \tag{9}$$

where $\boldsymbol{w}$ is the weight vector, $\boldsymbol{x}$ is the input feature vector, $\Phi(\boldsymbol{x})$ is the mapping of $\boldsymbol{x}$, and b is the bias term.

The SVR employs kernel functions to implicitly map the input features into a high-dimensional space, enabling the modeling of complex patterns. Moreover, the SVR utilizes the ε-insensitive loss function, which disregards errors within a margin of ε. This approach allows the model to focus on significant deviations, promoting robustness and reducing the impact of noise. The ε-insensitive loss for a single data point $(\boldsymbol{x}_i, y_i)$ is represented as:

$$L\big(y, f(\boldsymbol{x})\big) = \max(0, |y - f(\boldsymbol{x})| - \varepsilon) \tag{10}$$

The objective of the SVR is to find the function $f(\boldsymbol{x})$ that minimizes the complexity of the model while ensuring that the predictions lie within the ε-margin of the actual target values. This balance is achieved through the following constrained optimization:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{11}$$

subject to:

$$\begin{cases} y_i - (w^T \phi(\boldsymbol{x}_i) + b) \leq \varepsilon + \xi_i \\ (w^T \phi(x_i) + b) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{12}$$

where ξ_i and ξ_i^* are slack variables representing deviations outside the ε-margin, $C > 0$ is a regularization parameter used to control the trade-off between model complexity and the penalty for deviations beyond ε.

To deal with non-linear issues, the SVR leverages the kernel trick, allowing the algorithm to operate in a high-dimensional feature space without explicitly computing the coordinates in that space. In other words, the dual formulation of the SVR optimization problem can be addressed based on kernel functions. The Lagrange multipliers α_i and α_i^* are introduced for the inequality constraints. Then, by optimizing the Lagrangian and applying the Karush-Kuhn-Tucker (KKT) conditions, the dual problem can be derived.

$$\max_{\alpha,\alpha^*} -\frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)\big(\alpha_j - \alpha_j^*\big) K\big(\boldsymbol{x}_i, \boldsymbol{x}_j\big) + \varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) - \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*) \tag{13}$$

The dual problem can be typically solved using optimization algorithms such as the sequential minimal optimization (SMO). Once the optimal Lagrange multipliers α_i^* and α_i are determined, the regression function can be expressed in terms of these multipliers and the kernel function. After solving the dual problem, the regression function $f(\boldsymbol{x})$ can be reconstructed as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{14}$$

The bias term *b* can be calculated by:

$$b = y_i - \sum_{j=1}^{n} (\alpha_j - \alpha_j^*) K(\boldsymbol{x}_j, \boldsymbol{x}_i) \tag{15}$$

This calculation ensures that the regression function appropriately aligns with the support vectors. The conventional kernel functions used in the SVR are shown in Table 5. To ensure the reproducibility of the estimation results for the two aforementioned algorithms employed in this study, a fixed random seed is set during the coding process. Additionally, all codes are written using sci-kit learn 1.4.2 in Python 3. In this study, cross-validation is employed to select the optimal kernel function for the SVR. The penalty parameter C is tested with the values 0.1, 10, and 100, while γ is set to Python's built-in options 'scale' and 'auto'. The final results indicate that the linear kernel is the optimal kernel function for battery SOH estimation. Meanwhile, the penalty parameter C is set to 100, and the option of γ is 'auto'.

Table 5. Conventional kernel functions.

Kernel functions	Formulas
Linear kernel	$K(x, x') = x^T x'$
Polynomial kernel	$K(x, x') = (\gamma x^T x' + r)^d$
RBF kernel	$K(x, x') = \exp(-\gamma \|x - x'\|^2)$
Sigmoid kernel	$K(x, x') = \tanh(\gamma x^T x' + r)$

* In this table, all γ are kernel coefficients. r represents the constant term, and d denotes the polynomial degree.

4. Results and discussion

In this section, the data-driven methods, including RFR and the SVR, are used to conduct the comprehensive analyses for the battery SOH estimation accuracy based on the selected charging voltage interval and other charging voltage intervals. To quantitatively describe the estimation results, the mean absolute error (MAE) and the root-mean-square error (RMSE) would be used as indicators of the estimation accuracy. In addition, the model's coefficient of determination (R^2) would be used as the fitting accuracy indicator. The three metrics can be expressed as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{16}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{17}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{18}$$

where y_i, $\hat{y}_i$ and $\bar{y}$ represent the real value, estimated value, and average value, respectively; n represents the number of samples.

The estimated results for the SOH of the three LIBs are shown in Figure 5. It can be seen that the estimated results of Cell2 are the best among the three LIBs. Meanwhile, the comparative analyses of the RFR and the SVR algorithms are conducted based on the selected charging voltage interval. With the RFR algorithm, the estimated values of R^2 for the three cells are 0.9915, 0.9986, and 0.9950, respectively. Meanwhile, the MAE for the three LIBs would be 0.10%, 0.11%, and 0.23%, respectively. Furthermore, the RMSE for the three LIBs are 0.16%, 0.15%, and 0.34%, respectively. As a comparison, with the SVR algorithm, the estimated value of R^2 for the three LIBs would be 0.9876, 0.9975, and 0.9907, respectively. The corresponding MAE would be 0.15%, 0.16%, and 0.32%, while the corresponding RMSE would be 0.19%, 0.21%, and 0.46%, respectively. Therefore, the fitting accuracy of the RFR method is better than that of the SVR method.

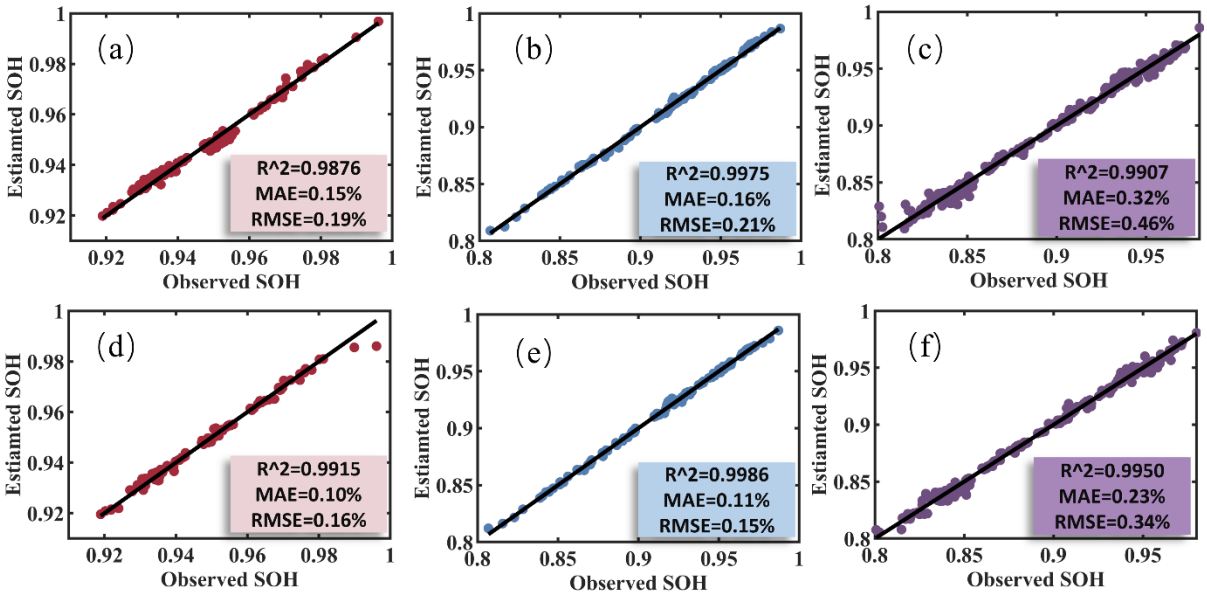

Figure 5. SOH estimation results with the SVR and the RFR. (a) cell1 with the SVR, (b) cell2 with the SVR, (c) cell3 with the SVR, (d) cell1 with the RFR, (e) cell2 with the RFR, and (f) cell3 with the RFR.

In addition, the remaining charging voltage intervals are also used with the SVR and the RFR data-driven methods to achieve the battery SOH estimation, respectively. The corresponding evaluation analyses are shown in Tables 6 and 7. Overall, the evaluation metrics for the SOH estimation with the S3 segment are the best among all charging voltage intervals. We use the blue color to mark this optimal charging voltage interval. Moreover, the S4 segment, which has a similar coverage area to the S3 segment, also demonstrates impressive estimation accuracy. Similarly, it can be observed that estimated SOH using the charging capacity and time of the whole voltage interval S11 as features also achieves satisfactory estimation accuracy. However, compared with the S3 segment, both the estimated MAE and estimated RMSE based on the S11 segment are increased. Moreover, collecting data for the S11 segment requires more than twice the amount of data compared with the S3 segment. Therefore, the battery SOH estimation based on the selected charging voltage interval not only reduces the time cost associated with data acquisition but also decreases the computational cost involved in regression estimation algorithms. Other charging voltage intervals, such as the S5, S6, and S7 segments, also

demonstrate relatively good accuracy for the battery SOH estimation. However, the corresponding accuracy is slightly inferior to that of the S3 segment. These results further support the necessity of selecting efficient health features and the optimal charging voltage interval for the battery SOH estimation.

Table 6. Estimated SOH evaluation analyses of the three LIBs cells with the RFR.

interval	Cell1			Cell2			Cell3		
	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
S1	0.6729	0.76%	0.99%	0.9643	0.57%	0.78%	0.9752	0.55%	0.76%
S2	0.9006	0.39%	0.55%	0.9894	0.29%	0.43%	0.9810	0.47%	0.66%
S3	0.9915	0.10%	0.16%	0.9986	0.11%	0.15%	0.9950	0.23%	0.34%
S4	0.9907	0.11%	0.17%	0.9983	0.12%	0.17%	0.9939	0.26%	0.37%
S5	0.9695	0.21%	0.30%	0.9957	0.19%	0.27%	0.9893	0.34%	0.50%
S6	0.9734	0.20%	0.28%	0.9978	0.15%	0.19%	0.9930	0.28%	0.40%
S7	0.9809	0.17%	0.24%	0.9981	0.13%	0.18%	0.9930	0.27%	0.40%
S8	0.8742	0.47%	0.61%	0.9893	0.32%	0.43%	0.9579	0.37%	0.99%
S9	0.9304	0.34%	0.46%	0.9928	0.26%	0.35%	0.9624	0.64%	0.93%
S10	0.6748	0.74%	0.99%	0.9781	0.46%	0.61%	0.9580	0.71%	0.98%
S11	0.9662	0.25%	0.32%	0.9965	0.16%	0.24%	0.9878	0.36%	0.53%

Table 7. Estimated SOH evaluation analyses of the three LIBs cells with the SVR.

interval	Cell1			Cell2			Cell3		
	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
S1	0.5901	0.87%	1.10%	0.9501	0.66%	0.91%	0.9629	0.70%	0.93%
S2	0.8840	0.46%	0.59%	0.9811	0.45%	0.57%	0.9758	0.57%	0.75%
S3	0.9876	0.15%	0.19%	0.9975	0.16%	0.21%	0.9907	0.32%	0.46%
S4	0.9864	0.16%	0.20%	0.9975	0.17%	0.21%	0.9902	0.33%	0.48%
S5	0.9591	0.25%	0.35%	0.9905	0.28%	0.40%	0.9846	0.42%	0.60%
S6	0.9744	0.22%	0.28%	0.9954	0.20%	0.28%	0.9886	0.34%	0.51%
S7	0.9832	0.17%	0.22%	0.9965	0.18%	0.24%	0.9882	0.34%	0.52%
S8	0.8416	0.55%	0.69%	0.9800	0.46%	0.58%	0.9618	0.69%	0.94%
S9	0.8909	0.46%	0.57%	0.9876	0.35%	0.46%	0.9651	0.66%	0.90%
S10	0.5995	0.80%	1.09%	0.9654	0.59%	0.77%	0.9465	0.84%	1.11%
S11	0.9617	0.29%	0.34%	0.9954	0.22%	0.28%	0.9845	0.43%	0.60%

Furthermore, the estimated battery SOH results of the absolute error with different charging current rates are shown in Figure 6. The box plots indicate that both the SVR and the RFR methods would experience increased prediction errors with high charging rates. However, the comparative results show that the SVR exhibits fewer outliers, but its interquartile range is wider compared with the RFR, indicating that the overall estimation error of the SVR is larger, while the corresponding variability of its errors might be smaller. These findings strongly support the argument presented in Section 3, which suggests that the SVR has stronger generalization ability, whereas the RFR possesses better fitting accuracy, respectively.

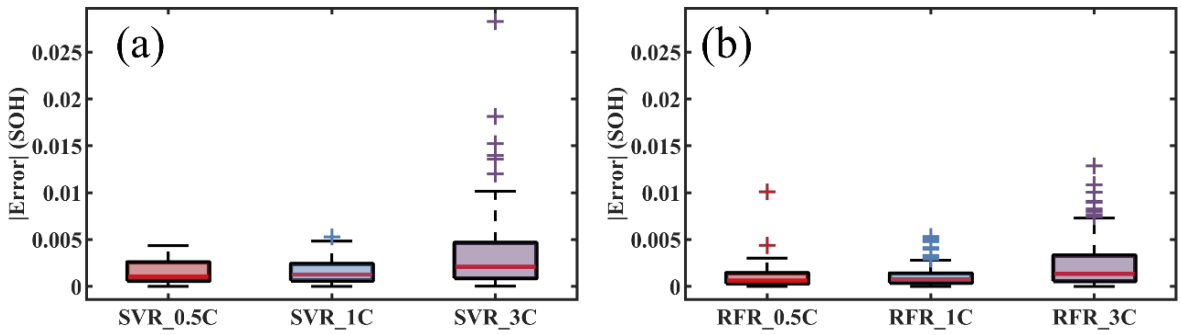

Figure 6. Absolute error analyses of the SOH estimation for the three LIBs based on box plots. (a) With the SVR. (b) With the RFR.

5. Conclusions

This study has successfully developed two data-driven methods for the SOH estimation of LIBs based on a selected charging voltage interval. At first, the IC curves and Pearson correlation analysis were used to select the optimal charging voltage interval. Then, two typical data-driven methods, including the RFR and the SVR, were employed for the SOH estimation of LIBs based on the extracted features from the selected charging voltage interval. The originality of this study is that the ICA and the Pearson correlation analysis are used as a bridge to select the optimal charging voltage interval, while the most relevant features for the battery SOH would be extracted from the selected charging voltage interval. In this case, the data-driven methods based on the most relevant features from the limited and optimal charging voltage interval can significantly improve computational efficiency and reduce the computational cost for the SOH estimation of LIBs. In addition, the accuracy for the battery SOH estimation can also be effectively guaranteed.

The comparative analyses between the simulation and experimental results were presented to verify the effectiveness and accuracy of the two data-driven SOH estimation methods based on the selected charging voltage interval. Both the SVR and the RFR methods demonstrated their respective advantages in the domain of the battery SOH estimation. Specifically, the SVR could produce fewer outliers in estimation errors, while the RFR could achieve higher overall estimation accuracy. With the RFR, the estimated values of R^2 for the three LIBs would be 0.9915, 0.9986, and 0.9950, respectively. Meanwhile, the MAE for the three LIBs would be 0.10%, 0.11%, and 0.23%, respectively. Furthermore, the RMSE for the three LIBs would be 0.16%, 0.15%, and 0.34%, respectively. As a comparison, with the SVR, the estimated values of R^2 for the three LIBs would be 0.9876, 0.9975, and 0.9907; the corresponding MAE would be 0.15%, 0.16%, and 0.32%; and the corresponding RMSE would be 0.19%, 0.21%, and 0.46%, respectively. Compared with other charging voltage intervals, the two data-driven methods with the selected charging voltage interval showed the best fitting performance for the SOH estimation of LIBs.

In summary, the selected optimal charging voltage interval exhibits the highest correlation for achieving SOH estimation compared with other charging voltage intervals. This study would provide a simple but effective way for estimating the SOH of LIBs, while the research findings would provide new references for the development and energy management optimization of battery energy storage systems in practical applications.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the China Electric Institute State Key Laboratory of Environmental Adaptability for Industrial Products (Grant No. 2024EASKJ-006), the Aeronautical Science Foundation of China (Grant No. 2024Z039070003), and the Key Research and Development Program of Shaanxi Province (Grant No. 2023-YBGY-376).

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Junguang Sun: Software, Writing—original draft, Formal analysis, Data curation, Visualization, Validation. Xiaodong Zhang: Conceptualization, Methodology, Investigation, Writing—review & editing. Wenrui Cao: Software, Formal analysis, Writing—review & editing. Lili Bo: Data curation, Visualization, Validation, Writing—review & editing. Changhai Liu: Data curation, Visualization, Validation, Writing—review & editing. Bin Wang: Resources, Supervision, Writing—review & editing, Funding acquisition.

References

1. Yu WH, Guo Y, Xu SM, et al. (2023) Comprehensive recycling of lithium-ion batteries: Fundamentals, pretreatment, and perspectives. *Energy Storage Mater* 54: 172–220. http://doi.org/10.1016/j.ensm.2022.10.033
2. Zhu AH, Bian XY, Han WJ, et al. (2023) The application of deep eutectic solvents in lithium-ion battery recycling: A comprehensive review. *Resour Conserv Recy* 188: 106690. http://doi.org/10.1016/j.resconrec.2022.106690
3. Naseri F, Gil S, Barbu C, et al. (2023) Digital twin of electric vehicle battery systems: Comprehensive review of the use cases, requirements, and platforms. *Renewable Sustainable Energy Rev* 179: 113280. http://doi.org/10.1016/j.rser.2023.113280
4. Che YH, Hu XS, Lin XK, et al. (2023) Health prognostics for lithium-ion batteries: Mechanisms, methods, and prospects. *Energy Environ Sci* 16: 338–371. http://doi.org/10.1039/d2ee03019e
5. Kong JZ, Yang FF, Zhang X, et al. (2021) Voltage-temperature health feature extraction to improve prognostics and health management of lithium-ion batteries. *Energy* 223: 120114. http://doi.org/10.1016/j.energy.2021.120114
6. Deng ZW, Xu L, Liu HA, et al. (2023) Prognostics of battery capacity based on charging data and data-driven methods for on-road vehicles. *Appl Energy* 339: 120954. http://doi.org/10.1016/j.apenergy.2023.120954

7. Wang JW, Deng ZW, Yu T, et al. (2022) State of health estimation based on modified Gaussian process regression for lithium-ion batteries. *J Energy Storage* 51: 104512. http://doi.org/10.1016/j.est.2022.104512
8. Geng CM, Zhang TH, Chen B, et al. (2023) Battery state of health estimation using GA-BP neural network on data feature mining. *Ieice Electron Expr* 20: 20230370. http://doi.org/10.1587/elex.20.20230370
9. Hu XS, Deng ZW, Lin XK, et al. (2021) Research directions for next-generation battery management solutions in automotive applications. *Renewable Sustainable Energy Rev* 152: 111695. http://doi.org/10.1016/j.rser.2021.111695
10. Zhang Y, Li YF (2022) Prognostics and health management of Lithium-ion battery using deep learning methods: A review. *Renewable Sustainable Energy Rev* 161: 112282. http://doi.org/10.1016/j.rser.2022.112282
11. He W, Williard N, Osterman M, et al. (2011) Prognostics of lithium-ion batteries based on Dempster-Shafer theory and the Bayesian Monte Carlo method. *J Power Sources* 196: 10314–10321. http://doi.org/10.1016/j.jpowsour.2011.08.040
12. Johnannes S, Stefan K, Madeleine E, et al. (2014) A holistic aging model for Li(NiMnCo)O_2 based 18650 lithium-ion batteries. *J Power Sources* 257: 325–334. http://doi.org/10.1016/j.jpowsour.2014.02.012
13. Zhu MY, Qian KF, Liu XT (2024) A three-time-scale dual extended Kalman filtering for parameter and state estimation of Li-ion battery. *Proc Inst Mech Eng D J Automob Eng* 238: 1352–1367. http://doi.org/10.1177/09544070231153440
14. Sun JG, Du R, Chen LX, et al. (2024) State of charge estimation for power batteries in new energy vehicles considering temperature fluctuations. *J Xi'an Jiaotong Univ* 58: 39–51. http://doi.org/10.7652/xjtuxb202411004
15. Ge B, Luo Y, Li LX, et al. (2023) Real-time state estimation of lithium batteries used for energy storage in electric vehicle charging stations with wind-solar complementary power system. *J Xi'an Jiaotong Univ* 57: 55–65.http://doi.org/10.7652/xjtuxb202301006
16. Ji SL, Zhang ZS, Stein HS, et al. (2025) Flexible health prognosis of battery nonlinear aging using temporal transfer learning. *Appl Energy* 377: 124766. http://doi.org/10.1016/j.apenergy.2024.124766
17. Zheng K, Meng JH, Yang ZP, et al. (2024) Refined lithium-ion battery state of health estimation with charging segment adjustment. *Appl Energy* 375: 124077. http://doi.org/10.1016/j.apenergy.2024.124077
18. Tang AH, Xu YC, Hu YZ, et al. (2024) Battery state of health estimation under dynamic operations with physics-driven deep learning. *Appl Energy* 370: 123632. http://doi.org/10.1016/j.apenergy.2024.123632
19. Wen JP, Chen X, Li XH, et al. (2022) SOH prediction of lithium battery based on IC curve feature and BP network. *Energy* 261: 125234. https://doi.org/10.1016/j.energy.2022.125234
20. Ren Y, Tang T, Jiang FS, et al. (2025) A novel state of health estimation method for lithium-ion battery pack based on cross generative adversarial networks. *Appl Energy* 377: 124385. https://doi.org/10.1016/j.apenergy.2024.124385
21. Li XB, Fan DQ, Liu XT, et al. (2024) State of health estimation for lithium-ion batteries based on improved bat algorithm optimation kernel extreme learning machine. *J Energy Storage* 101: 113756. https://doi.org/10.1016/j.est.2024.113756

22. Liu JZ, Liu XT (2023) An improved method of state of health prediction for lithium batteries considering different temperature. *J Energy Storage* 63: 107028. https://doi.org/10.1016/j.est.2023.107028
23. Ren Y, Tang T, Xia Q, et al. (2024) A data and physical model joint driven method for lithium-ion battery remaining useful life prediction under complex dynamic conditions. *J Energy Storage* 79: 110065. https://doi.org/10.1016/j.est.2023.110065
24. Ruan HK, He HW, Wei ZB, et al. (2023) State of health estimation of lithium-ion battery based on constant-voltage charging reconstruction. *IEEE J Em Sel Top Power Electron* 11: 4393–4402. http://doi.org/10.1109/jestpe.2021.3098836
25. Edge JS, O'Kane S, Prosser R, et al. (2021) Lithium ion battery degradation: what you need to know. *Phys Chem Chem Phys* 23: 8200–8221. http://doi.org/10.1039/d1cp00359c
26. Lin CP, Xu J, Jiang DL, et al. (2024) A comparative study of data-driven battery capacity estimation based on partial charging curves. *J Energy Chem* 88: 409–420. http://doi.org/10.1016/j.jechem.2023.09.025
27. Xiong R, Sun Y, Wang CX, et al. (2023) A data-driven method for extracting aging features to accurately predict the battery health. *Energy Storage Mater* 57: 460–470. http://doi.org/10.1016/j.ensm.2023.02.034
28. Li KQ, Wang YJ, Chen ZH (2022) A comparative study of battery state-of-health estimation based on empirical mode decomposition and neural network. *J Energy Storage* 54: 105333. http://doi.org/10.1016/j.est.2022.105333
29. Li Y, Abdel-Monem M, Gopalakrishnan R, et al. (2018) A quick on-line state of health estimation method for Li-ion battery with incremental capacity curves processed by Gaussian filter. *J Power Sources* 373: 40–53. http://doi.org/10.1016/j.jpowsour.2017.10.092
30. Lin CP, Xu J, Shi MJ, et al. (2022) Constant current charging time based fast state-of-health estimation for lithium-ion batteries. *Energy* 247: 123556. http://doi.org/10.1016/j.energy.2022.123556
31. Bian XL, Wei ZG, Li WH, et al. (2022) State-of-Health estimation of lithium-ion batteries by fusing an open circuit voltage model and incremental capacity analysis. *IEEE Trans Power Electr* 37: 2226–2236. http://doi.org/10.1109/tpel.2021.3104723
32. Li XY, Wang ZP, Zhang L, et al. (2019) State-of-health estimation for Li-ion batteries by combing the incremental capacity analysis method with grey relational analysis. *J Power Sources* 410: 106–114. http://doi.org/10.1016/j.jpowsour.2018.10.069
33. Lin MQ, Yan CH, Wang W, et al. (2023) A data-driven approach for estimating state-of-health of lithium-ion batteries considering internal resistance. *Energy* 277: 127675. http://doi.org/10.1016/j.energy.2023.127675
34. Yan YZ, Wang B, Wang CH, et al. (2024) Adaptive maximum available energy evaluation for lithium battery in hydrogen-electirc hybrid unmanned aerial vehicle applications considering dynamic ambient temperature and aging level. *Energy Convers Manage* 314: 118685. https://doi.org/10.1016/j.enconman.2024.118685
35. Peng KL, Deng ZW, Bao ZB, et al. (2023) Data-driven battery capacity estimation based on partial discharging capacity curve for lithium-ion batteries. *J Energy Storage* 67: 107549. http://doi.org/10.1016/j.est.2023.107549

36. Pan WJ, Luo XS, Zhu MT, et al. (2021) A health indicator extraction and optimization for capacity estimation of Li-ion battery using incremental capacity curves. *J Energy Storage* 42: 103072. http://doi.org/10.1016/j.est.2021.103072
37. Deng ZW, Xu L, Liu HA, et al. (2024) Rapid health estimation of in-service battery packs based on limited labels and domain adaptation. *J Energy Chem* 89: 345–354. https://doi.org/10.1016/j.jechem.2023.10.056
38. Liu KL, Li Y, Hu XS, et al. (2020) Gaussian process regression with automatic relevance determination kernel for calendar aging prediction of lithium-ion batteries. *IEEE Trans Ind Inform* 16: 3767–3777. http://doi.org/10.1109/tii.2019.2941747
39. Lin C, Xu J, Hou J, et al. (2023) A fast data-driven battery capacity estimation method under non-constant current charging and variable temperature. *Energy Storage Mater* 63: 102967. http://doi.org/10.1016/j.ensm.2023.102967
40. Sattianadan D, Sharma RK, Fernandez SG, et al. (2024) Estimation of state of charge and state of health of batteries using hybrid method and recurrent neural network. *AIP Conf Proc* 3037: 020016–020028. http://doi.org/10.1063/5.0196476
41. Wang FJ, Zhai Z, Liu BC, et al. (2024) Open access dataset, code library and benchmarking deep learning approaches for state-of-health estimation of lithium-ion batteries. *J Energy Storage* 77: 109884. http://doi.org/10.1016/j.est.2023.109884
42. Zhang SX, Liu ZT, Su HY (2023) State of health estimation for lithium-ion batteries on few-shot learning. *Energy* 268: 126726. http://doi.org/10.1016/j.energy.2023.126726

AIMS Press