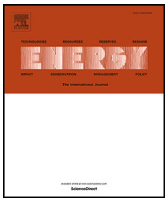
Contents lists available at ScienceDirect

Energy

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



State of health prediction of lithium-ion batteries based on autoregression with exogenous variables model

Zhelin Huang [a], Fan Xu [b,*], Fangfang Yang [c]

[a] *Department of Statistics, College of Economics, Shenzhen University, Shenzhen, China*
[b] *School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan, China*
[c] *School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, China*

ARTICLE INFO

Keywords:
AREV model
Feature extraction
State of health prediction
Lithium-ion battery
Data-driven method

ABSTRACT

The gradually decreasing capacity of lithium-ion batteries can serve as a health indicator for tracking their degradation. Therefore, it is important to predict the capacity of future cycles to assess the health condition of lithium-ion batteries. According to electrochemical theory and the characteristics of the data curves, this paper proposes several ideas for feature extraction. A novel fusion prognostic framework is proposed, in which a data-driven time series prediction model is adopted and combined with extracted features for lithium-ion battery capacity prediction. The proposed method is based on an autoregression with an exogenous-variable model that can self-adaptively update at each cycle and then predict the state of health in the next cycle and cycles in the near future. Under the assumption that the historical capacity data is available, the experimental results showed that by using the proposed autoregression with exogenous variables model, the root mean square error, mean absolute error, and mean absolute percentage error of the prediction results were 0.000963, 0.000562, and 0.000584, respectively, which indicated that the prediction results were precise.

1. Introduction

Lithium-ion batteries, because of their high energy densities, high galvanic potentials, wide temperature ranges, low self-discharge rates, and long lifetimes, are core components in a wide variety of systems. Therefore, the reliability of lithium-ion batteries has become a subject of great interest to the electronics industry. Safety management, charging and discharging control, performance degradation, capacity fade, and remaining useful life (RUL) estimation of lithium ion batteries have become important and challenging issues in the fields of reliability engineering, automatic testing, power sources, and electric vehicles. As a result, lithium-ion battery state of health (SOH) has become an important issue in the prognostics and health management (PHM) of electronics [1,2]. Prognostics and RUL estimation entail the use of the current and previous system states to predict the future states of a battery system. Reliable predicted information can be used to schedule repairs and maintenance in advance and provide an alarm before faults reach critical levels to prevent performance degradation, malfunction, or even catastrophic failures [3,4].

1.1. Literature review

The various approaches for battery SOH estimation can be generally classified into two categories: model-based approaches and data-driven approaches [5,6]. Model-based methods usually include establishing physical models to describe the physical process of the system state and fault evolution. However, model-based methods may not be suitable for many practical applications, in which the physical parameters may vary under various operating conditions [7]. Moreover, it is difficult to obtain an appropriate physical model to describe the dynamic characteristics of the system, and model-based approaches cannot be applied for those complex systems in which the internal state variables are inaccessible to direct testing and monitoring with general sensors.

Importantly, data-driven methods have attracted increasing attention owing to the increasing availability of a large amount of battery data. Furthermore, data-driven methods have attracted extensive interest because of their high precision and model-free characteristics. In short, the data-driven methods show advantages for battery prediction in the era of big data, and some new research progress has been made. For example, Xu et al. [8] used a multiscale dual extended Kalman filter to estimate the battery state. Ouyang et al. [9] applied a Gaussian linear model based on six commonly used open-circuit-voltage parameters to estimate the SOH. Shen et al. [10] proposed a deep convolutional neural network transfer learning algorithm that took the capacity, voltage, and current information as inputs. Liu et al. [11] introduced a new energy-based health index based on the voltage curve

[*] Corresponding author.
E-mail address: xufanfanxu@wuht.edu.cn (F. Xu).

https://doi.org/10.1016/j.energy.2022.125497
Received 19 May 2022; Received in revised form 1 August 2022; Accepted 18 September 2022
Available online 27 September 2022

of the battery; then, they designed an online SOH estimation framework based on a limit learning machine. Shu et al. [12] developed a fixed-size least squares support vector machine method to estimate the SOH using the charging time within a predefined voltage range. In recent years, a new data-driven method based on Gaussian process regression (GPR) has emerged. This Bayesian nonparametric probability method exhibits good performances in nonlinear mapping, and it can deal with problems such as high nonlinearity [13], battery degradation modeling, and a small sample and prediction uncertainty representation for PHM. Richardson et al. [14] automatically extracted the time value between equal width voltage points as the input and constructed a GPR model for field capacity estimation. Liu et al. [15] considered the temperature and discharge depth, modified the covariance function, and proposed two innovative models based on the GPR.

However, compared to model-based methods, data-driven methods rely more on large-scale data. Therefore, it is very important to broaden the range of battery features and include more extractable features to establish a data model and improve the accuracy of the prediction. After 2019, the multi-feature extraction scheme proposed by Severson et al. [16] attracted extensive attention. The general features in the recent literature were extracted based on the number of cycles or the voltage curve [17]. Incremental capacity (IC) analysis and differential voltage (DV) analysis are two electrochemical techniques recently introduced for battery diagnosis and prediction [18]. Various features have been extracted from the IC and DV curves, such as the strength, position, and area under the curves, and it was verified that they were highly correlated with battery degradation [19]. For example, Tang et al. [20] used the regional capacity derived from the IC peak during a constant-current charging process as the feature and employed a linear regression model for SOH estimation. Pei et al. [21] used the partially charged electric quantity derived from IC analysis as the feature and employed a linear model for capacity estimation. Li et al. [22] used IC values within a specific voltage interval as the features and employed a Gaussian process regression model for SOH estimation and RUL prediction. Li et al. [23] used the area, position, and height of the second IC peak as the features and employed a support vector machine for SOH estimation. Furthermore, to describe such variations, the origins of the electrochemical activation in lithium-ion batteries were introduced, which could help to understand the battery degradation process from the perspective of electrochemistry and further extract the corresponding health features [24–26].

There are many ways to apply data-driven technology, each of which is equivalent to different assumptions about the nature of the underlying process. One of the most common and simplest method is to use direct mapping from cycle to SOH [13]. This is equivalent to fitting a curve with the capacity cycle data and then predicting the future value by extrapolating the fitted curve. This means that accurate capacity data for the first few cycles of the battery life can be obtained.

However, this paper mainly focuses on capacity prediction, that is, estimating the future values of battery capacity. Therefore, we assume that the historical data of the capacity cycle is available. In practice, these data can be obtained by direct measurement (low-speed charge–discharge cycle specially used for capacity measurement at periodic intervals) or by various other techniques, which can avoid interfering with the system. However, the mapping from cycle to SOH is too simple because the battery capacity depends on various factors, and historical capacity data alone is unlikely to be sufficient to predict the future capacity. Moreover, it can be reasonably expected that there is a certain correlation between the previous capacity and the future capacity. Therefore, using a time series model, we can explore its ability to predict through historical data. In addition, the method applied to the capacity and cycle data can then be applied to inputs with a greater amount of information from the monitoring data, such as the current, voltage, time, and impedance. To take time series information and feature information together into consideration, an autoregression with exogenous variables (AREV) model is proposed in this paper. Moreover, in the feature extraction of external variables, we proposed several ideas based on the principles of electrochemistry and the observation of the IC and other related curves.

1.2. Our contributions

The methodology presented here produced robust and highly accurate SOH predictions. In addition, we also provide a theoretical framework. The method combines causality parameters through a hidden Markov model. The proposed AREV model achieves the following objectives:

1. When the sensor connected to the battery receives new available information, it dynamically combines this information. Therefore, we named our model AREV, which represents autoregression with battery SOH data combined with exogenous feature data extracted from other measurable battery data. AREV also uses L1 (and possibly L2) regularization to automatically select the most relevant information.
2. The latest battery degradation changes are dynamically captured using a 30-cycle moving window (immediately before the expected prediction date) during training.
3. In addition, for the problem of feature extraction of lithium-ion battery data, referring to a method proposed previously [16] and combined with the principles of electrochemistry, we propose several possible ideas to extract the features of lithium-ion batteries. Specific extraction ideas will be discussed in the second section of the paper.

The experimental results showed that the proposed method achieved great prediction results. Moreover, we verified the idea of model construction in this paper; that is, the historical data of the capacitance and other features extracted based on electrochemical analysis and curve analysis played a role in the capacitance prediction of the next cycle.

1.3. Organization of paper

The rest of this paper is organized as follows. The data and the proposed feature extraction ideas are discussed in Section 2. Technical details of the AREV model are introduced in Section 3. The experimental results and comparative studies are reported and discussed in Section 4, and our conclusions are drawn in Section 5.

2. Data

2.1. Data description

The dataset used in this study was presented previously [16]. A Massachusetts Institute of Technology (MIT) team completed experiments to acquire this dataset (therefore, it is referred to as the MIT dataset in this paper). Our method used this data set unless otherwise stated below. Due to the electrochemical mechanism and manufacturing variability of the capacity fade of lithium-ion batteries, the capacity decay is expected to be observed from multiple dimensions. The cells have a nominal capacity of 1.1 Ah and a nominal voltage of 3.3 V. In order to explore the fade process, commercial $LiFePO_4$ (LFP)/graphite cells manufactured by A123 Systems (APR18650M1 A) were cycled in horizontal cylindrical fixtures on a 48-channel Arbin LBT potentiostat in a forced-convection temperature-controlled environmental chamber (30 °C) under various fast-charging conditions but identical discharging conditions (4C to 2.0 V, where 1C was 1.1 A. While the chamber temperature was controlled, the cell temperatures varied by up to 10 (30 °C) within a cycle due to the large amount of heat generated during charge and discharge. This temperature variation was a function of the internal impedance and charging policy (supplementary information is available for this paper at https://doi.org/10.1038/s41560-019-0356-8). Because the graphite negative electrode dominated the degradation in these batteries, these results could be useful for other lithium-ion batteries based on graphite. By deliberately varying the charging conditions, the MIT team generated a dataset that captured a wide

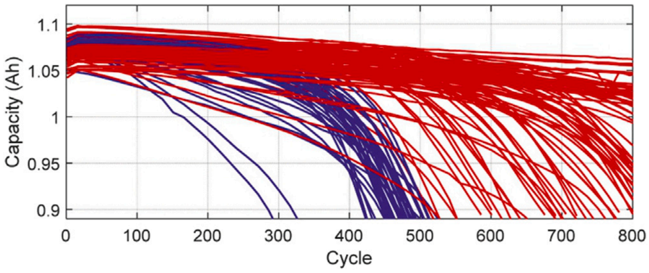


Fig. 1. Capacity curves of 123 batteries.

range of cycle lives, from approximately 150 to 2300 cycles (average cycle life of 806 with a standard deviation of 377).

The voltage, current, cell temperature, and internal resistance were continuously measured during cycling. The dataset contained approximately 96,700 cycles, and thus, it is the largest publicly available dataset for nominally identical commercial lithium-ion batteries cycled under controlled conditions. Fig. 1 shows the discharge capacity as a function of cycle number for the first 800 cycles. The capacity fade was negligible in the first 100 cycles and accelerated near the end of life, as is often observed in lithium-ion batteries (this dataset is available at https://data.matr.io/1).

2.2. Feature extraction

A set of battery features extracted from the temperature, voltage, and current information were adopted as degradation indicators. These indicators were obtained by applying signal processing methods, such as incremental capacity analysis, which are thought to be useful for evaluating battery health.

Previous battery degradation studies have shown that the loss of active material (LAM), the loss of lithium inventory (LLI), and the increase in the internal resistance (IR) are the three main processes that cause battery degradation [27]. Therefore, we look for possible health features from these perspectives.

2.2.1. Features from incremental capacity (IC) curve

The characteristics of the IC curve can indicate the internal electrochemical reactions of a battery and reflect the aging degree. Therefore, the battery health characteristics can be extracted from the IC curve [28]. The IC curve can be calculated by the following equation. Note that discharge capacity is proportional to the discharging time because of constant discharging current. The following relationships between the discharging voltage and capacity can be acquired:

$$V = f(Q), Q = f^{-1}(V), \tag{1}$$

$$(f^{-1})' = \frac{dQ}{dV} = \frac{I * dt}{dV} = I * \frac{dt}{dV}, \tag{2}$$

where V and Q are the discharge voltage and capacity under the constant current level, and there is a certain functional relationship between them, expressed by f, I is the discharge current, and t is the discharge time. Eq. (2) is regarded as an IC curve and is derived from differential calculus. The incremental capacity curve describes the relationship between a voltage change and a capacity change ($\triangle\ Q/\ \triangle\ V$) during a discharge process. Although the battery has a large charge and discharge current when used in the vehicle, as shown previously [29,30], the peak value on the IC curve can still be identified, which reveals the important characteristics of the battery health based on normal charge and discharge data. A study on the relationship between the coulomb efficiency and the capacity degradation of commercial lithium ion batteries was reported previously [31]. By observing the gradual development of the IC curve peak in the whole life cycle, we can understand the aging mechanism of the battery.

Fig. 2 illustrates the change of the IC peak value during battery degradation. Fig. 2(a) shows the data obtained from our previous study of the relationship between the coulombic efficiency and the capacity degradation of commercial lithium ion batteries [18]. To better show the change of the battery, this dataset was obtained by discharging at 1C and used in our previous research to illustrate the change process of the IC curve during battery aging. As shown in Fig. 2(a), there were three peaks on the IC curve of the fresh $LiFePO_4$ (LFP) battery. Fig. 2(a) also shows the relationship between the peak voltage position and the SOH. A disproportionate decrease in the intensity of peak 1 was observed between the fresh cell and the cell with a 90% SOH, which is expected to be caused by the loss of lithium inventory (LLI) process. In contrast, the intensities of peaks 2 and 3 showed little change, indicating that there was no significant loss of active material in the LFP batteries. The positions of all three peaks moved from 90% SOH to 80% SOH at a lower voltage.

However, in the MIT dataset that we used in this study, the data were obtained in 4C discharge mode. Because the current was too large, only one peak appeared on its IC curve, as shown in Fig. 2(b). According to the above analysis, it was still feasible to use the peak value in the IC curve as the data feature to predict the SOH value. Therefore, extracting features from the peak movement of the curve is proposed. In addition, the LAM factors can also be identified by the unbalanced decrease in the peak intensity (Y-axis of the IC curve), the decrease in the peak intensity ratio, and the offset of the peak voltage position (x-axis of the IC curve). On the IC curve, the maximum values of the abscissa and ordinate were calculated to obtain F1 and F2, respectively. Therefore, we propose F1 and F2 as key features for this work.

2.2.2. Features from discharge voltage curve

The loss of lithium inventory can also be observed on the change of discharge voltage curve. To capture the electrochemical evolution of a single cell during the cycle, several characteristics were calculated from the discharge voltage curve. Specifically, we consider the period-to-period evolution of $Q(V)$ and the discharge voltage curve as functions of the voltage for a given period, as shown in Fig. 3. Because the voltage range was the same for each cycle, we considered the capacity as a function of the voltage as the basis of the comparison cycle, and therefore, proposed feature F3, F4 and F5.

2.2.3. Features from internal resistance (IR) curve

The IR is obtained by applying 10 consecutive pulses to the battery, averaging the voltage drop of the pulse, and then dividing by the size of the current pulse. For the cycle life test of commercial LFP batteries, the increase in the IR is the main source of battery aging. The IR values of two LFP batteries with two different life cycles are shown in Fig. 4 (with 1748 and 426 charge–discharge cycles). The fluctuation rates of the two had significant differences in the early stage of battery life (the first 250 charge discharge cycles), which could be used in model construction and life prediction. Therefore, extracting features by mining and comparing the changes of the volatility of other features in different stages is proposed, as well as using F6 as a feature.

2.2.4. Features from temperature curve

Temperature is a key factor that may affect the decline of the battery performance. During charging and discharging, the battery temperature may fluctuate due to the switching of the charging/discharging strategy, internal chemical reactions, and environmental factors. From the perspective of electrochemistry, the temperature change has an impact on the internal ion mobility and electrolyte conductivity of the battery, which will affect the aging of the battery. Therefore, in our work, thermal factors are considered in degradation modeling. In general, thermal factors can be considered indirectly and directly. The indirect method is to construct temperature-related parameters in the degradation model. However, this method makes the model more

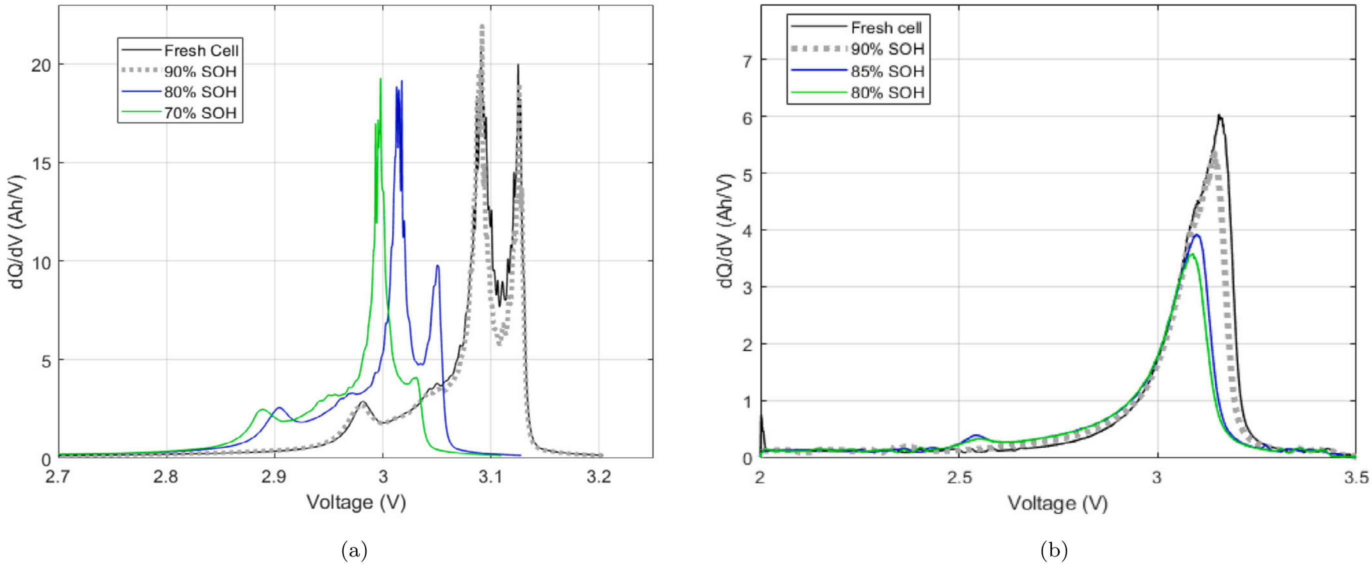


Fig. 2. Incremental capacity (IC) curves of an $LiFePO_4$ (LFP) battery: (a) battery discharge at 1C; (b) battery discharge at 4C.

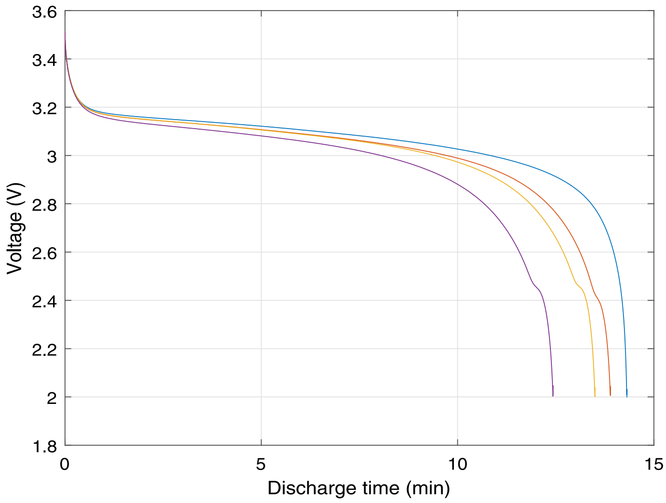


Fig. 3. Discharge voltage curves for battery in four different cycles. (MIT dataset).

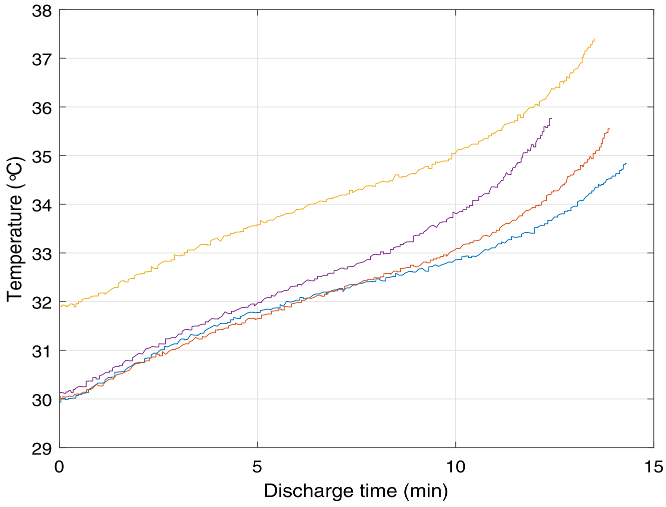


Fig. 5. Discharge temperature for battery in four different cycles. (MIT dataset).

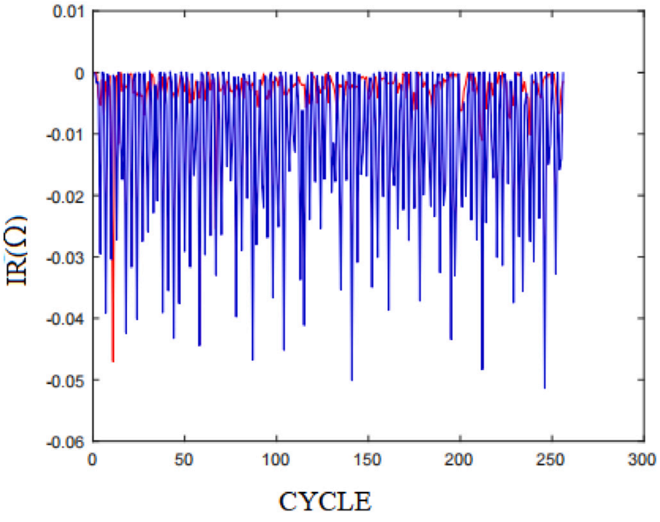


Fig. 4. Internal resistance (IR) values of two LFP batteries. Red line: battery with 1748 lifecycles; blue line: battery with 426 lifecycles.

complex and easy to over fit, and increases the computational burden of the parameter estimation. In addition, the effectiveness of this method depends on the construction method of the temperature-related parameters and their sensitivity to battery degradation. In the modeling process, the construction of parameters is still a difficult problem. In view of the above shortcomings of the direct method, the direct method was adopted in this work, which directly uses the temperature or simple temperature statistics as the health features. This method can not only consider the thermal factors in the degradation modeling, but also keep the model simple. To comprehensively consider the influencing factors of battery estimation and the prediction model, we included a thermal correlation measurement as an additional feature in the health feature set. Specifically, in this study, we selected the average surface temperature of the charge and discharge processes in each cycle as the thermally dependent features F7 and F8, respectively. From Fig. 5, we can see that the general trend of the temperature fluctuated with the increase in the number of cycles.

Table 1
Extracted features.

Notation	Details of extracted features
F1	$\frac{dQ}{dV}_{max}$, y-axis maximum values on the IC curve.
F2	$\frac{dQ}{dV}_{max}(V)$, x-axis of maximum values on the IC curve.
F3	$mean(\triangle Q)$, mean of the difference of Q values in adjacent cycles.
F4	$var(\triangle Q)$, variance of the difference of Q values in adjacent cycles.
F5	$min(\triangle Q)$, minimum value of the difference of Q values in adjacent cycles.
F6	IR, internal resistance.
F7	T_{max}, maximum temperature of the cell in each cycle.
F8	T_{min}, minimum temperature of the cell in each cycle.

2.3. Summary of extracted features

Therefore, in this work, eight different features were extracted, as shown in Table 1. As described in Section 3.2, in the process of model selection and updating, we adaptively selected the appropriate features through the optimization process of the lasso algorithm. Due to the characteristics of the algorithm, if some features did not play a role in the prediction process, their parameters would be adjusted to 0 in the training process, that is, the prediction model obtained from the training would not contain this feature. Therefore, we no longer artificially perform feature selection in the method presented in Section 2.

3. Battery state of health (SOH) prediction based on autoregressive with exogenous variables (AREV) model

3.1. AREV model

The proposed AREV model was first introduced for an influenza infection prediction problem. Based on the influenza data in the past few weeks (autoregressive term) and the search volume of influenza related keywords in Internet searches (external variables), the number of influenza infections in the next week was predicted. The prediction effect of this model was good, and the influence of the autoregressive term and external variables could be considered at the same time. Because the composition of this dataset was similar to the SOH value prediction process described in this paper, a combination of the two was used to form the AREV model.

The proposed AREV model was motivated by a hidden Markov model. The capacity value $\{y_t\}$ is the intrinsic time series of interest. We impose an autoregressive model with lag n, which implies that the collection of vectors $\{y_{(t-n+1):t}\}_{t\geq n}$ is a Markov chain (this captures the fact that the degradation process of the battery was gradual, and its SOH value was related to the SOH value of the previous cycles in a short period of time).

The vector of extracted features at a specific time, denoted as X_t, depends only on the battery SOH at the same time. The Markovian property on the block $y_{(t-n+1):t}$ leads to the (vector) hidden Markov model structure. Therefore, the mathematical assumptions of the proposed model are as follows:

Assumption 1.

$$y_t = \mu_y + \sum_{i=1}^{n} \alpha_j y_{t-j} + \epsilon_t,$$

where ϵ_t represents i.i.d. Gaussian variables.

Assumption 2.

$$X_t | y_t \sim N(\mu_x + y_t\beta, Q).$$

Assumption 3. Conditional on y_t, X_t is independent of $\{y_l, X_l : l \neq t\}$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_k)^T$, $\mu_x = (\mu_{x_1}, \mu_{x_2}, \ldots, \mu_{x_k})^T$, and Q is the covariance matrix.

Autoregression with exogenous variables model

We consider the following autoregression with exogenous variables:

$$y_{t+1} = \mu_y + \sum_{i=0}^{n} \beta_i y_{t-i} + \sum_{l=1}^{F} \eta_l, X_{l,t} + \epsilon_t, \tag{3}$$

where $\{y_t\}$ is the SOH value, and $\{X_{l,t}\}$ represents the extracted features from the raw battery data, μ_y is the average level of y, and $\epsilon_t \sim N(0, \sigma^2)$.

3.2. Model update

We chose $n = 3$ to capture the time series information of the capacity in the past four cycles, and $F = 8$ (eight other features we extracted from the battery data). Then, we imposed regularities for parameter estimation. In general, we have three kinds of penalties: the L_1 penalty, the L_2 penalty, and a linear combination of the L_1 and L_2 penalties.

An estimate of the parameters can be obtained by solving the following optimization problem:

$$argmin_{\mu_y,\beta,\eta} \sum_t (y_{t+1} - \mu_y - \sum_{i=0}^{n} \beta_i y_{t-i} - \sum_{l=1}^{F} \eta_l X_{l,t})^2 + \lambda_\beta \|\beta\|_1 + \gamma_\beta |\beta|^2 + \lambda_\beta \|\eta\|_1 + \gamma_\beta |\eta|^2, \tag{4}$$

where λ_α, λ_β, η_α, and η_β are hyperparameters.

In our implementation, we used a grid search procedure to determine the values of the hyperparameters λ_β, γ_β, λ_η, and γ_η. Ideally, we would like to use cross-validation to select all four hyperparameters. In each grid where all the hyperparameters were fixed, we used the quadprog package in MATLAB to solve the quadratic optimization problem. However, because we had only 30 training data points for a given battery, the cross-validation result was highly noisy. Thus, we needed to prespecify some of the hyperparameters. For model simplicity and sparsity, combined with the evidence from the cross-validation, we set $\eta_\alpha = \eta_\beta = 0$, leading to L1 penalization on both the autoregressive terms and the extracted features. With the remaining λ_α and λ_β, the cross-validation results still had considerable variance. By the same sparsity and simplicity considerations, we further constrained $\lambda_\alpha = \lambda_\beta$. Therefore, the AREV model we finally proposed is Eq. (3) with the constraints that $\eta_\alpha = \eta_\beta = 0$ and $\lambda_\alpha = \lambda_\beta$. The flowchart of the proposed methodology is shown in Fig. 6.

3.3. Training and testing process

The goal of a regression problem is to learn the mapping from the inputs x to the outputs y, given a labeled training set of input–output pairs $\{x_i, y_i\}_{i=1}^{m}$, where m is the number of training examples. In our case, the input x_i was the eight extracted features from the batteries combined with SOH values in the past n cycles. The learned model could then be used to make predictions at the test indices. Because there were many kinds of lithium batteries for this dataset and the operating conditions of the different batteries were different, the applicability of the model could be guaranteed by using the historical data of the same battery for training.

Because the objective of this study was to predict the SOH values in future cycles, we assumed that the SOH values of past cycles could be obtained. At a given cycle k, the goal was to predict the SOH value of cycle $k + 1$. In this paper, we set $m = 30$, $n = 3$, and $F = 8$. The training and testing processes were as follows:

Step 1: $(k - m, k - m - 1, \ldots .k - 1)$ was selected as a sliding window, which immediately preceded the cycle to be predicted, cycle k, for the training cycle to capture the most recent changes in the battery degradation mode and the time series behavior. The sliding window slid as cycle k increased.

Step 2: The SOH values of cycles $(k-n, k-n+1, \ldots, k-1)$ and the eight external feature variable values of cycle $k-1$ were selected as the input

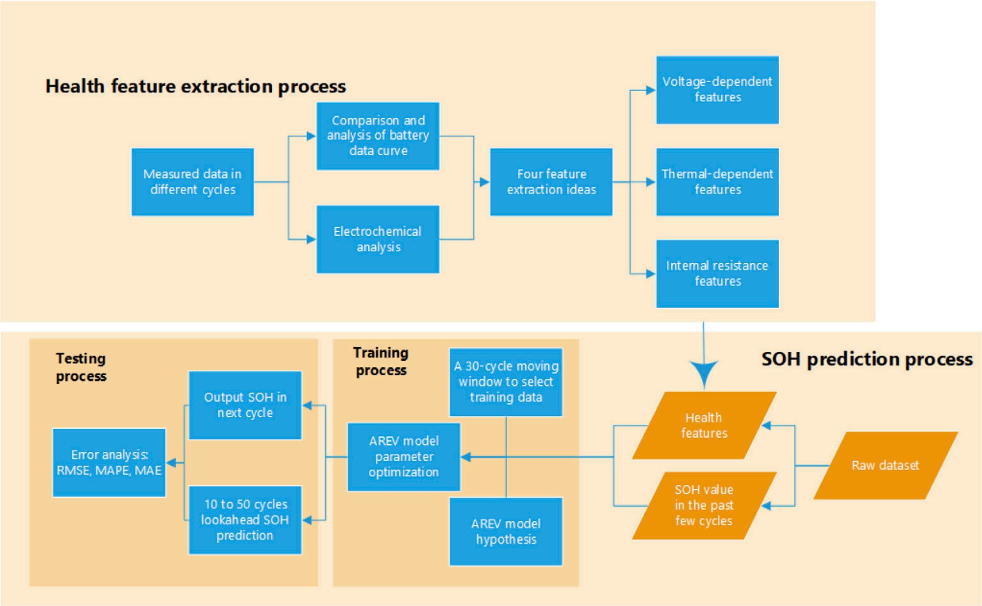

Health feature extraction process
Measured data in different cycles
Comparison and analysis of battery data curve
Electrochemical analysis
Four feature extraction ideas
Voltage-dependent features
Thermal-dependent features
Internal resistance features
SOH prediction process
Testing process
Training process
Output SOH in next cycle
Error analysis: RMSE, MAPE, MAE
10 to 50 cycles lookahead SOH prediction
A 30-cycle moving window to select training data
AREV model parameter optimization
AREV model hypothesis
Health features
SOH value in the past few cycles
Raw dataset

Fig. 6. Flowchart of the proposed methodology.

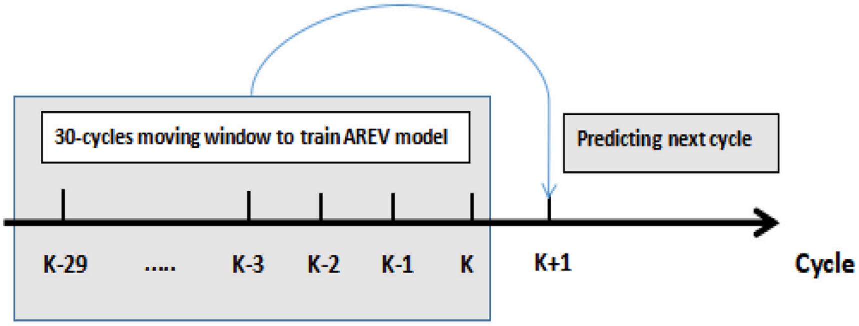


Fig. 7. Explanatory diagram of step 2.

data, and the SOH value of cycle k was selected as the output data. The above input and output formed the first group of training data. The SOH values of cycles $(k-n-1, k-n, \ldots, k-2)$ and the eight external feature variable values of cycle $k-2$ were selected as the input, and the SOH value of cycle $k-1$ was selected as the output. The above input and output formed the second group of training data. … The SOH values of cycles $(k-m-n+1, k-m-n+2, \ldots, k-m)$ and the eight external feature variable values of cycle $k-m$ were selected as the input, and the SOH value of cycle $k-m+1$ was selected as the output. The above input and output formed the thirtieth group of training data. The above historical data were obtained at the current time point, that is, at the end of the kth cycle. The AREV model was trained using the above data (a total m groups of training data) to obtain its model parameters (α and β in Eq. (3)).

Step 3: A test was performed on cycle $k+1$. The trained model with known parameters was used for the prediction of the next cycle, that is, the SOH values of $(k-n+1, k-n+2, \ldots, k)$ and the values of eight external feature variable values of cycle k were selected as inputs, and they were introduced into the trained AREV equation (Eq. (3)) to obtain the SOH prediction value of cycle $k+1$.

Step 4: $k = k+1$, and the above steps were repeated.

In step 2, the selection of training data, that is, the moving window of 30-cycles, is shown in Fig. 7.

3.4. Evaluation metrics

The root mean square error (RMSE), mean absolute error (MAE), and mean absolute percent error (MAPE) were the three evaluation metrics used to evaluate the prediction accuracy of the proposed method. They are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}-y)^2} \tag{5}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}-y| \tag{6}$$

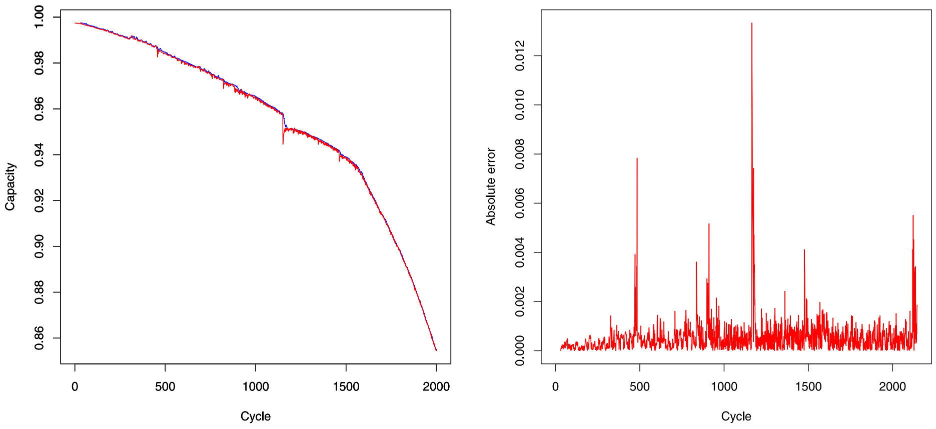
$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}-y}{y}\right| \tag{7}$$

where y is the real value of the capacity, $\hat{y}$ is the one-step-ahead predicted value of the capacity, and N is the number of time points of the prediction period.

4. Experimental results and discussion

4.1. SOH prediction

According to the process described in Section 3.3, first train the model parameters of AREV model through the past 30 groups of training data, and then use the obtained model with known parameters



(a) (b)

Fig. 8. Prediction results using proposed autoregressive with exogenous variables (AREV) model. (a) Red line: real value, blue line: prediction results; (b) absolute error.

Table 2
Summary of experimental errors of state of health (SOH) prediction.

Model	RMSE	MAPE	MAE
AREV	0.000963	0.000562	0.000584
Autoregressive model	0.000991	0.000571	0.000593
Model with only exogenous variables	0.001253	0.000717	0.000744
SVM with all variables	0.001017	0.000640	0.000612

Table 3
Prediction results of capacity in last 100 cycles with different look-ahead periods.

Look-ahead period	RMSE	MAPE	MAE
10 cycles	0.000548	0.000432	0.000513
30 cycles	0.000912	0.000517	0.000622
50 cycles	0.001116	0.000659	0.000799

to predict the SOH value of the next cycle. The output prediction results can be seen in Fig. 8. The prediction result is relatively accurate, as shown in Table 2, and its RMSE is 0.000963, MAE is 0.000562, MAPE is 0.000584. Because the model proposed in this paper is a combination of AR model and external variables, in the next section, we further compare the difference in prediction accuracy between the original model (AREV) and the individual AR model, as well as the model that only considers external variables.

4.2. Comparative study

The preliminary analysis of the battery health state showed the current capacity value was not only affected by the capacity value of the previous charge–discharge cycles but also by other external variables that affected the battery degradation. Therefore, the proposed AREV model was constructed by combining two influencing factors at the same time. In this experiment, we compared the AREV model with an autoregressive model that considered only the influence of autoregressive factors of the time series and with feature models considering only external variables. This was to compare the prediction accuracies of different models when predicting the capacity value of the next cycle. Therefore, three models, which were named the full model (AREV), autoregressive (AR) model, and model with only exogenous variables, were applied and compared. In addition, a machine learning approach, the support vector machine (SVM) method, was conducted using the same procedure as that described in Section 3.3 as a comparison.

In the AR model, the inputs of the model were the capacities at cycles k, $k-1$, $k-2$, and $k-3$, and the output was the capacity value at cycle $k+1$. In the model with only exogenous variables (feature model), the values of the eight features extracted from the IC curve and the IR curve (as described in Section 3.1) in the kth cycle were used as the inputs of the model, and the output was still the capacity at cycle $k+1$.

As shown in Fig. 9 and Table 2, the proposed AREV model had better prediction results than the model considering only autoregressive terms or the model with only external variables in terms of all three metrics. Therefore, we verified the idea of model construction in this paper; that is, the historical data of the capacitance and other features extracted based on electrochemical analysis and curve analysis played a role in the capacitance prediction of the next cycle. Furthermore, under this circumstance (using all autoregressive and external variable data as inputs), the prediction accuracy of the AREV model was still better than that of the SVM model.

4.3. Short-term look-ahead SOH prediction

In this experiment, we only considered the prediction accuracy of the capacity value at the end of the battery life (the last 100 cycles), because this result would be helpful in the subsequent remaining useful life prediction. Under the general definition, when the SOH value decayed to less than 80% of its initial value, the battery life can be regarded as at its end. Therefore, if the capacity value can be predicted accurately by a certain look-ahead period, the remaining useful life of the battery can also be predicted more accurately at this time.

Fig. 10 shows the performance of the n-cycle look-ahead prediction, and Table 3 records the prediction errors of the last 100 cycles. The training and testing process was the same, while we changed the output SOH value by setting an n-cycle look-ahead period. For each cycle number, the model parameters were obtained by training with the data before the current cycle. A 30-cycle moving window was also used in this experiment. Three look-ahead periods were obtained and compared, which were 10, 30, and 50 cycles ahead.

It can be seen in Fig. 10 that the proposed method had high accuracy for relatively small look-ahead periods n, but the performance decreased with the increase in n. This is not surprising, because the longer the time interval was, the lower the accuracy of forward-looking prediction became. Based on the results, the proposed model could accurately predict the capacity after 50 cycles.

5. Conclusions

In this paper, a data-driven method based on the AREV model was proposed to process battery capacity data. Our proposed model aimed to predict the SOH of a lithium-ion battery using data from both the SOH at the previous time point and other features that we extracted from the raw battery data.

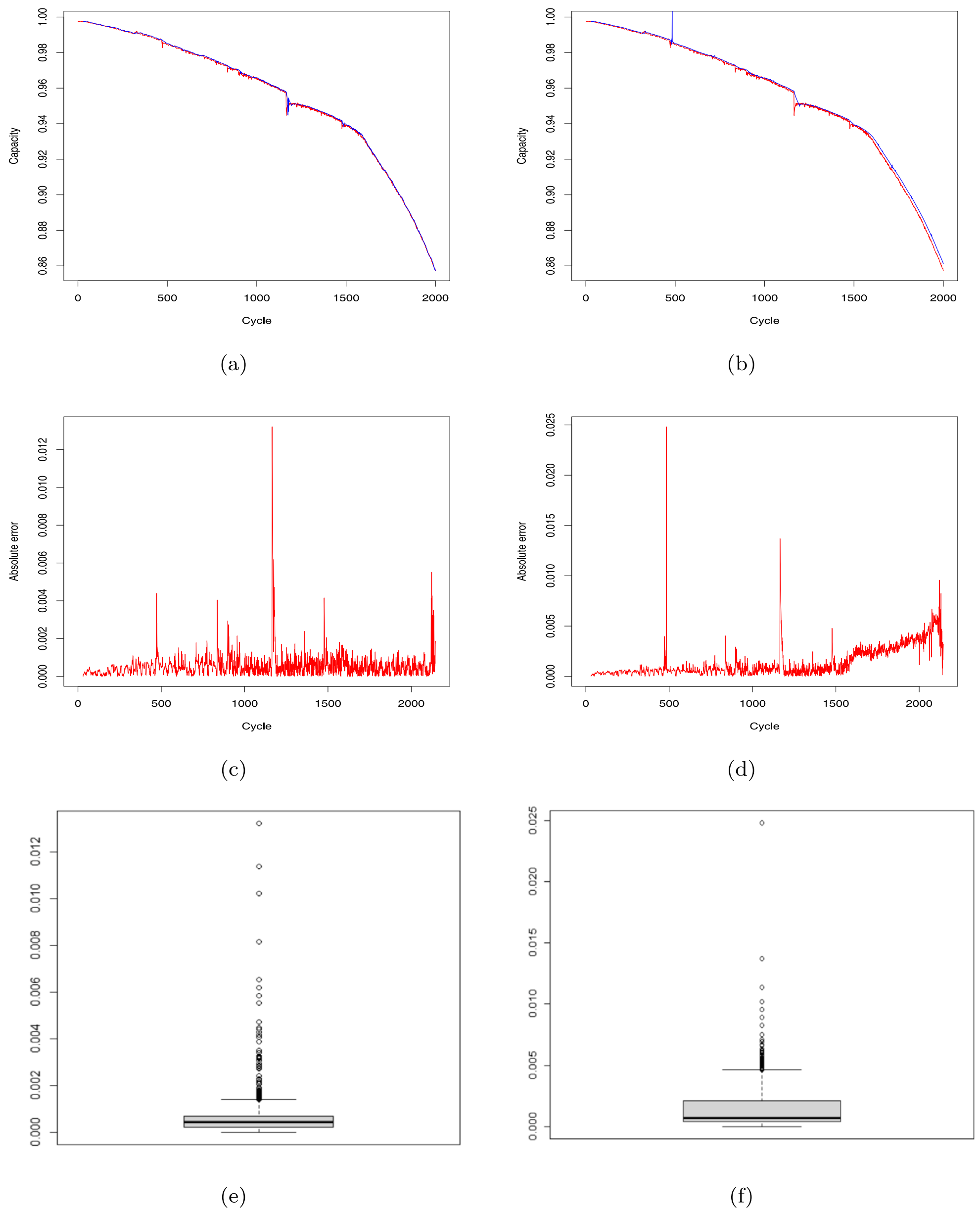


Fig. 9. Comparison of different models. Autoregressive model: (a) red line: real value, blue line: prediction results; (c) absolute error; (e) boxplot of absolute error. Model with only exogenous variables: (b) red line: real value, blue line: prediction results; (d) absolute error; (f) boxplot of absolute error.

Through our experiments, we drew the following conclusions. First, compared with models with only time series or feature information, the proposed AREV model, by combining time series information of the past three time points and other eight features that we extracted, achieved more accurate prediction results. Second, a moving window was proposed to select the training data from each battery and update the model at each time point so that we could obtain an online model. In addition, we used training data from the same data set and avoided the influence of different working conditions on different batteries. Third, in the feature extraction section, we extracted related features from battery data and used a penalized regression model to select the most related features to perform predictions. These features are based on the mechanism of battery decay in electrochemistry, as described in Section 2.1 of this article. Thus, they have strong interpretability and predictive effectiveness.

A methodology was proposed that optimally combined the information from multiple extracted features to produce more accurate and robust real-time predictions than any other existing system. Moreover, our ensemble approach was capable of using real-time and historical information to accurately predict the SOH several cycles ahead. Therefore, it could effectively monitor the decline of battery health in advance.

Furthermore, our experimental data were obtained by a constant-current discharge. At present, the battery data obtained in the laboratory cannot fully simulate the real dynamics of the charge and discharge cycle in the practical application of batteries, and the data are limited (from the experimental results of the MIT team). Although the charging and discharging of batteries are dynamic in practical applications, in general, the daily or weekly load is approximately the same. Therefore, in terms of applications, the following process can be used: regularly discharge the battery at a constant current, obtain relevant data, and extract and predict features to ensure the accuracy of the predictions.

Finally, in the previous mechanism analysis and curve comparison, we found that the curves of some data had a distinct downward trend or fluctuations with the aging of the battery, that is, the extracted features

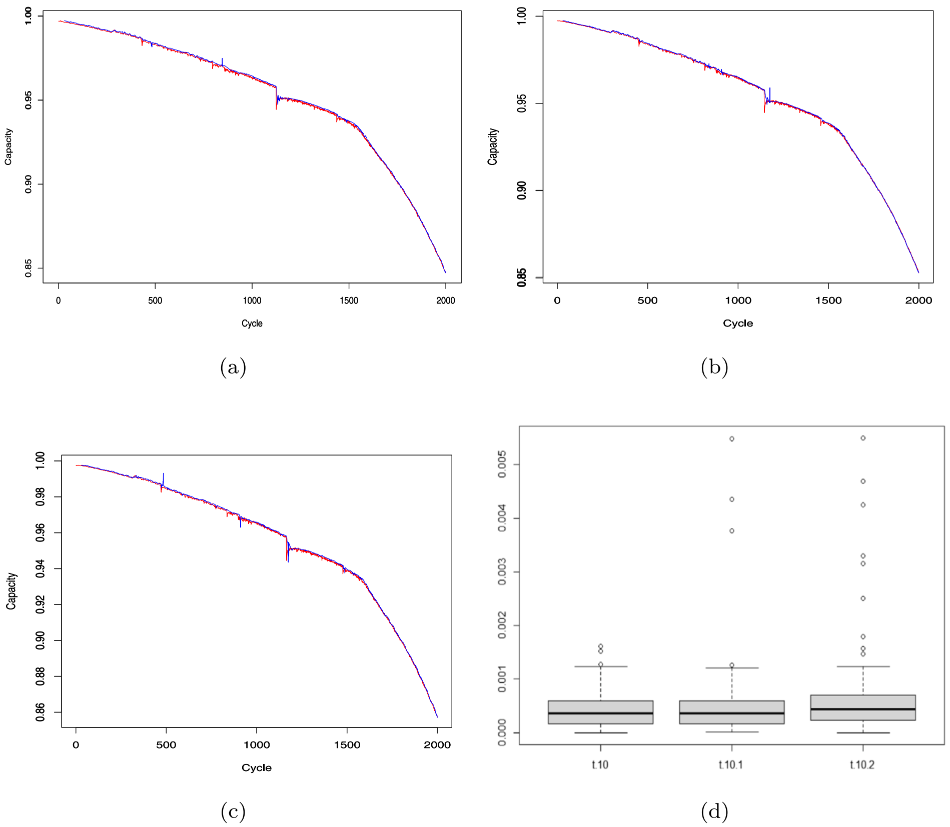


Fig. 10. Comparison of different look-ahead periods, where the red lines represent the real values and the blue lines represent the prediction results: (a) 50-cycle look-ahead prediction results; (b) 30-cycle look-ahead prediction results; (c) 10-cycle look-ahead prediction results; (d) box plot of absolute error for short-term look ahead prediction.

could clearly reflect the degradation process of the lithium battery. Therefore, the following points could be used as ideas to further explore the feature extraction of lithium batteries in the future:

1. The characteristics of directly measured variables, such as the aging cycle, charging time, and open circuit voltage, have been widely involved in previous research, so the variables with better effects could be selected as the basic characteristics.
2. According to the characteristics of the original voltage capacity–time curve, four characteristics were extracted from the constant-current charging curve, including the charging and voltage duration, slope, and vertical slope. The slope, intercept, and other parameters could be obtained by a simple machine learning model or a multiple regression model. By using the slope of the existing regression model and recording the change of its slope in each cycle, the degradation trends of batteries in different stages could be observed; accordingly, they were arranged to form a new sequence to reflect the change of the degradation trends over different degradation time periods.

   For example, by approximating the derivative, the IC curves can be converted to a difference form, and the difference capacity (DC) curves can be obtained, which can be computed by

   $$\triangle Q(V)_i = Q_{i+1}(V_{i+1}) - Q_i(V_i), \tag{8}$$

   where i is the index number of the data. Then, the DC curve can be linearized as follows:

   $$\triangle Q(V) = w(c)(a(c) - Q(V)) * Q(V) + B(c) + \epsilon(c), \tag{9}$$

   where $Q(V)$ represents the difference capacity; $a(c)$, $w(c)$, and $b(c)$ are the function model parameters of cycle number C; and $\epsilon(c) \sim N(0, \theta^2)$. The mean value is 0, and the variance is θ^2. Using the proposed difference model, the nonlinear relationship between the battery deterioration and potential information in the DC curve can be mined through the above transformed linear equation. Then, the model parameters $a(c)$, $w(c)$, and $b(c)$ of the whole curve can be obtained and used as health features in future work. Further details can be found elsewhere [32].
3. Voltage curve features can be processed, such as the IC and DV curves. In particular, because the peak intensity, peak offset, and other values in the IC curve have a great impact on the degradation of a battery, we can focus on extracting the characteristics related to it.
4. The characteristics of statistical indicators, such as the parameters of linear capacity fitting, the sample entropy of the voltage sequence, the internal resistance, and the polarization internal resistance, were analyzed in previous work. It was found that the internal resistance of the battery showed significant changes in different stages, so it can be used to reflect the degradation process. According to electrochemical theory, the sample entropy and polarization internal resistance also have some characteristics reflecting the degradation process, which have not been involved in previous work. Therefore, we can also focus on the feature extraction of the two curves.

CRediT authorship contribution statement

Zhelin Huang: Formal analysis, Methodology, Writing – original draft. **Fan Xu:** Data curation, Software. **Fangfang Yang:** Visualization, Writing – review & editing.

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

I have written the website to download data in the paper

Acknowledgments

This study was supported by National Natural Science Foundation of China (62203482), Shenzhen Institutions Stability Support Program #20200812150851001, National Natural Science Foundation of China (52205168), the Fundamental Research Funds for the Central Universities (WUT:223104001), the Guangdong Basic and Applied Basic Research Foundation (2021A1515110354), and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (22qntd1704).

References

[1] Shi Y, Escobar LA, Meeker WQ. Accelerated destructive degradation test planning. Technometrics 2009;51(1):1–13.

[2] Meeker WQ, Escobar LA. Statistical methods for reliability data. John Wiley & Sons; 2014.

[3] Whitmore G. Estimating degradation by a Wiener diffusion process subject to measurement error. Lifetime Data Anal 1995;1(3):307–19.

[4] Wang X. Wiener processes with random effects for degradation data. J Multivariate Anal 2010;101(2):340–51.

[5] Nelson WB. Accelerated testing: statistical models, test plans, and data analysis. Vol. 344. John Wiley & Sons; 2009.

[6] Park C, Padgett W. Accelerated degradation models for failure based on geometric Brownian motion and gamma processes. Lifetime Data Anal 2005;11(4):511–27.

[7] Wang X, Xu D. An inverse Gaussian process model for degradation data. Technometrics 2010;52(2):188–97.

[8] Xu P, Hu X, Liu B, Ouyang T, Chen N. Hierarchical estimation model of state-of-charge and state-of-health for power batteries considering current rate. IEEE Trans Ind Inf 2022;18(9):6150–9.

[9] Ouyang T, Xu P, Lu J, Hu X, Liu B, Chen N. Coestimation of state-of-charge and state-of-health for power batteries based on multithread dynamic optimization method. IEEE Trans Ind Electron 2022;69(2):1157–66.

[10] Shen L, Lu J, Geng D, Deng L. Peak traffic flow predictions: Exploiting toll data from large expressway networks. Sustainability 2020;13.

[11] Liu W, Xu Y. Data-driven online health estimation of li-ion batteries using A novel energy-based health indicator. IEEE Trans Energy Convers 2020;(99):1.

[12] Xing SA, Gl B, Js A, Zl C, Zheng C, Yl D. A uniform estimation framework for state of health of lithium-ion batteries considering feature extraction and parameters optimization. Energy 2020;204.

[13] Richardson RR, Osborne MA, Howey DA. Gaussian process regression for forecasting battery state of health. J Power Sources 2017;357:209–19.

[14] Richardson RR, Birkl CR, Osborne MA, Howey DA. Gaussian process regression for in-situ capacity estimation of lithium-ion batteries. IEEE Trans Ind Inf 2019;15(1):127–38.

[15] Liu K, Hu X, Wei Z, Li Y, Jiang Y. Modified Gaussian process regression models for cyclic capacity prediction of lithium-ion batteries. IEEE Trans Transp Electrif 2019;5(4):1225–36.

[16] 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.

[17] Wang Z, Zeng S, Guo J, Qin T. State of health estimation of lithium-ion batteries based on the constant voltage charging curve. Energy 2019;167(JAN.15):661–9.

[18] Yang F, Wang D, Zhao Y, Tsui KL, Bae SJ. A study of the relationship between coulombic efficiency and capacity degradation of commercial lithium-ion batteries. Energy 2018;145(FEB.15):486–95.

[19] Jiang B, Dai H, Wei X. Incremental capacity analysis based adaptive capacity estimation for lithium-ion battery considering charging condition. Appl Energy 2020;269.

[20] Tang X, Zou C, Yao K, Chen G, Liu B, He Z, et al. A fast estimation algorithm for lithium-ion battery state of health. J Power Sources 2018;396(aug.31):453–8.

[21] Pp A, Qz A, Lei WB, Zw A, Jh B, Hf C. Capacity estimation for lithium-ion battery using experimental feature interval approach - ScienceDirect. Energy 2020;203.

[22] Li X, Wang Z, Yan J. Prognostic health condition for lithium battery using the partial incremental capacity and Gaussian process regression. J Power Sources 2019;421(MAY 1):56–67.

[23] Li X, Yuan C, Wang Z. State of health estimation for li-ion battery via partial incremental capacity analysis based on support vector regression. Energy 2020;203:117852.

[24] Zhang D, Lu J, Pei C, Ni S. Electrochemical activation, sintering, and reconstruction in energy-storage technologies: Origin, development, and prospects. Adv Energy Mater 2022;12(19).

[25] Xu J, Liang P, Zhang D, Pei C, Zhang Z, Yang S, et al. A reverse-design-strategy for C@Li3VO4 nanoflakes toward superb high-rate Li-ion storage. J Mater Chem A 2021;9(32):17270–80.

[26] Xu Z, Zhang D, Lu J, Pei C, Li T, Xiao T, et al. Neural-network design of Li3VO4/NC fibers toward superior high-rate Li-ion storage. J Mater Chem A 2021;9.

[27] Kassem M, Bernard J, Revel R, Pelissier S, Duclaud F, Delacourt C. Calendar aging of a graphite/LiFePO4 cell. J Power Sources 2012;208:296–305.

[28] He J, Wei Z, Bian X, Yan F. State-of-health estimation of lithium-ion batteries using incremental capacity analysis based on voltage–capacity model. IEEE Trans Transp Electrif 2020;6(2):417–26.

[29] Han X, Ouyang M, Lu L, Li J, Zheng Y, Li Z. A comparative study of commercial lithium ion battery cycle life in electrical vehicle: Aging mechanism identification. J Power Sources 2014;251:38–54.

[30] Weng C, Sun J, Peng H. A unified open-circuit-voltage model of lithium-ion batteries for state-of-charge estimation and state-of-health monitoring. J Power Sources 2014;258:228–37.

[31] Yang F, Wang D, Zhao Y, Tsui K-L, Bae SJ. A study of the relationship between coulombic efficiency and capacity degradation of commercial lithium-ion batteries. Energy 2018;145:486–95.

[32] Kong JZ, Yang F, Zhang X, Pan E, Wang D. Voltage-temperature health feature extraction to improve prognostics and health management of lithium-ion batteries. Energy 2021;223(6):120114.