Health Prognosis With Optimized Feature Selection for Lithium-Ion Battery in Electric Vehicle Applications

Ji Wu, *Member, IEEE*, Xuchen Cui, Hui Zhang, and Mingqiang Lin

***Abstract*—The widespread use of lithium-ion batteries in electric vehicles has attracted widespread attention in both academia and industry. Among them, lithium-ion batteries' prognosis and health management are important research problems that need to be resolved urgently. This article proposes a novel computationally efficient data-driven state-of-health (SOH) estimation approach based on an optimized feature selection method. The difficulty of feature acquisition is defined based on voltage data distribution from more than 11 000 charging processes. The ridge regression is applied to model the battery aging process with features obtained from the charged capacity and incremental capacity data. Afterward, the feature set is downsized by solving a multiobjective optimization issue with the particle swarm optimization algorithm. The comparison experiments validate that our proposed optimized feature set performs better than the conventional feature set via manually selecting. Moreover, by collaborating with the selected features, the ridge regression can provide more reliable SOH estimation results using fewer computing resources than some nonlinear algorithms. Our approach is the first application of quantified feature acquisition difficulty in battery SOH estimation to the best of authors' knowledge.**

***Index Terms*—Difficulty of feature acquisition (DFA), electric vehicle (EV), particle swarm optimization (PSO), ridge regression, state of health (SOH).**

NOMENCLATURE

R^2	Coefficient of determination.
BMS	Battery management system.
BTR	Bagged tree regression.
C-curve	Capacity curve.
CDF	Cumulative probability function.
DFA	Difficulty of feature acquisition.
EV	Electric vehicle.
GEV	Generalized extreme value.
GPR	Gaussian process regression.
IC	Incremental capacity.
ICA	Incremental capacity analysis.
MAE	Mean absolute error.
PHM	Prognosis and health management.
PSO	Particle swarm optimization.
RMSE	Root-mean-square error.
SOC	State of charge.
SOH	State of health.
SSE	Sum of squared errors.
SVR	Support vector regression.

Manuscript received December 31, 2020; revised March 22, 2021; accepted April 22, 2021. Date of publication April 27, 2021; date of current version July 30, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61903114, in part by the Anhui Provincial Natural Science Foundation under Grant 2008085QF301, in part by the Youth Science, and Technology Talents Support Program (2020) by Anhui Association for Science and Technology under Grant RCTJ202008, and in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2019-002. Recommended for publication by Associate Editor S. Williamson. *(Corresponding author: Mingqiang Lin.)*

Ji Wu and Xuchen Cui are with the Department of Vehicle Engineering, Hefei University of Technology, Hefei 230009, China, and also with the Anhui Intelligent Vehicle Engineering Laboratory, Hefei 230009, China (e-mail: wu.ji@hfut.edu.cn; 2020170942@mail.hfut.edu.cn).

Hui Zhang is with the Department of Computer Science, KTH Royal Institute of Technology, 11428 Stockholm, Sweden (e-mail: huiz2@kth.se).

Mingqiang Lin is with the Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Sciences, Jinjiang 362200, China (e-mail: kdlmq@fjirsm.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPEL.2021.3075558.

Digital Object Identifier 10.1109/TPEL.2021.3075558

I. INTRODUCTION

PROGNOSIS and health management (PHM) of the lithium-ion battery in the electric vehicle (EV) application is a vital mission of the battery management system. The healthy state can directly affect the battery system's stability, reliability, safety, or even the entire EV. Therefore, accurately estimating the state of health (SOH) is eagerly anticipated by researchers and engineers [1], [2].

Instead of building a battery aging model based on the prior knowledge of the battery's chemical mechanism, the data-driven method can establish a model with sufficient experimental cycle data and then predict the battery change during the entire cycle with few selected features [3]. Besides, open-source datasets, such as [4] and [5], are also the driving force for developing data-driven PHM.

Due to nonlinear characteristics, machine learning algorithms frequently appear in the literature for their excellent ability to fit complex systems and complex relationships. A feed-forward neural network proposed by Wu *et al.* [6] model a relationship between the battery terminal voltage and the battery's remaining useful life with the nonlinear hidden neuro layers. The neural network [7] was also employed to estimate battery SOH together with the Markov chain to restrain the random estimation errors from the network mentioned above. A support vector regression (SVR) was utilized in [8] to model the capacity degradation

and describe the lithium-ion battery's aging process. The model was evaluated by the root-mean-square error (RMSE) and was found to have about 20 mAh RMSE. Yang *et al.* [9] proposed a battery aging model by a modified Gaussian process regression (GPR) with the similarity measurement of input variables and covariance kernel function. The RMSE of this model was about 1%. However, to apply the battery aging model in real EV applications, the abovementioned complicated algorithms should be simplified. An identical and straightforward dynamically driven recurrent network-based SOH estimation method was developed in [10] to overcome the battery's nonlinear dynamic nature. Tang *et al.* [11] developed a linear model using the charging battery's regional capacity and obtained an estimated SOH with less than 2.5% error. It is evident that both nonlinear and linear models can achieve acceptable SOH estimation results while using the ideal health indicators.

Vast quantities of indicators have been selected to depict the battery degradation process and set as the aging models' parameters. Features were selected from battery terminal voltage curves in the constant current charging process via differential geometric analysis in [12]. Similarly, the relation between the health indicators from the constant voltage charging curves and the SOH was developed in [13] by considering the incomplete discharging process's challenge. However, for most past research, the situation of the real vehicular applications is not contemplated. Dong *et al.* [14] combined the incremental capacity analysis (ICA) of constant-current charging and the time constant of constant-voltage to describe the battery's degradation. Four metric points of the ICA curve were gathered in the voltage interval of [3.3 V, 4.1V] by Daniel *et al.* [15]. These points were then used to relate with battery capacity fade for SOH estimation. Sample entropy of the voltage sequences under the hybrid pulse power characterization profiles in the hybrid pulse tests, which can hardly be realized in real operations, was used as the signature of battery degradation [16].

However, influenced by the randomness of EV drivers' behavior, features selected from some particular state of charge (SOC) or specific voltages may be impractical for the SOH estimation in real EV applications. The frequencies of these SOC and voltages' occurrence are different regardless of the charging or discharging process. This begs an undersolved problem: Which features are more comfortable to obtain in real EV applications? Furthermore, as stated in [17], "bad features may require a much more complicated model to achieve the same level of the performance." Features would strictly influence the complexity of the estimation model and the accuracy of the estimation results. This begs another issue: How to find the optimal features with the battery aging model's synergy in SOH estimation?

To address these issues, we proposed a novel optimized feature selection method in this article. First, the initial voltage distribution at the constant charge process is analyzed based on actual EV data from more than 11 000 charging operations. Afterward, the difficulty of feature acquisition (DFA) is defined by the voltage distribution's cumulative probability function (CDF). Second, computation-friendly ridge regression is employed to model the relation between the selected features and the SOH for its simplicity and effectiveness. Third, features are selected by solving a multiobjective optimization problem using the particle swarm optimization (PSO) algorithm. Therefore, the DFA and the model's estimation error listed in the objective function can be minimized. Finally, a reliable SOH estimation result could be achieved by combining the optimal features and the linear model. The main contributions of this article can be summarized as follows.

1) Definition of feature acquisition difficulty for actual EV application is proposed using voltage distribution from more than 11 000 charging processes.
2) Solution of feature selection by solving a multiobjective optimization problem is presented and implemented.
3) Accurate SOH estimation is achieved by using the optimized features and linear ridge regression synergistically.

The rest of this article is organized as follows. In Section II, the operational data from real EV applications are analyzed. An optimized feature selection method, which consists of data extraction, feature generation, and feature wrapper, is developed in Section III. The ridge regression and solution of multiobjective optimization are introduced in Section IV. Experiments with three Li-ion batteries are conducted in Section V. Finally, the conclusion is given in Section VI.

II. Data Analysis for Real EV Applications

The lithium-ion battery's constant current charging process in the laboratory environment is moderately similar to the real EV applications in the charging current and mode. Moreover, benefit from the advanced thermal management method in the EVs, battery temperature can also be controlled as the experimental environment [18]. Hence, the cyclic data from the experimental platform can be applied to model the battery degradation process and estimate the SOH for real EV applications. However, there are still several differences between experimental conditions and practical applications.

Batteries in different EVs may have various driving cycles due to different user habits. The difference in driving conditions will lead to differences in the battery's initial SOC and voltage when it starts to charge. At the beginning of charging, the state of the battery will, therefore, show a certain degree of randomness [19]. As a result, voltage and current data within a long timescale can hardly be procured in every charging process. Moreover, the constant voltage charge is usually replaced by a stair-stepping constant current charge to make the charging process more manageable. Usually, the battery would be charged to the cut-off voltage with a continual 0.5 C and then with 0.25 C or 0.3 C (1 C means the battery can be fully discharged with a constant current in 1 h). Finally, the current can be reduced to 0.05 C or lower. Raw current and voltage data during a typical charging process are plotted in Fig. 1(a) and (b). However, most of the actual charging processes may not behave so ideally. Users often stop charging before the SOC becomes 100%. Furthermore, the ending time can vary with each individual. Therefore, data from more than 11 000 charging operations of about 100 electric vans are analyzed in this article. The statistical results are demonstrated in

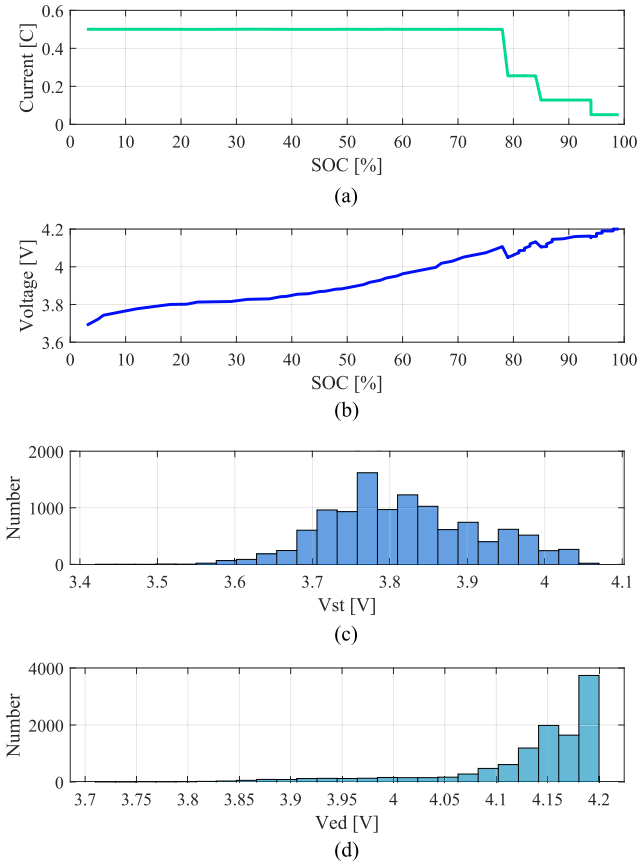


Fig. 1. Typical charging process: (a) charging current, (b) battery terminal voltage. Distribution of the charging process: (c) voltage at the beginning, (d) voltage at the end.

Fig. 1(c) and (d). V_{st} and V_{ed} represent battery voltages when start and end the charge, respectively. It is shown that the charge starting voltages distribute dispersedly at the range of 3.4–4.1 V. On the contrary, due to the users' range anxiety, the charging processes are more intensively ended at high SOC ranges and high voltages [20]. As can be seen from Fig. 1(d), about 81% of the V_{ed} are higher than 4.1 V. Therefore, the distribution of V_{ed} will not be considered in this article.

As shown in Fig. 1, some voltage ranges occur more frequently than others during the EVs' charging process. In other words, some data are more comfortable to be collected in the real application. Therefore, the difficulty of information acquisition should be considered to make the health prognosis method more practical when selecting health indicators or features.

III. Optimized Feature Selection

Features are detected from the raw charging data to build the aging model and estimate the battery's healthy state. A feature selection method consists of capacity curve (C-curve) extraction, feature generation, and feature wrapper is developed to find the most appropriate features for SOH estimation in this article. The

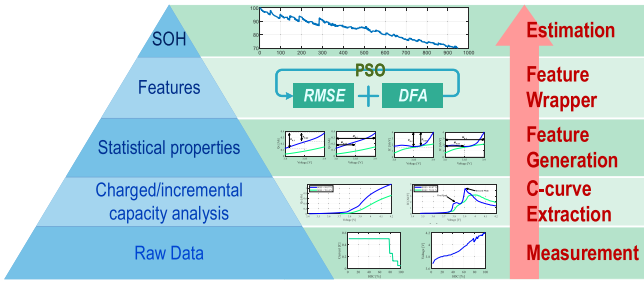


Fig. 2. Procedure of the developed feature selection method.

SOH is defined as follows

$$\text{SOH} = \frac{C_t}{C_0} \times 100. \tag{1}$$

where C_t is the maximum available capacity currently and C_0 represents the capacity when the battery starts to work.

The procedure of this feature selection method is illustrated in Fig. 2. The charged capacity (Qc) and incremental capacity (IC) are analyzed based on the raw data during C-curve extraction. Potential features are then generated from these capacity-related data. Finally, optimal features for SOH estimation are selected by the PSO coordinating with the linear model.

Experimental data from the lithium-ion battery's cyclic constant charge processes are utilized to provide a visualized description of the feature selection procedure.

A. C-Curve Extraction and Feature Generation

According to the raw data's high density, it is impractical to establish the aging model using these data directly. Therefore, the C-curves, e.g., charged capacity and IC, are extracted from the collected current, voltage, and time data by the EV sensors. The IC is defined as the capacity increment for a specific voltage step during the charging process and is expressed as follows:

$$\text{IC} = \frac{\mathrm{d}Q}{\mathrm{d}V} = \frac{I \cdot \mathrm{d}t}{\mathrm{d}V} \tag{2}$$

where I is the charging current, $\mathrm{d}t$ is the sampling time interval, $\mathrm{d}V$ is the voltage step and is set to 0.004 V.

Characteristics of the charged capacity and IC curves are utilized to generate potential features for the battery aging model and health prognosis. Since voltage is a measurable variable, the charged capacity and IC are both displayed as the functions of the voltage under different SOHs in Fig. 3. In other words, the feature obtained from the capacity and IC data is also considered a function of voltage.

It is found that the charged capacity is obviously decreased throughout the aging process. Hence, as shown in Fig. 3 (b) and (c), ranges and means of the capacity variations within specific voltage intervals are employed as SOH estimation features. $R_{0.1}$ represents the difference between the maximum and minimum values of the charged capacity variation within 0.1 V. $M_{0.1}$ is the mean value of the capacity variation in 0.1 V voltage

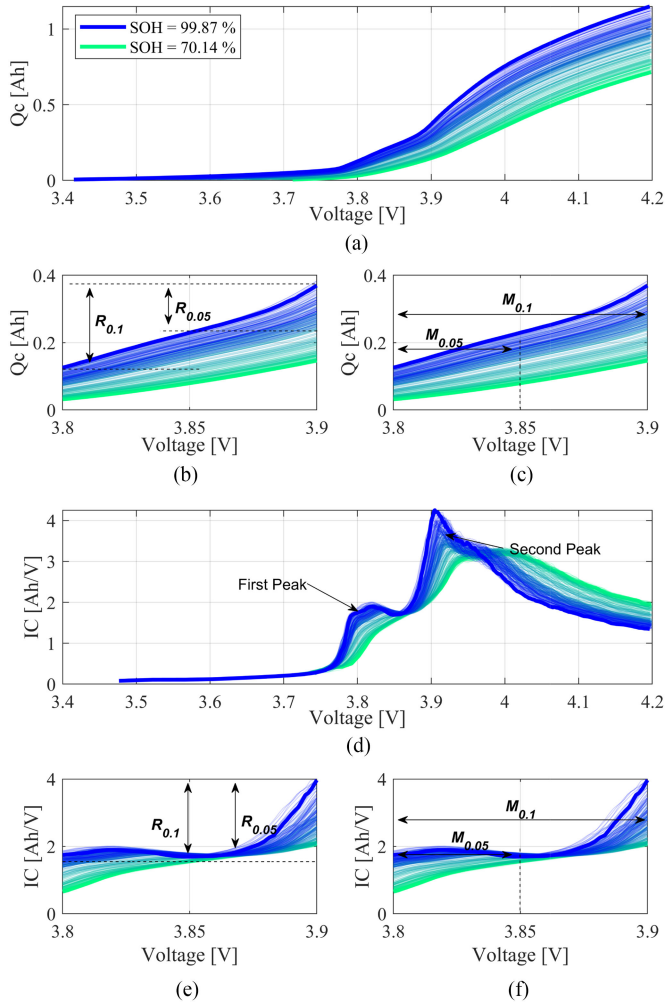


Fig. 3. (a) Charged capacity at different SOH; (b) range and (c) mean of the charged capacity in specific voltage intervals; (d) IC at different SOH; (e) range and (f) mean of the IC in specific voltage intervals.

interval. The generated range-related and mean-related features are computed in voltage intervals of 0.025, 0.05, 0.075, and 0.1 V in this article, respectively.

Similarly, the ranges and means are also applied in the IC curves to create features. Furthermore, as shown in Fig. 3(d), there are two peaks in most IC curves. The peaks' intensity and position can reveal the capacity degradation resulting from loss of Li+ or active material [21]. The IC curves' peak reflects a phase change in the intercalation material where two or more phases with different lithium concentrations coexist with the same chemical potential [22]. From the perspective of external battery characteristics, the IC peaks can detect the battery's ability to absorb energy under a particular potential and would change with the variation of the battery's internal material during the aging process. Therefore, characteristics of these peaks are also used for SOH estimation. Specifically, the maximum values, the range values, the arithmetic mean values, and the geometrical mean values in voltage intervals of [3.75 V, 3.85V] and [3.85 V, 4.1V] are calculated. Afterward, a feature set, denoted by $\boldsymbol{F}$, consists of more than 250 features are obtained from the measured data with the extraction and generation operations.

B. Feature Wrapper

Features obtained from the extraction and generation process are still exceeded for the model establishment and SOH estimation. To reduce the features' quantity while finding the optimal feature for the aging model, feature wrapper is treated as a multiobjective optimization issue in this article.

First, selected features should cooperate with the data-driven machine learning model to achieve a reliable SOH estimation performance. Hence, the RMSE of the SOH estimation results is set as one of the objectives. The expression of the RMSE is given as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathrm{SOH}_j - \hat{\mathrm{SOH}}_j)^2} \tag{3}$$

where RMSE is the RMSE of estimation result; SOH_j and $\hat{\mathrm{SOH}}_j$ are the real SOH and the estimated SOH, respectively; m is the element number for validation.

Second, as mentioned in Section II, some data are more comfortable to be obtained. It is observed that features from different voltage intervals also have different difficulty degrees for acquisition. Therefore, the DFA is defined according to the battery voltage distribution at the beginning of the charging process. The cumulative probability of the voltage corresponding to the potential feature is employed to quantify the DFA. Features generated from voltage ranges that occur more frequently would have less DFA. For example, as shown in Fig. 1(c), features from [3.8 V, 3.9V] may have a smaller DFA than the feature obtained from [3.4 V, 3.5V]. The DFA is expressed as follows:

$$\mathrm{DFA}(x) = 1 - \mathrm{CDF}_{\boldsymbol{V}}(x) = 1 - \int_0^x f_{\boldsymbol{V}}(v)\mathrm{d}v \tag{4}$$

being

$$f_{\boldsymbol{V}}(v) = \frac{1}{\sigma}\exp\left(-\left(1+\mathrm{K}\frac{(v-\mu)}{\sigma}\right)^{-\frac{1}{k}}\right)\left(1+\mathrm{K}\frac{(v-\mu)}{\sigma}\right)^{-1-\frac{1}{k}}$$

where $\mathrm{DFA}(x)$ is the DFA at voltage x; $\mathrm{CDF}_{\boldsymbol{V}}$ is the CDF, which is fitted by the GEV distribution in this article; $\boldsymbol{V}$ is the range of voltage at the beginning of the charge and is equal to [3.4 V, 4.1V]; $f_{\boldsymbol{V}}(v)$ is the probability density function of the V_{st}; K, μ, and σ are the unknown parameters of the GEV distribution.

Therefore, to minimize the RMSE and DFA, the feature wrapper's objective function can be expressed as follows:

$$J = w_1\mathrm{RMSE}\left(\boldsymbol{F}_s\right) + w_2\mathrm{DFA}\left(V_{\min}\right) \tag{5}$$

where w_1 and w_2 are the weights of RMSE and DFA, respectively; $\boldsymbol{F}_s$ is the set of selected features. DFA is calculated by the minimum voltage corresponding to these features, denoted by $V_{\min}$.

Two constraints should be satisfied during the optimization.

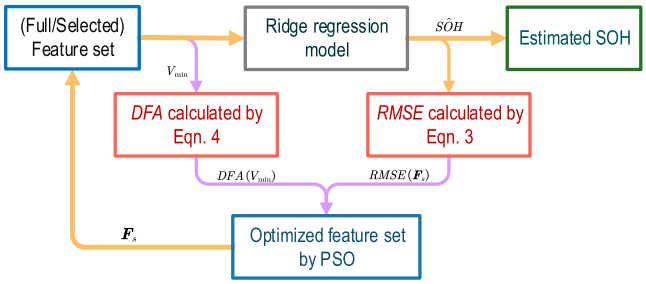


Fig. 4. Procedure of SOH estimation.

1) features should be selected from the potential feature set $\boldsymbol{F}$;
2) the minimum voltage of the selected features should locate in the duration of the charge starting voltage $\boldsymbol{V}$.

Finally, the minimization of the RMSE from the ridge regression and the feature set's DFA is addressed as a multiobjective optimization issue. By combining the objective and constraints proposed above, this optimization problem's mathematical expression can be concluded as follows:

$$\begin{aligned} \min \quad & J = w_1 \mathrm{RMSE}\left(\boldsymbol{F}_s\right) + w_2 \mathrm{DFA}\left(V_{\min}\right) \\ \text{s.t.} \quad & \boldsymbol{F}_s \subseteq \boldsymbol{F} \\ & V_{\min} \in \boldsymbol{V}. \end{aligned} \tag{6}$$

IV. SOH Estimation

The procedure of the presented SOH estimation method is shown in Fig. 4. Battery SOH is estimated by a ridge regression model with selected features. The PSO algorithm and ridge regression are cooperated to find the most suitable features according to the DFA and RMSE. The optimized feature set will be iteratively propagated back to the linear model until a satisfactory result is procured.

A. Ridge Regression

First, the SOH is detected by the ridge regression for its simpleness and high efficiency. Therefore, to compute the RMSE, the estimated SOH, $\hat{\mathrm{SOH}}$ in (3), from the linear model can be expressed as follows:

$$\hat{\mathrm{SOH}} = \boldsymbol{F}_s^{\mathrm{T}} \boldsymbol{\beta} \tag{7}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots \beta_n)^{\mathrm{T}}$ is an n-dimensional row vector of parameters to be identified; n is also the number of the selected features.

In the ridge regression, to prevent model overfitting, a penalty is added to the least-squared algorithm when identifying the $\boldsymbol{\beta}$. Therefore, the sum of squared errors (SSE) with a regularization item is minimized and given as follows:

$$\mathrm{SSE}_{\mathrm{ridge}=} \|\mathrm{SOH} - \hat{\mathrm{SOH}}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_2^2 \tag{8}$$

where α is a nonnegative complexity parameter and is equal to 0.1, in this article, $\|\cdot\|_2^2$ is the $L2$ norm.

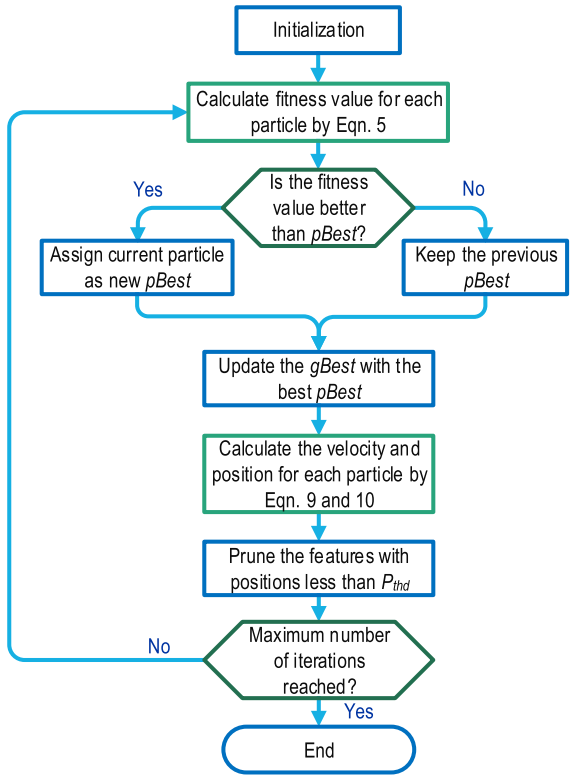


Fig. 5. Procedure of PSO-based feature wrapper.

B. Particle Swarm Optimization

The PSO addresses the multiobjective optimization problem with a weighted sum of the RMSE and DFA. PSO is an evolutionary algorithm that optimizes a problem by improving a swarm of particles' best position iteratively [23]. Particles in the swarm are moving around in the search space according to the formulas of their position and velocity. For every single particle, its position is impacted by the best-known positions, locally and globally. After repeating the searching process, the swarm eventually moves toward satisfactory solutions and finally solves the optimization problem. In this article, the PSO is employed to find the best features. A schematic diagram describing the optimization procedure is shown in Fig. 5.

Random values initialize the particles' positions and velocities with ranges of [0, 1]. Each particle is formed by the potential features, which is an N-dimensional vector. N is the number of potential features. Hence, position and velocity are also N-dimensional. The position of the particle is utilized to choose the suitable features with a threshold, denoted by P_{thd}. Moreover, the velocity can help the particle moving toward the best position. The particle with a better fitness value, which is also a better feature set, can be discovered by the defined objective function (5) in this article. The velocity and position of each particle are calculated by

$$\begin{aligned} v_i\left(k+1\right) = {} & w(k) v_i(k) + c_1 r_1 \left(p_i - x_i(k)\right) \\ & + c_2 r_2 \left(p_g - x_i(k)\right) \end{aligned} \tag{9}$$

$$x_i\left(k+1\right) = x_i(k) + v_i\left(k+1\right) \tag{10}$$

TABLE I
SPECIFICATIONS OF THE EXPERIMENTAL BATTERIES

Capacity [Ah]	50
Nominal voltage [V]	3.65
Voltage range [V]	2.5 to 4.2
Charging temperature range [°C]	0 to 55
Discharging temperature range [°C]	-20 to 45

TABLE II
MAJOR EQUIPMENT PARAMETERS

Device	Characteristic
CETC battery testing system	Voltage: 0V to 5V (± 0.1% FS) Charging current: 1A to 100A (± 0.1% FS) Discharging current: -1A to -100A (± 0.1% FS)
Thermal chamber	Temperature: -40°C to 150°C (± 1.5°C)

being

$$w(k) = w_{\max} - \frac{(w_{\max} - w_{\min})\, k^2}{k_{\max}^2}$$

where $v_i(k)$ and $x_i(k)$ are, respectively, the velocity and position of ith particle at kth iteration; c_1 and c_2 are the acceleration coefficients, and are both equal to 2; r_1 and r_2 are the random numbers in the range of [0, 1]; p_i is the best-known position of the ith particle, named as pBest in Fig. 5; p_g is the best position of the particle in the entire swarm, named as gBest; $w(k)$ is a linear differential decreasing inertia weight with a maximum value of $w_{\max}$ and a minimum value of $w_{\min}$; $k_{\max}$ is the maximum number of iterations.

V. EXPERIMENTS AND ANALYSIS

Experiments are conducted to verify the proposed feature selection method and also the SOH estimation approach. The operational data from real EVs are gathered by the new energy vehicle remote monitoring center at the Hefei University of Technology. Three lithium-ion batteries, which are of the same type produced by the same manufacturer, denoted by batteries A, B, and C, are circularly charged and discharged with a continuous 0.5 C until the available capacity decreased to 70% of its nominal capacity. The specifications of the experimental Li-ion batteries are given in Table I. All the experiments are conducted and recorded by the CBTS-50A-05 V type battery testing system manufactured by the China electronics technology group corporation (CETC), as shown in Fig. 6. The environmental temperature of the tested batteries is controlled by the HD-E702-100 type thermal chamber from the Haida International Equipment Co., Ltd. Experimental data are collected by the devices listed in Table II and then processed by the MATLAB 2018b on a computer with an AMD 3950X processor and 64 GB memory.

A. Distribution of the Starting Voltage

Three different distributions are utilized to fit the data of the V_{st} from more than 11 000 charging operations. The cumulative probability and probability density are plotted in Fig. 7. "GEV" represents the generalized extreme value distribution, a continuous probability distribution. The "Normal" represents the normal distribution widely used to fit the initial SOC distribution when charging [24]. The "Weibull" is the Weibull distribution, an effective tool for reliability characteristics and trends determination [25]. The maximum likelihood estimation is employed to identify the parameters of the probability density function. In Fig. 7, the GEV and normal distributions are demonstrated to have similar fitting accuracy, both better than the Weibull distribution.

Fig. 6. Battery testing platform.

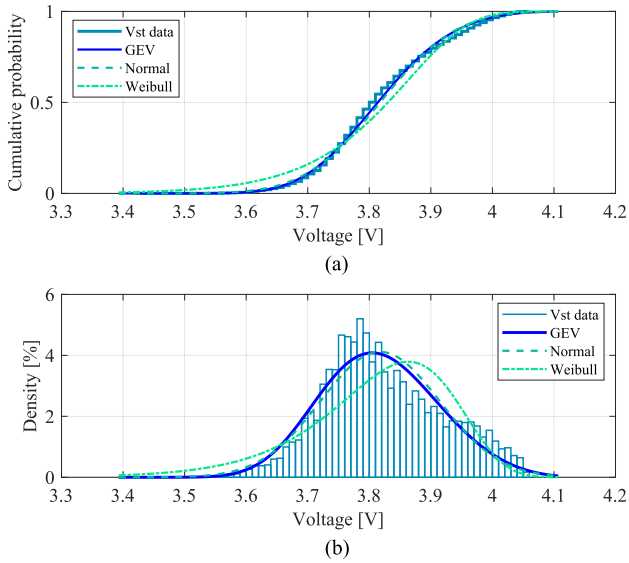


Fig. 7. Distribution of the V_{st}: (a) cumulative probability; (b) probability density.

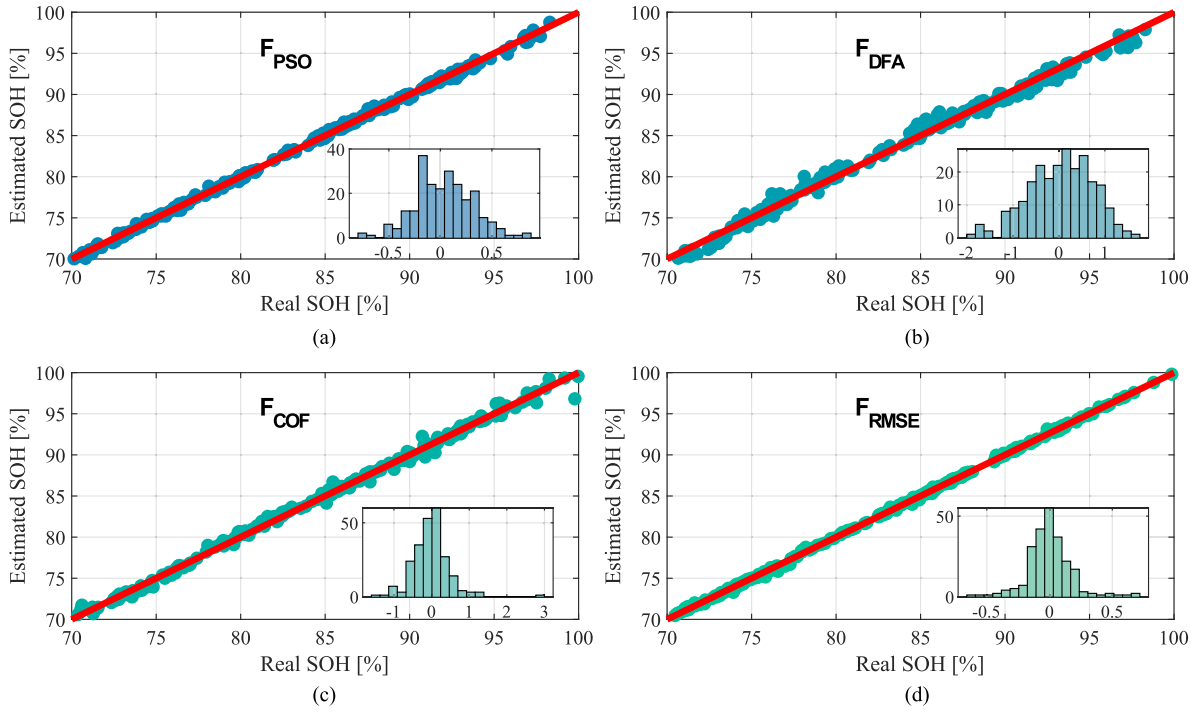


Fig. 8. Estimation results of different feature sets for battery A: (a) $\boldsymbol{F}_{\mathrm{PSO}}$; (b) $\boldsymbol{F}_{\mathrm{DFA}}$; (c) $\boldsymbol{F}_{\mathrm{COF}}$; (d) $\boldsymbol{F}_{\mathrm{RMSE}}$.

TABLE III
COMPARISON OF DIFFERENT DISTRIBUTION FUNCTIONS

Distribution	GEV	Normal	Weibull
Log-likelihood	10527.20	10395.30	9454.07
RMSE for CDF	0.0217	0.0279	0.0541

Afterward, the fitting results of these three distributions are evaluated by the log-likelihood and the RMSE. The numerical result is given in Table III. The GEV's log-likelihood is about 1.3% and 11.3% greater than the normal distribution and Weibull distribution, respectively. The fitting error of the GEV is the least among these three approaches. It is about 22% less than the normal distribution and is near half of the Weibull distribution. Hence, the GEV is applied to simulate the CDF and quantify the DFA.

B. Comparison of Different Features

A comparison of SOH estimation with different feature sets is established to validate the proposed PSO-based feature selection method. To be fair, all the SOHs will be estimated by the ridge regression models. The number of the features is also set equal for all of the comparative feature sets. A total of 75% of battery A's cyclic data are randomly selected to train the linear model. Then, the rest, 25%, are used to verify the estimation accuracy. Seven optimal features, such as the mean values of IC in [3.93 V, 4.005V] and [3.95 V, 4.15 V], the range of the charged capacity in [4.05 V, 4.075V], the mean values of the charged capacity in [3.97 V, 4.03V], [4.01 V, 4.06V], [3.93 V, 4.005V], and [3.97 V, 4.07V], denoted by $\boldsymbol{F}_{\mathrm{PSO}}$, are finally selected by optimizing the feature set with the PSO algorithm after 300 iterations. Features are selected from different kinds of extracted data. Among the optimal features, two of them are obtained from the IC data, and the rest are generated from the charged capacity data. The optimal result is inseparable from the cooperation of the IC and Qc-based features.

Furthermore, features with minimum DFA, denoted by $\boldsymbol{F}_{\mathrm{DFA}}$, with the highest correlation coefficients with the SOH [26], denoted by $\boldsymbol{F}_{\mathrm{COF}}$, and with minimum SOH estimation RMSE, denoted by $\boldsymbol{F}_{\mathrm{RMSE}}$, are selected for comparison. As displayed in Fig. 8, results from the features with minimum DFA have the worst performance. Most of the errors are distributed in the range of -1% to 1%. $\boldsymbol{F}_{\mathrm{PSO}}$ and $\boldsymbol{F}_{\mathrm{RMSE}}$ may have relative SOH estimation accuracy, with absolute errors around 0.5%.

More specifically, the RMSE, mean absolute error (MAE), R^2, and DFA are utilized to evaluate different feature sets' performance. MAE and R^2 are defined as follows:

$$\mathrm{MAE} = \frac{1}{m}\sum_{j=0}^{m-1}\left|\mathrm{SOH}_j - \hat{\mathrm{SOH}}_j\right| \tag{11}$$

$$R^2 = 1 - \frac{\sum_{j=1}^{m}(\mathrm{SOH}_j - \hat{\mathrm{SOH}}_j)^2}{\sum_{i=1}^{n}(\mathrm{SOH}_j - \overline{\mathrm{SOH}})^2} \tag{12}$$

being

$$\overline{\mathrm{SOH}} = \frac{1}{m}\sum_{j=1}^{m}\mathrm{SOH}_j.$$

Numerical results are given in Table IV. The $\boldsymbol{F}_{\mathrm{RMSE}}$ wined the best RMSE, MAE, and R^2 among these four feature sets. However, these features are also the most difficult to acquire.

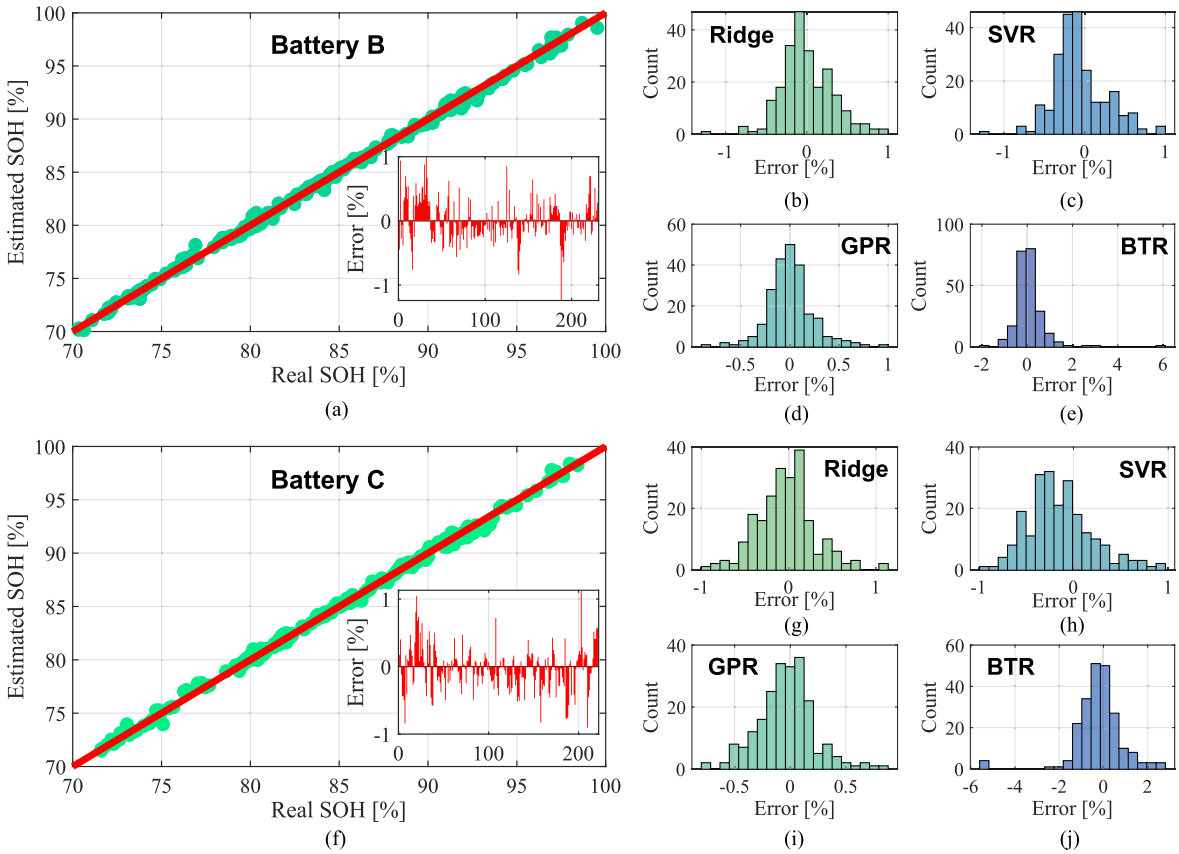


Fig. 9. Estimation results of different models: (a)–(e) battery B; (f)–(j) battery C. (a) and (f) Estimation results of ridge regression. (b)–(e) and (g)–(j) Error distributions of different models.

TABLE IV
NUMERICAL RESULTS OF DIFFERENT FEATURE SETS FOR BATTERY A

Feature set	$\boldsymbol{F}_{\mathrm{PSO}}$	$\boldsymbol{F}_{\mathrm{DFA}}$	$\boldsymbol{F}_{\mathrm{COF}}$	$\boldsymbol{F}_{\mathrm{RMSE}}$
RMSE [%]	0.2855	0.7182	0.4414	0.1744
MAE [%]	0.2214	0.5902	0.3174	0.1250
R^2	0.9986	0.9914	0.9968	0.9995
DFA	0.1261	0.0083	0.5968	0.7491

The features in set $\boldsymbol{F}_{\mathrm{DFA}}$ can be generated through about 99% charging operations, but its SOH estimation accuracy is unacceptable. The DFA of $\boldsymbol{F}_{\mathrm{COF}}$ is also too high to be employed in real applications. With the help of the optimization process, features with balanced DFA, and estimation accuracy are selected. According to the DFA in Table IV, the $\boldsymbol{F}_{\mathrm{PSO}}$ can be obtained from most of the charging processes. The MAE and RMSE of the optimized features are also acceptable for actual EV applications. Therefore, it is indicated that features selected by the PSO can accomplish reliable SOH estimation while being obtained in most practical operations.

C. Comparison of Different Models

Several commonly used machine learning algorithms, e.g., SVR, GPR, and bagged tree regression (BTR) [27], are implemented to estimate the SOH comparatively. The SVR's kernel is a linear function, and the GPR's kernel is a rational quadratic function. Similarly, 75% of the cyclic data from batteries B and C are used to train the models, and the rest are for verification. Features selected by PSO based on the linear model are inputting to the regressors mentioned above to build the aging model. It is noted that model parameters are identified by MATLAB's statistics and machine learning toolbox adaptively.

The SOH estimation results of different models for two lithium-ion batteries are shown in Fig. 9. Models based on the SVR, GPR, and ridge regression are found to have a better performance than the BRT. The ridge regression can achieve a similar SOH estimation result to the SVR. Most of the estimation errors are locating at the range of -1% to 1% by using the two models abovementioned. The GPR model is the most accurate among all four methods. However, as can be seen from the numerical results given in Table V, the GPR also takes much more operational time for model building and SOH estimation than the other algorithms, more than 5000 ms. Benefit from the optimized feature selection, the ridge regression is observed to have less RMSE and MAE than the SVR and BTR while with less computation requirement. For battery B, the ridge regression with optimized features can obtain a reliable SOH estimation result with just 2 ms, which is much shorter than the GPR and the SVR. Its MAE is about 6.8% less than the SVR and about 38.3% lower than the BTR. The same consequence can also be detected from the results of battery C. The selected features

TABLE V
SOH ESTIMATION RESULTS FROM DIFFERENT MODELS FOR BATTERIES B AND C

	Battery B				Battery C			
	Ridge	**SVR**	**GPR**	**BTR**	**Ridge**	**SVR**	**GPR**	**BTR**
RMSE [%]	0.3050	0.3262	0.2425	0.5649	0.3415	0.3766	0.2781	0.8677
MAE [%]	0.2323	0.2492	0.1706	0.3768	0.2548	0.3020	0.2021	0.6240
R^2	0.9980	0.9977	0.9988	0.9933	0.9975	0.9969	0.9983	0.9839
Time [ms]	2	125	8501	467	2	217	5047	471
DFA	0.0257				0.0409			

help improve the ridge regression's fitting performance to better estimate results than some nonlinear models.

Furthermore, with the help of the PSO, the DFA is synchronously minimized with the estimation errors. Both the DFA of batteries B and C are less than 0.05, indicating that the estimation results mentioned above can be procured with charging data from more than 95% of the actual operations.

VI. CONCLUSION

An optimized feature selection approach has been proposed in this article. Optimal features for linear model-based SOH estimation are obtained by solving a multiobjective function with the PSO algorithm. The features' acquisition difficulty and the ridge regression's estimation error are considered in the optimization process. The experimental results show that the developed feature selection method can optimize the DFA and the RMSE and obtain a better feature set than traditional manual selection. It is also validated that, comparing with the SVR and GPR, the ridge regression using selected features can provide similar performance with a much simpler structure and less computation cost. Furthermore, benefit from the DFA consideration, the proposed method can realize an accurate and high-effective SOH estimation for most practical charging operations (cover about 90% of the charging processes).

One limitation of the proposed method is that the battery pack's inconsistency is not considered. "More is different." In future work, the characteristics of the battery pack will be studied for the implementation of the presented SOH estimation approach.

ACKNOWLEDGMENT

Data Availability: Some of the datasets used in this article are available at https://github.com/CEL-HFUT/Voltage-distribution-for-health-prognosis.

REFERENCES

[1] X. Hu, L. Xu, X. Lin, and M. Pecht, "Battery lifetime prognostics," *Joule*, vol. 4, no. 2, pp. 310–346, 2020.
[2] Y. Li *et al.*, "Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review," *Renewable Sustain. Energy Rev.*, vol. 113, 2019, Art. no. 109254.
[3] L. Cai, J. Meng, D. I. Stroe, J. Peng, G. Luo, and R. Teodorescu, "Multi-objective optimization of data-driven model for Lithium-Ion battery SoH estimation with short-term feature," *IEEE Trans. Power Electron.*, vol. 35, no. 11, pp. 11855–11864, Nov. 2020.
[4] K. Goebel, B. Saha, A. Saxena, J. R. Celaya, and J. P. Christophersen, "Prognostics in battery health management," *IEEE Instrum. Meas. Mag.*, vol. 11, no. 4, pp. 33–40, Aug. 2008.
[5] Y. Xing, E. W. Ma, K.-L. Tsui, and M. Pecht, "An ensemble model for predicting the remaining useful performance of Lithium-Ion batteries," *Microelectronics Rel.*, vol. 53, no. 6, pp. 811–820, 2013.
[6] J. Wu, C. Zhang, and Z. Chen, "An online method for Lithium-Ion battery remaining useful life estimation using importance sampling and neural networks," *Appl. Energy*, vol. 173, pp. 134–140, 2016.
[7] H. Dai, G. Zhao, M. Lin, J. Wu, and G. Zheng, "A novel estimation method for the state of health of lithium-Ion battery using prior knowledge-based neural network and Markov chain," *IEEE Trans. Ind. Electron.*, vol. 66, no. 10, pp. 7706–7716, Oct. 2019.
[8] J. Wei, G. Dong, and Z. Chen, "Remaining useful life prediction and state of health diagnosis for Lithium-Ion batteries using particle filter and support vector regression," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5634–5643, Jul. 2018.
[9] D. Yang, X. Zhang, R. Pan, Y. Wang, and Z. Chen, "A novel Gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve," *J. Power Sources*, vol. 384, pp. 387–395, 2018.
[10] H. Chaoui and C. C. Ibe-Ekeocha, "State of charge and state of health estimation for lithium batteries using recurrent neural networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 8773–8783, Oct. 2017.
[11] X. Tang *et al.*, "A fast estimation algorithm for lithium-ion battery state of health," *J. Power Sources*, vol. 396, pp. 453–458, 2018.
[12] J. Wu, Y. Wang, X. Zhang, and Z. Chen, "A novel state of health estimation method of li-ion battery using group method of data handling," *J. Power Sources*, vol. 327, pp. 457–464, 2016.
[13] Z. Wang, S. Zeng, J. Guo, and T. Qin, "State of health estimation of lithium-ion batteries based on the constant voltage charging curve," *Energy*, vol. 167, pp. 661–669, 2019.
[14] G. Dong, W. Han, and Y. Wang, "Dynamic Bayesian network based Lithium-Ion battery health prognosis for electric vehicles," *IEEE Trans. Ind. Electron.*, to be published. doi:10.1109/TIE.2020.3034855.
[15] D. Stroe and E. Schaltz, "Lithium-Ion battery state-of-health estimation using the incremental capacity analysis technique," *IEEE Trans. Ind. Appl.*, vol. 56, no. 1, pp. 678–685, Jan. 2020.
[16] X. Hu, J. Jiang, D. Cao, and B. Egardt, "Battery health prognosis for electric vehicles using sample entropy and sparse Bayesian predictive modeling," *IEEE Trans. Ind. Electron.*, vol. 63, no. 4, pp. 2645–2656, Apr. 2016.
[17] A. Casari and A. Zheng, *Feature Engineering for Machine Learning*, 1st ed. Newton, MA, USA: O'Reilly Media, Inc., Apr. 2018.
[18] M. R. Amini, H. Wang, X. Gong, D. Liao-McPherson, I. Kolmanovsky, and J. Sun, "Cabin and battery thermal management of connected and automated HEVs for improved energy efficiency using hierarchical model predictive control," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 5, pp. 1711–1726, Sep. 2020.
[19] P. Kollmeyer and M. Skells, "Samsung inr21700 30t 3ah li-ion battery data," *Mendeley Data*, 2020, doi: 10.17632/9xyvy2njj3.1.
[20] M. Xu, Q. Meng, K. Liu, and T. Yamamoto, "Joint charging mode and location choice model for battery electric vehicle users," *Transp. Res. Part B: Methodological*, vol. 103, pp. 68–86, 2017.
[21] M. Dubarry *et al.*, "Identifying battery aging mechanisms in large format li ion cells," *J. Power Sources*, vol. 196, no. 7, pp. 3420–3425, 2011.
[22] M. D. Levi and D. Aurbach, "Simultaneous measurements and modeling of the electrochemical impedance and the cyclic voltammetric characteristics of graphite electrodes doped with Lithium," *J. Phys. Chem. B*, vol. 101, no. 23, pp. 4630–4640, 1997.
[23] Z. Zhan, J. Zhang, Y. Li, and H. S. Chung, "Adaptive particle swarm optimization," *IEEE Trans. Syst., Man, Cybern., Part B. (Cybern.)*, vol. 39, no. 6, pp. 1362–1381, Dec. 2009.
[24] M. S. Islam, N. Mithulananthan, and D. Q. Hung, "A day-ahead forecasting model for probabilistic EV charging loads at business premises," *IEEE Trans. Sustain. Energy*, vol. 9, no. 2, pp. 741–753, Apr. 2018.

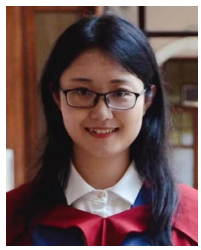
[25] Q. Xia *et al.*, "A modified reliability model for lithium-ion battery packs based on the stochastic capacity degradation and dynamic response impedance," *J. Power Sources*, vol. 423, pp. 40–51, 2019.
[26] S. Bashash, S. J. Moura, and H. K. Fathy, "On the aggregate grid load imposed by battery health-conscious charging of plug-in hybrid electric vehicles," *J. Power Sources*, vol. 196, no. 20, pp. 8747–8754, 2011.
[27] S. Voronov, E. Frisk, and M. Krysander, "Data-driven battery lifetime prediction and confidence estimation for heavy-duty trucks," *IEEE Trans. Rel.*, vol. 67, no. 2, pp. 623–639, Jun. 2018.

Ji Wu (Member, IEEE) received the B.E. degree in automation from the Hefei University of Technology (HFUT), Hefei, China, in 2011, and the Ph.D. degree in control science and technology from the University of Science and Technology of China, Hefei, China, in 2018.

From 2016 to 2017, he was a Guest Ph.D. Student with the Department of Energy Technology, Aalborg University, Aalborg, Denmark. Since 2018, he has been a Lecturer with the Department of Vehicle Engineering, HFUT, where he leads the Cyber Energy Laboratory. His areas of research include the modeling, control and optimization of the complex systems, including battery energy storage systems, electric vehicles and microgrids.

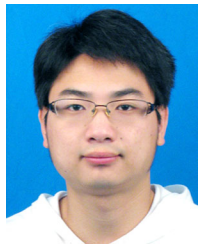
Xuchen Cui was born in Anhui, China, in 1998. He received the B.E. degree in mechanical design and manufacturing from the Tongling University, Tongling, China, in 2020. He is currently working toward the master's degree with the Department of Vehicle Engineering, Hefei University of Technology, Hefei, China.

His research interests include modeling and state estimation of lithium-ion batteries.



Hui Zhang received the M.S. degree in control science and technology from the University of Science and Technology of China, Hefei, China, in 2015, and the Ph.D. degree from the Department of Computer Science, The University of Hong Kong, Hong Kong, in 2020.

She has been a Visiting Scholar with the University of Stuttgart, Stuttgart, Germany, and the University of Tokyo, Tokyo, Japan, in 2017 and 2018, respectively. She is currently a Postdoctoral Fellow with the Department of Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. Her research interests include computer vision and graphics, deep learning, intelligent system control.



Mingqiang Lin received the B.E. degree in automation and the Ph.D. degree in control science and engineering from the University of Science and Technology of China, Hefei, China, in 2011 and 2016, respectively.

He is currently an Associate Professor with the Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Sciences, Jinjiang, China. His research interests include computer vision, pattern recognition, and health prognosis of energy storage system.