Remaining Useful Life Prediction for Lithium-Ion Batteries Based on Capacity Estimation and Box-Cox Transformation

Qiao Xue, Shiquan Shen, Guang Li, *Member, IEEE*, Yuanjian Zhang, *Member, IEEE*, Zheng Chen, *Senior Member, IEEE*, and Yonggang Liu, *Senior Member, IEEE*

***Abstract*—Remaining useful life (RUL) prediction of lithium-ion batteries plays an important role in intelligent battery management systems (BMSs). The current RUL prediction methods are mainly developed based on offline training, which are limited by sufficiency and reliability of available data. To address this problem, this paper presents a method for RUL prediction based on the capacity estimation and the Box-Cox transformation (BCT). Firstly, the effective aging features (AFs) are extracted from electrical and thermal characteristics of lithium-ion batteries and the variation in terms of the cyclic discharging voltage profiles. The random forest regression (RFR) is then employed to achieve dependable capacity estimation based on only one cell's degradation data for model training. Secondly, the BCT is exploited to transform the estimated capacity data and to construct a linear model between the transformed capacities and cycles. Next, the ridge regression algorithm (RRA) is adopted to identify the parameters of the linear model. Finally, the identified linear model based on the BCT is employed to predict the battery RUL, and the prediction uncertainties are investigated and the probability density function (PDF) is calculated through the Monte Carlo (MC) simulation. The experimental results demonstrate that the proposed method can not only estimate capacity with errors of less than 2%, but also accurately predict the battery RUL with the maximum error of 127 cycles and the maximum spans of 95% confidence of 37 cycles in the whole cycle life.**

***Index Terms*—Lithium-ion battery, remaining useful life, random forest regression, Box-Cox transformation, ridge regression, Monte Carlo simulation.**

Manuscript received June 10, 2020; revised October 7, 2020; accepted November 17, 2020. Date of publication November 23, 2020; date of current version January 22, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 51775063 and 61763021, in part by the National Key R&D Program of China under Grant 2018YFB0104000, and in part by the EU-funded Marie Skłodowska-Curie Individual Fellowships Project under Grant 845102-HOEMEV-H2020-MSCA-IF-2018. The review of this article was coordinated by Dr. Matthias Preindl. *(Corresponding authors: Zheng Chen; Yonggang Liu.)*

Qiao Xue and Shiquan Shen are with the Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming 650500, China (e-mail: strxue_qiao@163.com; shiquan219@gmail.com).

Guang Li is with the School of Engineering and Materials Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: g.li@qmul.ac.uk).

Yuanjian Zhang is with the Sir William Wright Technology Center, Queen's University Belfast, BT9 5BS Belfast, U.K. (e-mail: y.zhang@qub.ac.uk).

Zheng Chen is with the Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming 650500, China, and also with the School of Engineering and Materials Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: chen@kust.edu.cn).

Yonggang Liu is with the State Key Laboratory of Mechanical Transmissions & School of Automotive Engineering, Chongqing University, Chongqing 400044, China (e-mail: andylyg@umich.edu).

Digital Object Identifier 10.1109/TVT.2020.3039553

NOMENCLATURE

A. Acronyms

EVs	electric vehicles
DOD	depth of discharge
EOL	end of life
RUL	remaining useful life
PF	particle filter
EIS	electrochemical impedance spectroscopy
ML	machine learning
RFR	random forest regression
NN	neural network
LSTM	long short-term memory
SVM	support vector machine
GPR	Gaussian process regression
RVM	relevance vector machine
EKF	extended Kalman filter
IC	incremental capacity
SWS	sliding window size
AR	autoregression
PSO	particle swarm optimization
AFs	aging features
BCT	Box-Cox transformation
Ah	Ampere hour
CC	constant current
CV	constant voltage
SOH	state of health
IR	internal resistance
DIC	discharge incremental capacity
EW	entropy weight
GRA	grey relational analysis
MC	Monte Carlo
RF	random forest
CART	classification and regression tree
OOB	out-of-bag
MAE	mean absolute error
PDF	probability density function
ME	maximum absolute error
RMSE	root-mean-square error
STD	standard derivation

B. Symbols

$cycle_{EOL}$	cycle number of the end of life
$cycle_{now}$	current cycle number
F_1	battery internal resistance
F_2	average temperature of each cycle
F_3	peak absolute value of discharge incremental capacity curves
ξ	grey correlation grades
i	cycle number
X_i	aging features data set
Y_i	capacity data set corresponding to the X_i
S_t	original sample set
q	number of features
t	number of samples
Θ_t	a family of independent and identically distributed random vectors
T	number of prediction trees
λ	transformation parameter of Box-Cox transformation
$Q(\lambda)$	transformation values of Box-Cox transformation
β	coefficients of the linear model
ε_i	independent random error
σ^2	variance of ε_i
$J(\lambda, Q)$	Jacobian matrix corresponding to λ and Q
I	identity matrix
k	ridge regression coefficient
ρ	correlation coefficient
$\hat{y}$	fitting value of linear model
y	observation value of capacity
μ_y	mean value of y
σ_y	standard deviation of y
$\mu_{\hat{y}}$	mean value of $\hat{y}$
$\sigma_{\hat{y}}$	standard deviation of $\hat{y}$
$\hat{f}_h(\tilde{Q})$	probability density function of RUL prediction
$K_p(\cdot)$	Gaussian kernel function
h_p	band width of $K_p(\cdot)$
U_c	upper bounds of Monte Carlo simulation.
L_c	lower bounds of Monte Carlo simulation
$\tilde{Q}_i$	the ith result of RUL prediction
R^2	goodness-to-fit parameter

I. Introduction

TO MITIGATE worldwide energy crisis, environmental pollution and global warming problems, electric vehicles (EVs) are being rapidly developed [1]. Lithium-ion batteries have been widely considered as suitable power sources of EVs due to their high energy density, long cycle life, lower self-discharge rate, light weight and no memory effect [2]. However, complex operation conditions such as different load current rate, varying temperature and stochastic depth of discharge (DOD) generate significant influence on electrical performance of lithium-ion batteries [3]. Thus lithium-ion batteries applied in EVs can reach their end of life (EOL) [4] earlier than intended. Generally, lithium-ion batteries reach their EOL when the capacity drops to 80% of rated value in vehicular applications [5]. To monitor proper operation of batteries, it is necessary to develop advanced techniques to predict remaining useful life (RUL) of lithium-ion batteries so that end-users can know the operating status in advance and can replace the batteries timely.

Prediction of RUL can be made by regression analysis based on historical operation data. Currently, RUL prognostics methodologies can be divided into mechanism analysis methods and data-driven methods [6]. Mechanism analysis methods are generally implemented to predict battery RUL based on a nonlinear aging model combined with an effective filter. Lyu *et al.* [7] exploits the particle filter (PF), together with the electrochemical model, to predict RUL of batteries. Selina *et al.* [8] investigates the modeling of battery degradation under different operation conditions and ambient temperatures and proposes a simple Bayes model for RUL prediction considering different ambient temperature and discharge current. Bhaskar *et al.* [9] develops a Bayesian learning framework for RUL prediction, where the aging mode is constructed based on the features extracted from the electrochemical impedance spectroscopy (EIS), and the PF is leveraged to update model parameters and predict the battery RUL. In [10], an exponential model for lithium-ion battery capacity is first constructed to assess capacity degradation. Then, a spherical cubature-based PF is introduced to solve the exponential model. After that, the model extrapolation to a specified failure threshold is performed to infer the RUL of lithium-ion batteries. Wang *et al.* [11] develops a conditional three-parameter capacity degradation model for RUL prediction. The parameters of established model are calculated by nonlinear least squares regression based on capacity degradation training data, and then the RUL is estimated via extrapolating the model. Yang *et al.* [12] establishes a coulombic efficiency model to capture the convex degradation trend of lithium-ion phosphate batteries, and the PF framework is constructed to update the model parameters. Then, the RUL is predicted by extrapolating the models with renewed parameters. Although mechanism analysis methods are clear to describe the degradation trend of batteries, they involve a number of parameters and complex calculation for accurate modeling of the RUL variation law. In consequence, it is not quite suitable for real-time prediction instead it is more appropriate for theoretical research on battery designation [6].

The data-driven methods do not require accurate analysis of degradation mechanism. Such methods can capture effective feature information from battery operation data that can be measured by external sensors and then predict battery RUL based on machine learning (ML) algorithms [13]. Li *et al.* [14] extracts feature vectors from partial charging voltage curves and attains precise capacity estimation based on the extracted features and random forest regression (RFR). Li *et al.* [15] proposes a fusion method for battery RUL prediction by combining the Elman neural network (NN) and long short-term memory (LSTM) to predict high and low frequency sub-layers. Support vector machine (SVM) [16] and Gaussian process regression (GPR) [17] are two commonly employed methods in terms of RUL prediction. Patil *et al.* [18] presents a multistage SVM approach for RUL prediction of lithium-ion batteries. It inherits the classification and regression attributes of SVM, and the classification model provides general estimation and the regression

model refines the RUL prediction in turn. Guo *et al.* [19] introduces a remaining capacity estimation method based on fourteen health features extracted from the charging data. These health features are determined using principal component analysis, and then relevance vector machine (RVM) is employed to attain capacity estimation. Zhou *et al.* [20] combines the extended Kalman filter (EKF) with GPR to estimate the available capacity online according to the daily partial charging data. Li *et al.* [21] extracts the health features from partial incremental capacity (IC) curves, and then the GPR is implemented to achieve the short-term SOH estimation and long-term RUL prediction. In addition, NN [22] and time series methods [23] are also employed to predict RUL of lithium-ion batteries. Ren *et al.* [24] investigates a fused deep learning approach, combining auto-encoder with deep NN, for battery's RUL prediction. To address the selection principle of sliding window sizes (SWS), which is often defined empirically, Ma *et al.* [25] applies the false nearest neighbor method to calculate the SWS required for prediction and employs a hybrid NN to predict the battery RUL. Long *et al.* [26] establishes an autoregression (AR) model for RUL prediction of lithium-ion batteries and leverages the particle swarm optimization (PSO) algorithm to optimize the order of AR model. Compared with mechanism analysis methods, data-driven methods usually entail a large amount of offline training data to construct an accurate online RUL predictor [27].

To accelerate the modeling process and reduce computation burden, limited offline data are usually utilized to construct and train the degradation model in practical applications, and typical works only exploit part of the battery's capacity degradation data to build RUL prediction models, such as exponential models [28] and polynomial models [29]. However, the degradation rate of capacity varies significantly throughout the whole cycle life. Generally, the capacity degradation slope is relatively gentle in the early life phase and yet shows an exponential decline trend with faster dropping speed in the later life stage [30]. From this point of view, the model based on partial lifecycle data cannot accurately track the degradation trend in the whole lifespan and will lead to increase of RUL and EOL prediction error. To cope with this limitation, the whole lifecycle capacity data should be trained comprehensively, such as by ML algorithms, and then a prediction model can be constructed to effectively estimate the RUL of battery. This two-step prediction process can not only achieve the target of accurately predicting RUL, but also diagnose the health status of the battery in real time through the estimated capacity. Motivated by this, a RUL prediction method based on capacity estimation is developed. To precisely estimate the battery capacity for RUL prediction, three aging features (AFs) are extracted from electric and thermal characteristics curves and discharge IC curves of batteries. Owing to the qualified estimated performance and reliable identification ability of relevant variables and interactions, the RFR is exploited to estimate the capacity of battery under the whole lifespan [14]. However, the process of capacity degradation is nonlinear, and it leads to difficulty of capture the degradation trend in a mathematical manner. To cope with it, a linear model between the estimated capacities and cycle number is established by means of the Box-Cox transformation (BCT), which can contribute to the prediction accuracy improvement of RUL based on the estimated capacity [13]. Finally, the battery RUL can be predicted through extrapolating the linear model. The main contributions of this study can be attributed to the following three aspects: 1) Three AFs are extracted from electrical and thermal characteristics curves and discharge IC curves to improve the precision of capacity estimation. 2) The RFR is applied to achieve the precise capacity estimation of other cells by training only one cell's data. 3) The BCT is employed to construct a linear model between the estimated capacity and cycles to ensure the RUL prediction accuracy. It enables prediction of the battery EOL in its early life stage based on the constructed linear model.

TABLE I
THE SPECIFICATIONS OF TEST BATTERY

Type	APR18650M1A
Material	$LiFePO_4$/graphite
Dimension (D×H)	18 mm×65 mm
Nominal Capacity	1.1 Ah
Nominal Voltage	3.3 V
Allowed voltage range	2.0-3.6 V
Charge/Discharge Temperature	-30 °C-60 °C
Storage Temperature	-50 °C-60 °C

The remainder of this study is arranged as follows. The battery life cycle test is introduced, and the experimental data is analyzed in Section II. Section III illustrates the detailed algorithms for RUL prediction. The capacity estimation process is elaborated, and the estimation results are discussed in Section IV, followed by the analysis and discussion with respect to the RUL prediction in Section V. Finally, Section VI concludes the study.

II. BATTERY AGING TESTING AND DEGRADATION ANALYSIS

In this paper, the RUL is studied to assess the battery operating performance and estimate the available remaining service time left before EOL. In this study, RUL is defined as the difference between the cycle number of EOL $cycle_{EOL}$ and the current cycle number $cycle_{now}$, as:

$$RUL = cycle_{EOL} - cycle_{now} \tag{1}$$

A. Battery Aging Experiment and Degradation Data Analysis

In this study, the cyclic aging data of lithium-ion batteries are obtained from an open source [31], which was collected by cyclic life tests of a variety of commercial lithium iron phosphate/graphite batteries. The rated capacity of cells is 1.1 Ampere hour (Ah), the rated voltage is 3.3 V and their specifications are tabulated in Table I. These cells were cycled in horizontal cylindrical fixtures on an Arbin battery test equipment after being placed in a thermal controlled chamber, whose temperature is set to 30 °C. The detailed program of battery life cycle test is shown in Fig. 1. As can be found, the experiments adopt two-step fast-charging policy to charge the battery, and the upper and lower cut-off voltages are set to 3.6 V and 2.0 V, respectively. The charging policy specifies a C1(Q1)-C2 mode, where C1 and C2 denote the first step and second step current, respectively; and Q1 is the SOC at which the current changes. The second current

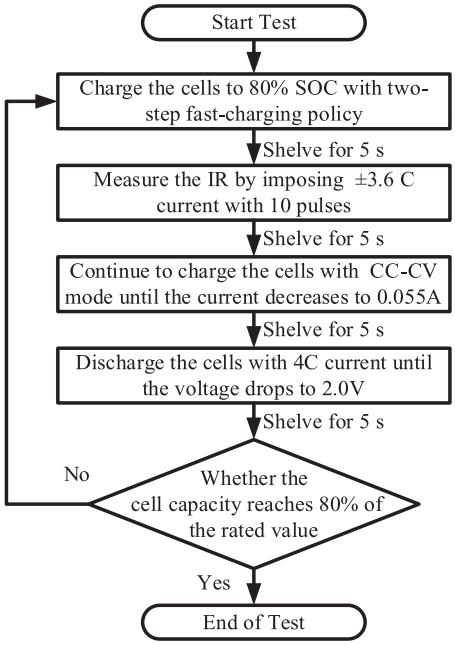

Fig. 1. The flowchart of battery aging test procedure.

step ends at 80% SOC, after that the cell is charged with 1C constant current (CC)-constant voltage (CV) mode, followed by the discharge test with 4C current, where C denotes the rated capacity value. During the experiment, the surface temperature and internal resistance of the battery are also measured. The temperature measurement is performed by attaching a Type-T thermocouple to the exposed surface, and the internal resistance measurement is conducted during charge at 80% SOC by imposing some pulses. Since large current excitation can lead to more obvious voltage variation and consequent more precise internal resistance estimation, the manufacturer's recommended fast-charging rate, i.e., 3.6C was chosen as the pulse current rate. In this study, 10 charge/discharge current (±3.6 C) pulses, each of which lasts 33 ms, are imposed to achieve the internal resistance measurement. Moreover, the charge/discharge current rate and the voltage cutoffs used in this work also follow the recommendation supplied by the manufacturer.

In this paper, the cyclic experiment data of 7 batteries (labeled as Cells 1 to 7) are selected from the data repository to construct and evaluate the RUL prediction algorithm. The curves of degradation capacity are shown in Fig. 3(a), which highlights that the degradation trajectories of seven cells remain almost the same, indicating that the degradation mechanism is nearly consistent for the same type of lithium-ion batteries. The cycle life experiments for all batteries are terminated when the batteries reached 80% of nominal capacity, i.e., 0.88 Ah. It can also be found that the degradation slope is relatively small before 90% state of health (SOH), which is defined as the ratio of current maximum available capacity over the nominal value, as shown in (2). To intuitively show the capacity decline speed, the degradation rate is calculated according to (3), and the relationship between the degradation rate and SOH is shown in Fig. 2. As can be found, the capacity degradation rate is smaller, i.e., less than 0.04%, before 90% SOH; whereas the capacity degradation rate shows a faster speed when SOH drops less than 90%. Besides,

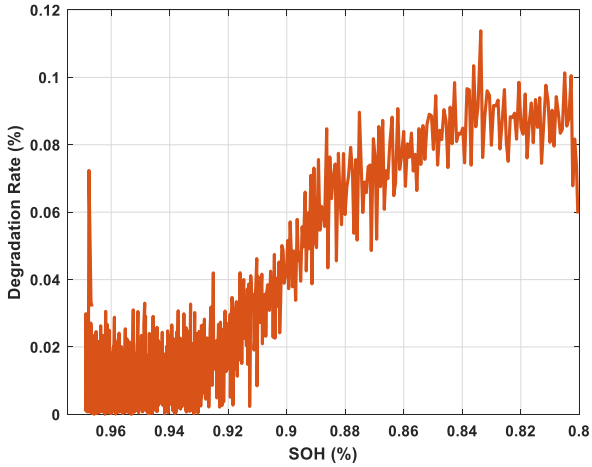

Fig. 2. Capacity degradation rate with different SOH.

the electric characteristics will gradually deteriorate during the aging process, the mechanical and thermal characteristics of batteries also vary with aging. For example, the thickness may increase due to gas generation; the heat transfer coefficient and entropic potential may also change during the degradation [30]. Next, the AFs will be extracted from electric characteristics and thermal characteristics variation of the battery.

$$SOH = \frac{Q_{cur}}{Q_{nom}} \times 100\% \tag{2}$$

$$Rate = \frac{Q_i - Q_{i+1}}{Q_i} \times 100\% \tag{3}$$

B. Extraction of Aging Features

From the perspective of electric characteristics, one main change during degradation is that the internal resistance (IR) will gradually increase, as shown in Fig. 3(b). Therefore, the battery IR, denoted by F_1, can be selected as one AF. Considering the battery's thermal characteristics, the battery surface temperature at each moment is recorded by attaching a T-type thermocouple to the battery surface during experiment. On this basis, the variation of battery surface temperature at each cycle is utilized to characterize the battery thermal characteristics, instead of establishing a heat transfer model. The variation of average temperature with different cycle times is shown in Fig. 3(c). It is obviously observed that the average temperature increases progressively with the cycle number. Intuitively, the average temperature of each cycle can be selected as another AF F_2. Meanwhile, the discharge incremental capacity (DIC) curves at different cycles are shown in Fig. 3 (d). Distinct variation can easily reveal that the absolute value of peak decreases with the increment of cycle number and reduction of capacity. Thus, it can also be considered as one AF, called F_3.

To sum up, three AFs, including the IR F_1, average temperature F_2 and the absolute value of DIC peak F_3, are extracted to estimate battery capacity based on the tested data set. These three AFs with respect to cycle number are shown in Fig. 3(b), (c) and (e). In practice, it is often difficult to determine a proper weight due to the sparsity of the indexes. To determine the contribution weight and intuitively evaluate the dispersion degree of AFs, the

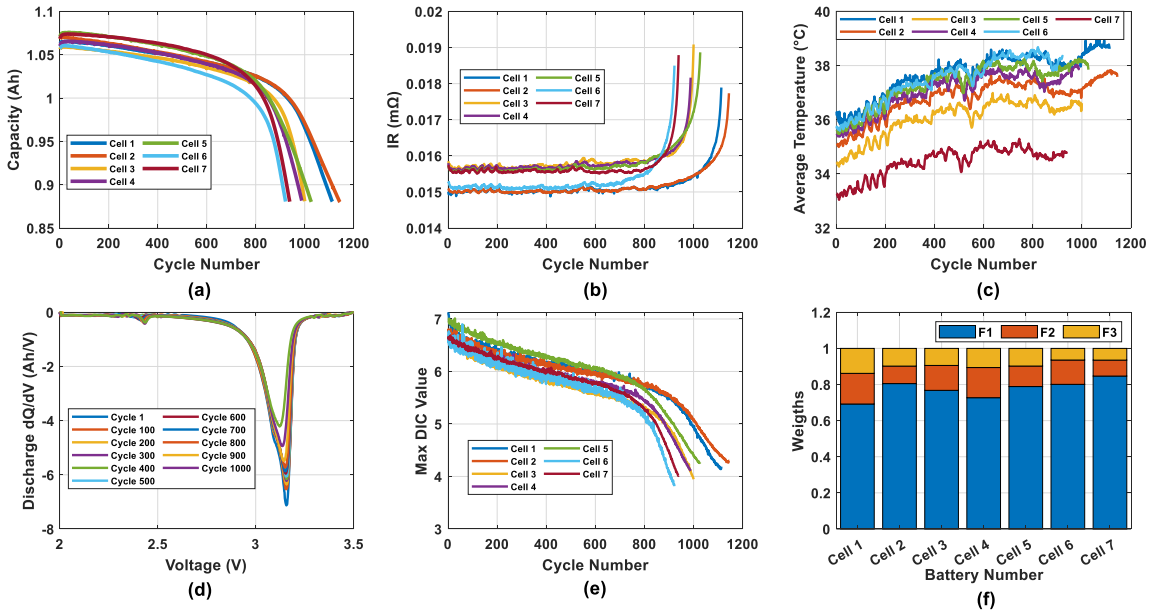

Fig. 3. The evolution trend with cycle number of capacity and AFs.

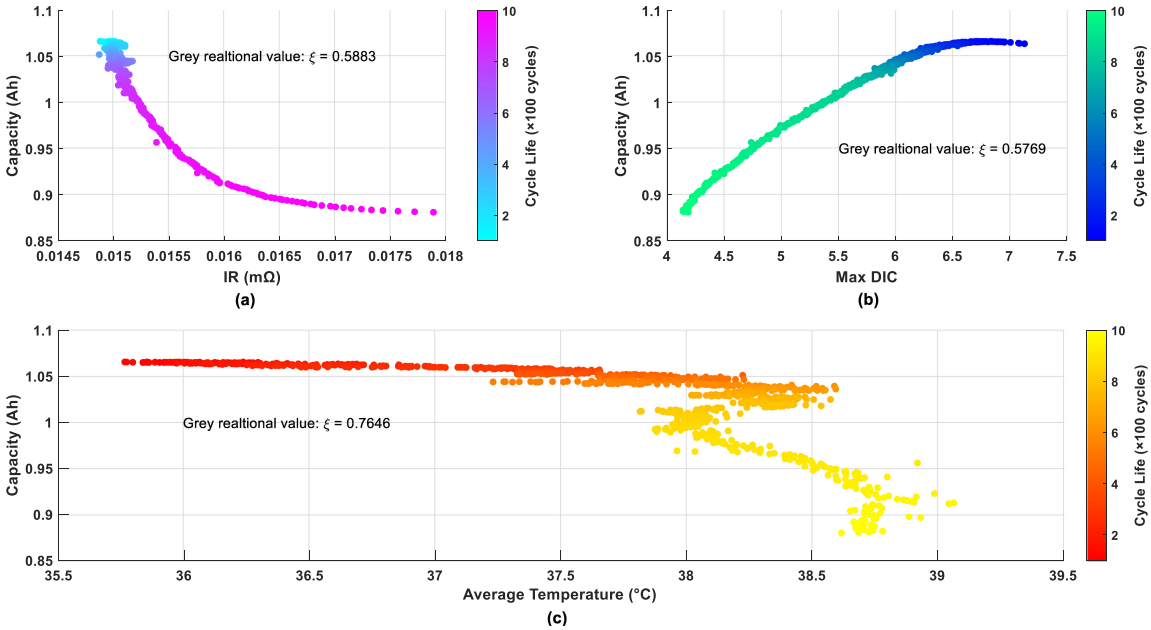

Fig. 4. The evolution relationship between capacity and AFs with cycle numbers of cell 1.

entropy weight (EW) method is firstly introduced to analyze the correctness of AFs extraction. Smaller entropy value indicates higher dispersion of corresponding AF and more impact on capacity, and vice versa. The detailed calculation process of EW method can be referred to [32]. The sum of weights of three AFs is equivalent to 1, and the entropy weight of each AF is evaluated and compared, as shown in Fig. 3(f). It can be observed that the EW value of F_1 is largest and greater than 0.7; the EW value of F_3 is least and lower than 0.2, indicating that F_1 contributes the most weight to the capacity prediction, whereas F_3 raises the least weight. It can be also seen that the EW values distribution of three AFs for seven cells are consistent. The EW value of AFs indicates that they have different impact weights on the capacity. Next, the implied relationships between AFs and capacity are analyzed.

C. Analysis of Aging Features Based on GRA

To analyze the relationship between AFs and capacity, we take cell 1 as an example, and the variation relationships between AFs and capacity with respect to cycle life are shown in Fig. 4, where the color scale represents cycle life. As can be found, the extracted AFs highlight different variation trend with the decrease of capacity. Among them, F_1 and F_2 increase and F_3

TABLE II
THE DIVISION OF VALUE INTERVAL FOR RELATIONAL GRADE

Value Interval	Relational Grade
[0 0.2)	Very weak or no correlation
[0.2 0.4)	Weak correlation
[0.4 0.6)	Moderate correlation
[0.6 0.8)	Strong correlation
[0.8 1.0]	Extremely strong correlation

TABLE III
GRA BETWEEN AGING FEATURES AND CAPACITY

Battery Number	Aging Features		
	F_1	F_2	F_3
Cell 1	0.5883	0.7646	0.5769
Cell 2	0.5902	0.7819	0.6022
Cell 3	0.5690	0.7818	0.5793
Cell 4	0.5760	0.7707	0.5873
Cell 5	0.5836	0.7593	0.5976
Cell 6	0.5753	0.7463	0.5873
Cell 7	0.5625	0.7778	0.5982

decreases with the capacity degradation. Additionally, three AFs show different increase/decrease rates with different cycle life phase. The different segments corresponding to the specific cycle region, such as early/middle/late phases, is determined based on the capacity degradation curves with respect to the cycle number. It can be seen from Fig. 4 that in the early and middle phases of cycle life (1 to 600 cycles), the capacity degrades with a slow speed, so that F_1 remains almost unchanged, and in contrast, F_2 increases obviously and F_3 gradually decreases with the increasing of cycle numbers. Comparatively, in the later phase of cycle life (600 to 1000 cycles), the capacity degradation and the increase of F_1 are rapid, and the increase rate of F_2 becomes slower and more stabilized; however, F_3 still decreases obviously. It can be concluded that the change of F_1 is not obvious, while the variation of F_2 is relative larger in the early cycle life. In the later cycle life stage, the changes of F_1 and F_2 are opposite to that of the early stage. Moreover, there exists obvious variation in F_3 throughout the whole cycle life. In this study, the correlation between AFs and battery capacity is further evaluated by grey relational analysis (GRA). As a crucial method based on the grey system theory, the GRA evaluates the correlation among the elements according to the similarity and dissimilarity of their variation trend. The intension of employing GRA is to evaluate the relationship between different curves by studying the geometric proximity, and higher proximity implies stronger correlation. For battery capacity estimation, the AF curves extracted from new cells are defined as the reference for capacity estimation. The quantitative analysis based on the GRA is to obtain the correlations between reference and comparative sequences, as detailed in [33]. By the GRA, the correlation grades, namely ξ, between the three AFs and capacity of each cell are acquired. To more precisely evaluate the correlation grade, the value interval of corresponding to the specific relational grade is further divided, as shown in Table II. As can be seen, [0 0.2) represents very weak or no correlation, and [0.8 1.0] means extremely strong correlation. The value of for three AFs are shown in Table III, highlighting that F_1 and F_3 have moderate correlation with capacity, but F_2 shows strong correlation with capacity. Particularly, the ξ for F_2 is greater than 0.75 for most of the cells, which means the selection of AFs is effective for capacity estimation.

D. The Framework and Flowchart for RUL Prediction

In this study, the capacity is firstly estimated and then utilized to predict the battery RUL. The prediction framework is illustrated in Fig. 5. As can be seen, the whole prediction process contains the capacity estimation module and the RUL prediction module. In the capacity estimation module, the AF data set $X_i = [F_{1i}, F_{2i}, F_{3i}]$ is firstly extracted from the aging experimental data set. Then, the whole cycle life data of cell 1 is used as the training set to build the RFR model, the X_i data set and corresponding capacity Y_i are considered as the RFR model's input and output, respectively; and i denotes the cycle number. The optimal model parameters are searched via testing and cross validation. The well-tuned model is applied to attain the precise estimation of capacity. In the RUL prediction module, the BCT is introduced to transform the estimated capacity data to construct a linear model between the transformed capacities and cycles. The RRA is then employed to identify the linear model parameters. The constructed linear model using BCT is extrapolated to predict the battery RUL, and the RUL prediction uncertainties are generated using the MC simulation.

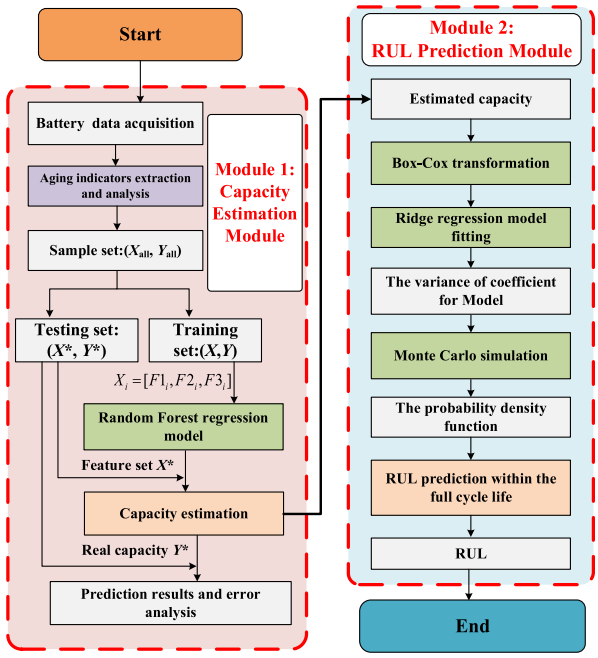

Fig. 5. The framework of capacity estimation and RUL prediction.

III. ALGORITHMS

This section introduces the related algorithms, including the RFR, BCT, RRA and MC simulation, for capacity estimation and RUL prediction of lithium-ion batteries.

A. Random Forest Regression

Random forest (RF) [34] is a ML algorithm based on decision trees, which generates hundreds and even thousands of decision trees in the classification or regression process. A decision tree, also called classification and regression tree (CART), is a nonparametric model that consists of decision nodes and leaf nodes. Based on the bagging algorithm [35], RF combines multiple weak classifiers to make the whole model possess with higher accuracy and generalization ability. Bagging algorithm (also called bootstrap aggregation) can improve the prediction performance of regression methods via reducing the variance. On this account, the Bagging algorithm is employed, together with the RFR, to improve the prediction performance of RUL by combining all the generated decision trees. During the algorithm training, multiple subsets are collected by randomly sampling with replacement from the original sample data set to train the classifiers. Here, one assumes the original sample set S_t as:

$$S_t = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_t, Y_t)\}, X \in R^t, Y \in R \tag{4}$$

where X represents the input vector containing q features, i.e., $X = \{x_1, x_2, \ldots, x_q\}$, Y is the output scalar and t denotes the number of samples. In each random sampling process, every observation may not be selected with the probability of $(1 - \frac{1}{t})$, thus the unselected probability of each observation in t times is $(1 - \frac{1}{t})^t$. When $t \to \infty$, we can attain:

$$\lim_{t\to\infty} \left(1 - \frac{1}{t}\right)^t = \frac{1}{e} \approx 0.368 \tag{5}$$

In consequence, about 36.8% of original sample data will not be selected in the bagging process. The samples that are not selected are included as part of another subset called out-of-bag (OOB) samples. Accordingly, when a regression tree is constructed, two thirds of the training samples are exploited to construct the regression function, and the remaining one third data are used to constitute the OOB sample. Since the OOB samples are not leveraged to train and fit the model, they can be used to evaluate the performance of the regression tree. By this manner, RFR can give an unbiased estimation for the generalization error without the help of external data subsets, compared with other regression methods, such as SVM and GPR. Moreover, the built-in validation attribute of RFR can largely reduce the possibility of overfitting and improve the generalization capability.

RFR is an algorithm composed of a set of regression decision subtrees $\{h(X, \Theta_t), t = 1, 2, \ldots, T\}$, where Θ_t is a family of independent and identically distributed random vectors, and T indicates the number of decision trees. The flowchart of constructing the RFR model is shown in Fig. 6. Note that the randomly collected sample process of RFR is called 'bootstrap'. A prominent advantage of RFR is that it only needs to tune two parameters, i.e., number of trees n_{tree} and number of random features n_{fea} for each split in the forest to build [14]. Consequently, only limited effort is imperative for fine-tuning parameters to achieve anticipated performance. It can be seen from Fig. 6 that the first step of building an RFR model is to draw n_{tree} bootstrap samples from the original data set S_t. Secondly, an unpruned regression tree will be grown using the bootstrap

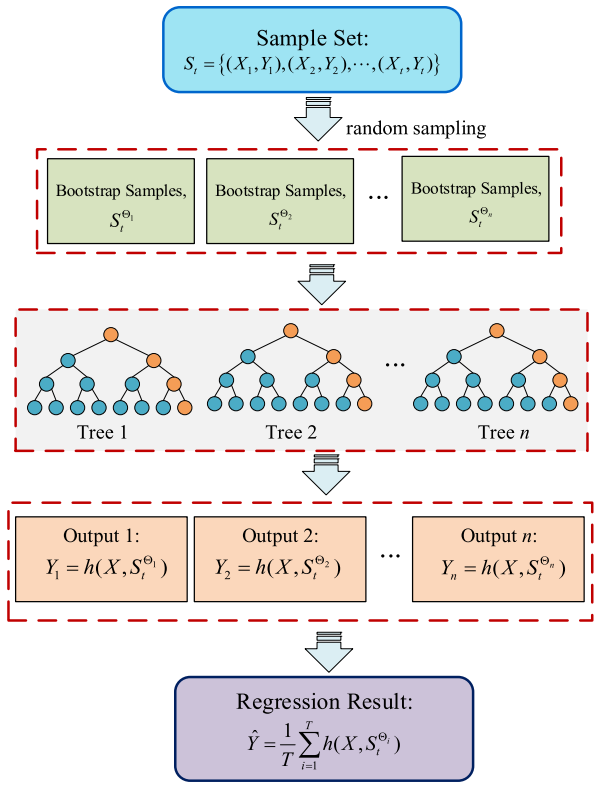

Fig. 6. The illustration of random forest regression model.

TABLE IV
THE STEPS OF RFR FOR PREDICTION

Step 1. Draw bootstrap samples set $(S_t^{\Theta_1}, S_t^{\Theta_2}, \cdots, S_t^{\Theta_T})$ based on the Bagging thought;
Step 2. Construct the regression decision subtree by randomly sampling feature to split the node of each tree;
Step 3. Repeat steps 1) and 2) to grow T regression trees, each tree grows randomly without pruning, and finally generate a 'forest';
Step 4. The aggregation is performed by averaging the outputs of T trees, and the estimation output can be obtained by (6).

sample $S_t^{\Theta_i}$. In this process, n_{fea} samples of the predictors will be randomly selected, and the best split will be chosen from those n_{fea} variables at each node, rather than from all predictors. Finally, new data will be predicted by aggregating the predictions of n_{tree} trees. The basic steps of applying RFR for prediction are also concisely summarized in Table IV. For the randomly collected sample process, a bootstrap sample is obtained by randomly selecting t observations with replacement from original sample set S_t. The bagging algorithm selects bootstrap samples $(S_t^{\Theta_1}, S_t^{\Theta_2}, \ldots, S_t^{\Theta_T})$ and applies the previous tree decision algorithm to construct a collection of T prediction trees$\{h(X, S_t^{\Theta_1}), \ldots, h(X, S_t^{\Theta_T})\}$. The ensemble produces T outputs corresponding to each tree, as:$Y_1 = h(X, S_t^{\Theta_1})$, $Y_2 = h(X, S_t^{\Theta_2})$,...,$Y_n = h(X, S_t^{\Theta_n})$. The prediction output of RFR is attained by performing the aggregation for the average of outputs of all trees, as:

$$\hat{Y} = \frac{1}{T}\sum_{i=1}^{T} Y_i = \frac{1}{T}\sum_{i=1}^{T} h(X, S_t^{\Theta_i}) \tag{6}$$

where Y_i is the output of the ith tree, and $i = 1, 2, \ldots, T$.

B. Box-Cox Transformation

The main target of BCT is to conduct monotonic transformation of data, thereby achieving normality in highly skewed imputed values [36]. Due to the nonlinear capacity degradation trend and the linear RUL decline rate with cycle number, accurate RUL prediction based on only the original capacity degradation data is rather difficult to attain. Therefore, this study exploits the BCT, which needed only one parameter to be identified, to transform nonlinear capacity degradation into linear degradation to improve the RUL prediction performance. The BCT, as originally introduced in [37], applies the following equation when $Q > 0$, as:

$$Q(\lambda) = \begin{cases} \frac{Q^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ \log Q, \lambda = 0 \end{cases} \tag{7}$$

where $Q(\lambda)$ represents the transformed values and λ is the transformation parameter that needs to be identified. Through applying the BCT, a linear model corresponding to the observations can be constructed, i.e., $Q(\lambda) \sim N(X, \beta, \sigma^2)$, as:

$$\begin{cases} Q(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_n + \varepsilon_i \\ \varepsilon_i = N(0, \sigma^2), i = 1, 2, \ldots, n \end{cases} \tag{8}$$

where X is a design matrix with $X = (x_1, x_2, \ldots, x_n)^T$, $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_m)^T$ are the coefficients of the linear model, and ε_i denotes the independent random error that is normally distributed with zero mean and variance of σ^2. Generally, λ is identified by the maximum likelihood method [38]. Since the transformation responses $Q(\lambda) \sim N(X\beta, \sigma^2)$, for the fixed λ, the log-likelihood function of β and σ^2 is expressed as:

$$L(\beta, \sigma^2) = \frac{\exp\left\{-\frac{1}{2\sigma^2}[Q(\lambda) - X\beta]^T[Q(\lambda) - X\beta]\right\}}{\left(\sqrt{2\pi}\sigma\right)^n} J(\lambda, Q) \tag{9}$$

where $J(\lambda, Q)$ is the Jacobian matrix of the transformation process, as:

$$J(\lambda, Q) = \prod_{i=1}^{n} \left| \frac{dQ_i(\lambda)}{dQ_i} \right| = \prod_{i=1}^{n} Q_i^{\lambda - 1} \tag{10}$$

In (9), when λ is fixed, J is a constant factor that is independent of β and σ^2. Taking the partial derivatives of $L(\beta, \sigma^2)$ with respect to β and σ^2, and setting each of the resulting equations to zeros, we can get:

$$\hat{\beta}(\lambda) = (X^T X)^{-1} X^T Q(\lambda) \tag{11}$$

$$\hat{\sigma}^2(\lambda) = \frac{[Q(\lambda) - X\hat{\beta}]^T [Q(\lambda) - X\hat{\beta}]}{n} \tag{12}$$

By Substituting $\hat{\beta}(\lambda)$ and $\hat{\sigma}^2(\lambda)$ into (9), the corresponding maximum likelihood value of λ can be calculated, as:

$$L_{\max}(\lambda) = L\left(\hat{\beta}(\lambda), \hat{\sigma}^2(\lambda)\right) = (2\pi e) \cdot J(\lambda, Q) \cdot [\hat{\sigma}^2(\lambda)]^{\frac{n}{2}} \tag{13}$$

Take the logarithm transformation of both sides, we can get:

$$\log\left(L_{\max}(\lambda)\right) = \log\left((2\pi e) \cdot J(\lambda, Q) \cdot [\hat{\sigma}^2(\lambda)]^{\frac{n}{2}}\right) \tag{14}$$

By submitting (10) into (14) and omitting the constant term irrelevant to λ, equation (14) can be changed into:

$$L^*(\lambda) = \log\left(L_{\max}(\lambda)\right) = \frac{n}{2}\log[\hat{\sigma}^2(\lambda)] + (\lambda - 1)\sum_{i=1}^{n}\log(Q_i) \tag{15}$$

Since $\log(x)$ is a monotone increasing function, when the unary function of λ reaches the maximum value $L_{\max}(\lambda)$, $\log(L_{\max}(\lambda))$ also gets the maximum value. Therefore, maximizing (13) is equivalent to maximizing (15). After λ is determined, equation (7) is exploited to transform the battery capacity variation.

C. Ridge Regression Algorithm

In the anterior step, a linear model between the transformed capacities and cycles is constructed through the BCT. Next, the model parameters need to be identified precisely. RRA sacrifices unbiasedness to gain high numerical stability, and thus obtains higher calculation accuracy. As an effective parameter estimation method, RRA is commonly used to address the collinearity problem frequently arising in multiple linear regression problems [39]. Note that the L2 regularization is added to the loss function of RRA to avoid overfitting. In view of the robust linear regression capability, the RRA is employed to estimate the parameters of linear model between the transformed capacities and cycles. The parameters of the regression equation are solved as follows. A standard model for the linear regression is considered, as:

$$\hat{y} = X\beta + \varepsilon \tag{16}$$

The corresponding loss function can be expressed as:

$$D(\beta) = \|\hat{y} - y\|_2^2 + k\|\beta\|_2^2 = (X\beta - y)^T(X\beta - y) + k\beta^T\beta \tag{17}$$

Taking the partial derivative of $D(\beta)$ with respect to β, as:

$$\frac{\partial D(\beta)}{\partial \beta} = 2X^T X\beta - X^T y - X^T y + 2k\beta \tag{18}$$

and setting $\frac{\partial D(\beta)}{\partial \beta} = 0$, the model parameter of (16) can be yielded:

$$\beta = (X^T X + kI)^{-1} X^T y \tag{19}$$

where I is the identity matrix and k is the ridge regression coefficient. The correlation coefficient ρ and mean absolute error (MAE) are employed to quantitatively evaluate the effectiveness of linear relationship fitting, as:

$$\rho(y, \hat{y}) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\overline{y_i - \mu_y}}{\sigma_y}\right)\left(\frac{\hat{y}_i - \mu_{\hat{y}}}{\sigma_{\hat{y}}}\right) \tag{20}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{21}$$

where y represents the observation value, $\hat{y}$ is the fitting value of linear model, μ_y and σ_y are the mean and standard deviation of y respectively, $\mu_{\hat{y}}$ and $\sigma_{\hat{y}}$ are the mean and standard deviation of $\hat{y}$ respectively.

D. Monte Carlo Simulation

The transformed capacities are not strictly linear with the cycles, there still exist deviations between the established linear model and the actual value. The uncertainties caused by the model error will be exaggerated with the prediction algorithm, and it will eventually lead to the prediction error of the RUL. Therefore, it is of crucial significance to describe the uncertainties of RUL prediction results as much as possible. The MC simulation can propagate the input uncertainties into prediction uncertainties [40]. Usually, MC simulation is combined with different prediction methods to calculate the probability density function (PDF) of prediction. Therefore, this study employs the MC simulation to calculate the PDF for RUL prediction. The procedures of calculating PDF for RUL prediction is described as:

1) Determine the main source of uncertainties of RUL prediction method. In this study, the uncertainties are generated in the process of parameters identification of linear regression;
2) Determine the distribution regularities of uncertainties based on the mean and variance of linear model parameters during the fitting;
3) Randomly generate several samples according to distribution regularities of uncertainties, and then perform simulation prediction in terms of each generated sample using the linear regression model.
4) Based on the simulation prediction results solved by step 3, the PDF of RUL prediction can be calculated as:

$$\hat{f}_h(\tilde{Q}) = \frac{1}{N}\sum_{i=1}^{N}\left[K_p\left(\frac{\tilde{Q}-\tilde{Q}_i^-}{h_p}\right) + K_p\left(\frac{\tilde{Q}-\tilde{Q}_i}{h_p}\right) + K_p\left(\frac{\tilde{Q}-\tilde{Q}_i^+}{h_p}\right)\right] \tag{22}$$

where $\hat{f}_h(\tilde{Q})$ is the PDF of RUL prediction, $K_p(\cdot)$ denotes the Gaussian kernel function, and h_p is the band width. $\tilde{Q}_i^-$ and $\tilde{Q}_i^+$ can be calculated by

$$\begin{cases}\tilde{Q}_i^- = 2L_c - \tilde{Q}_i \\ \tilde{Q}_i^+ = 2U_c - \tilde{Q}_i\end{cases} \tag{23}$$

where U_c and L_c are the upper and lower bounds of the MC simulation, respectively, and $\tilde{Q}_i$ represents the ith result of RUL prediction. In the next step, a series of capacity estimation are conducted, and the detailed discussions are performed.

IV. Results and Discussion of Capacity Estimation

To obtain the capacity degradation data in the whole lifespan of lithium-ion batteries, the RFR is firstly employed to estimate the battery capacity using only one cell data for model training, and then the built model will be validated in other cells.

A. Capacity Estimation Based on RFR

In this study, we employed the experimental data of cell 1 as the training data and other cells' data for test. Figs. 7 and 8 show the estimation results and corresponding errors respectively. As can be seen from Fig. 7, the estimated results of cells 2 to 7 all track the degradation trajectory of real capacity variation. In the stage where the capacity degrades exponentially, the estimated capacity also approximates the actual value. As can be obviously seen from Figs. 8(a) to (g), the maximum estimation error of all batteries is less than 2%. By comparing with the results listed in [41], of which the estimation error by conventional SOH and RUL prediction methods is more than 2% in most cases, we can conclude that the proposed RFR can estimate the capacity with higher accuracy. Next, the estimation performance of RFR is further evaluated by different criteria.

B. Error Analysis of Capacity Estimation

To quantitatively evaluate the estimation performance of RFR model, the maximum absolute error (ME), root-mean-square error (RMSE) and goodness-of-fit are considered as the criteria. Among them, ME and RMSE comprehensively represent the average estimation performance, and smaller value implies better estimation precision. A goodness-to-fit parameter R^2, varying within [0, 1], is a measure of how the predicted value derived by the model tracks the referred value. The higher value (closer to 1) of R^2 indicates more similar prediction result, compared with the real value. These three criterions are defined as follows:

$$\begin{cases}ME = \max|y_i - \hat{y}_i| \\ RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \\ R^2 = 1 - \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 / \sum_{i=1}^{n}(y_i-\bar{y}_i)^2\end{cases} \tag{24}$$

where n represents the total sample number; y and $\hat{y}$ are the real value and estimated value of target variable, respectively; and $\bar{y}$ represents the average value of response variables.

The detailed results for cells 2 to 7 are show in Table V, from which we can find that the maximum and minimum value of ME are 1.92% and 0.89%, respectively, and the RMSE of all cells is less than 1%. It can therefore be indicated that the RFR model can estimate the battery capacity with high accuracy. Since we employ only the data of cell 1 for model training, the estimation results indicate that the RFR model shows strong robustness and

TABLE V
THE CAPACITY PREDICTION ERRORS OF CELLS 2 TO 7

Battery Number	Error Criterion		
	ME (%)	*RMSE* (%)	R^2
Cell 2	1.36	0.57	0.9844
Cell 3	1.33	0.44	0.9836
Cell 4	0.89	0.30	0.9941
Cell 5	1.43	0.82	0.9692
Cell 6	1.92	0.92	0.9323
Cell 7	1.56	0.84	0.9542

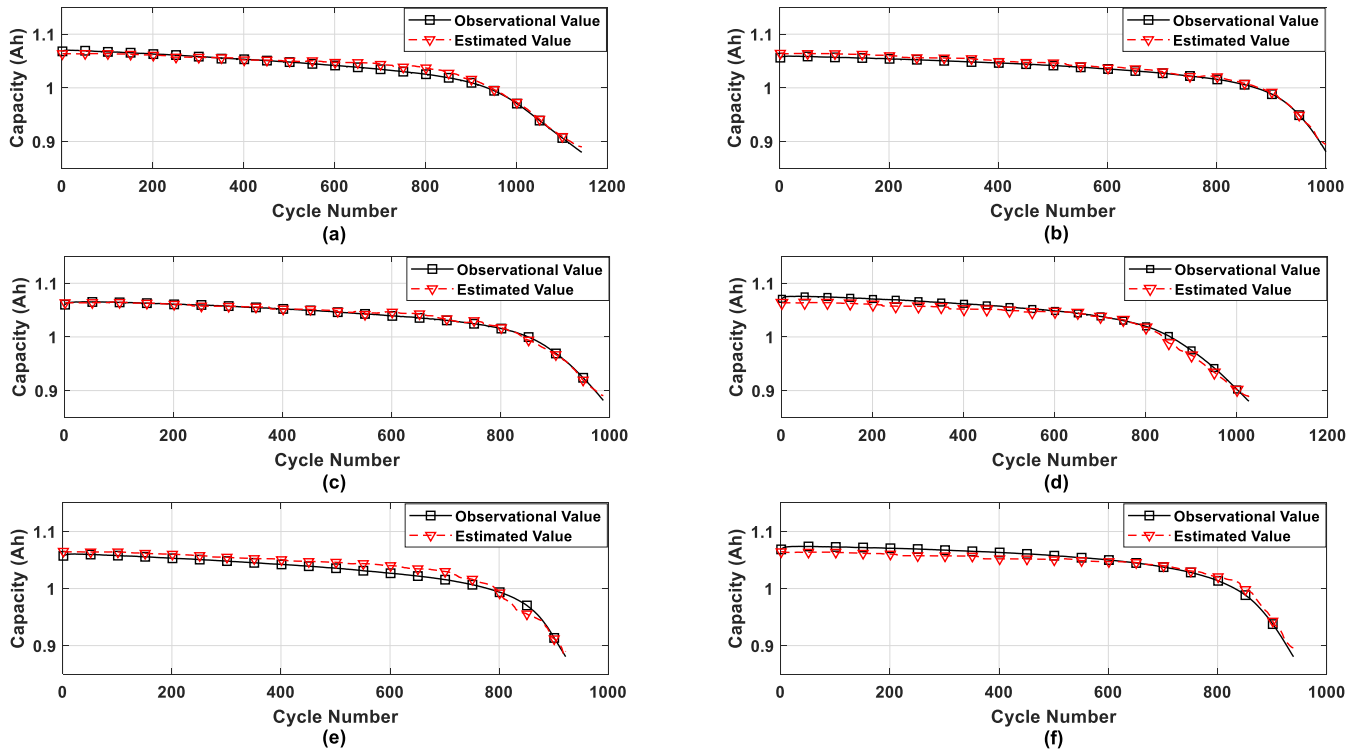

Fig. 7. The capacity estimation results with data of cell 1 for training. (a)–(f) capacity estimation results for cells 2 to 7.

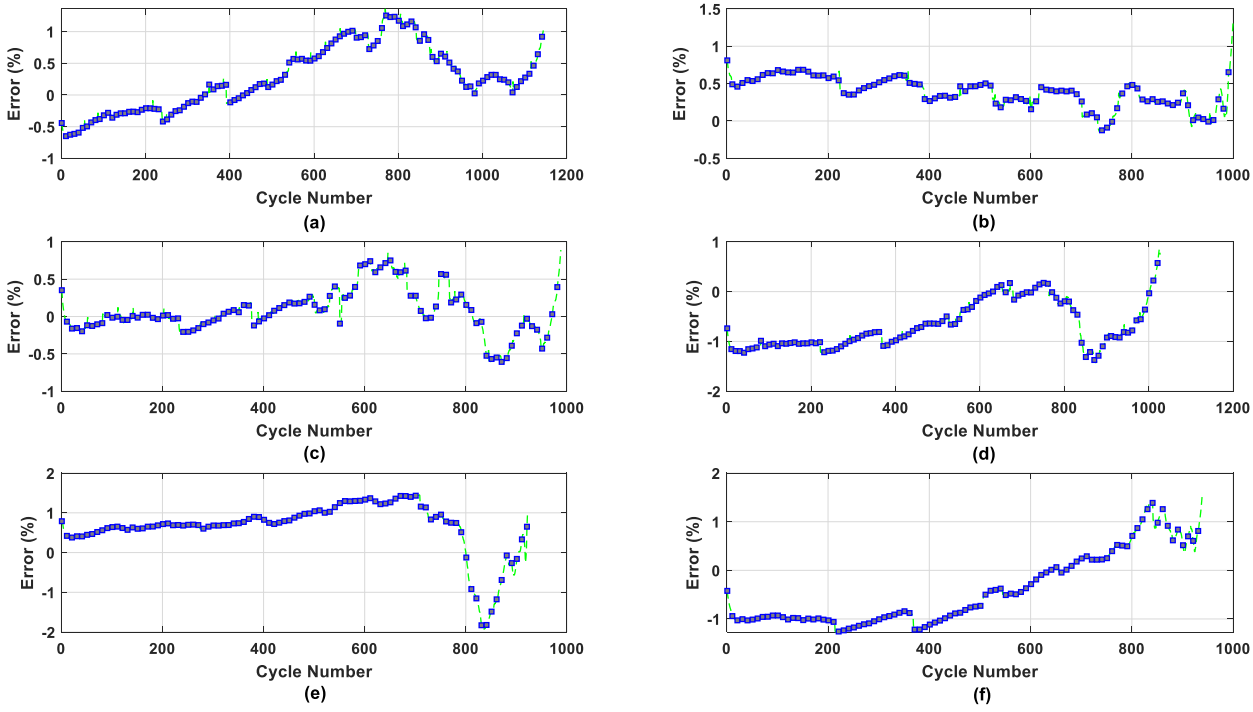

Fig. 8. The corresponding capacity estimation errors of cells 2 to 7.

can effectively capture the degradation mechanism for the same type battery. The R^2 for cells 2 to 4 are greater than 0.98, close to 1, indicating that the estimated values are similar to real values. In addition, the R^2 of cells 5 to 7 is 0.9692, 0.9323 and 0.9542, respectively, which is smaller than that of cells 2 to 4. It can be noted that the ME and RMSE of cell 6 are larger than the other cells' estimation error and its R^2 is the least. It can be found from Fig. 3(a) that the number of cycle life for cell 6 is shortest, which will increase the global estimation error of the model to a certain extent. As mentioned before, ME denotes the maximum difference between the estimated values and the observed values, and RMSE is utilized to evaluate the average difference between

TABLE VI
λ REFERENCE VALUES OF BCT FOR DIFFERENT CELLS

Cell No.	Cell 2	Cell 3	Cell 4
λ	25.9614	24.6964	27.3004
Cell No.	Cell 5	Cell 6	Cell 7
λ	25.7438	26.8264	35.2493

the estimated values and the observed values. R^2 measures how closely the estimated values match the observed values. R^2 equaling 1 indicates that the model can explain all the variability of the objective category. From the perspective of maximum error, average error and the similarity between the estimated values and observed values, the RFR algorithm leads to accurate estimation of battery capacity in the whole lifespan. To sum up, the built RFR model that relies only on the training data of one cell can well track the capacity degradation trend with the acceptable error for other cells with the same type. The overall estimation error is less than 2%, and all the results are within a reasonable range. By this manner, the reliability and robustness of the proposed algorithm is proved.

V. RESULTS AND DISCUSSION OF RUL PREDICTION

To evaluate its effectiveness and performance, the developed method is applied for predicting the battery RUL based on the estimated capacity data. First, the results of BCT and RRA fitting are analyzed. Then, the developed method is performed for RUL prediction.

A. Results of BCT and RRA Fitting

It can be extrapolated from (1) that the battery RUL has a linear relationship with the cycle number, whereas the capacity degradation is nonlinear, as shown in Fig. 3(a). To employ the developed method to predict the battery RUL, the estimated capacity data is firstly transformed. The BCT is applied to construct a linear relationship between the transformed capacities and cycles. Therefore, there is only one independent variable that represents the cycle number and thus $m = 1$ in (8). Now, equation (8) can be rewritten as

$$\begin{cases} C_i(\lambda) = \beta_0 + k_i\beta_1 + \varepsilon_i \\ \varepsilon_i = N(0, \sigma^2), i = 1, 2, \ldots, n \end{cases} \tag{25}$$

where C represents the transformed capacity using BCT; k denotes the cycle number; β_0 and β_1 are the coefficients of linear model, and ε_i is the random error. The values of λ in (7) for Cells 2–7 are calculated according to (9)–(15), which are listed in Table VI. For the linearized capacity values, the RRA is utilized to identify the linear model parameters.

The parameters of linear model expressed in (25) are identified though (16) to (19), and the identification results are shown in Table VII. Theoretically, there is a similar linear relationship of the transformed capacity with the cycle for the same type of battery, due to the similar degradation mechanism. It can be seen from Table VII that the linear model parameter β_0 of all batteries is 0.2751, 0.2696, 0.2828, 0.2803, 0.2792 and 0.2783, and β_1 is -2.692×10^{-4}, -3.065×10^{-4}, -3.155×10^{-4}, -3.057×10^{-4}, -3.368×10^{-4} and -2.956×10^{-4}. It can be found that the parameters of linear models for all the cells remain close and comply with the hypothesis that the same type battery exhibits similar linear relationships between the transformed capacities and cycles. The calculation results also indicate that the linear model established between the transformed capacities and the cycles using BCT is reliable. To evaluate the effectiveness of RRA, the evaluation criterions ρ and MAE are obtained via (20) and (21) and the corresponding results are show in Fig. 9. As can be seen, the correlation coefficient ρ of all the linear models is 0.9877, 0.9975, 0.9900, 0.9852, 0.9949 and 0.9757, and all the MAE is less than 0.015, manifesting that a strong linear relationship exists between the transformed capacities and cycles. Moreover, the fitted values are quite close to the transformed capacities, manifesting the feasibility of the linear fitting of RRA. The fitted effectiveness of RRA also indicates that there exists a strong linear relationship between the transformed capacities and cycle number. Next, the constructed linear model is extrapolated to predict the battery RUL.

TABLE VII
THE LINEAR MODEL PARAMETERS AND RIDGE REGRESSION COEFFICIENTS OF DIFFERENT CELLS

Cell No.	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7
β_0	0.2751	0.2696	0.2828	0.2803	0.2792	0.2783
β_1 ($\times 10^{-4}$)	-2.692	-3.065	-3.155	-3.057	-3.368	-2.956
k	-0.089	-0.088	-0.090	-0.091	-0.089	-0.080

B. RUL Prediction Within Whole Cycle Life

In this study, the battery RUL is predicted by extrapolating the constructed linear model. To evaluate the predicted performance of developed method within the entire lifespan, the whole estimated capacity of cells 2 to 4 is utilized to realize the RUL prediction. Fig. 10 shows the RUL prediction results and errors for cells 2 to 4. It can be seen form Fig. 10(a), (c) and (e) that there exists a linear relationship between the real RUL and cycle number. The prediction results show similar trend with real RUL curve but with mild partial oscillation, indicating the prediction error is relatively large. The RUL predicted values are obtained by extrapolating the linear model between the transformed capacities and cycles, and the prediction performance of the developed method is mainly dependent on the constructed linear model. The RUL prediction results of cells 2 to 4, as shown in Fig. 10(a), (c) and (e), verifying the feasibility of the constructed linear model based on BCT. Fig. 10(b), (d) and (f) show the prediction error at each cycle. Except individual cycles where the prediction error is great than 100 cycles, the prediction errors of cells 2 and 4 are mostly less than 100. Additionally, the prediction error of cell 3 is smaller, and the prediction error is less than 50 cycles overall. As can be seen from Fig. 9, the correlation coefficient ρ of cell 3 reaches 0.9975, which is the highest value among those of cells 2 to 7. Moreover, the RUL prediction error of cell 3 is smaller than that of other cells. Thus, the RUL prediction results justify the previous hypothesis that the RUL prediction accuracy is mainly dependent on the linear

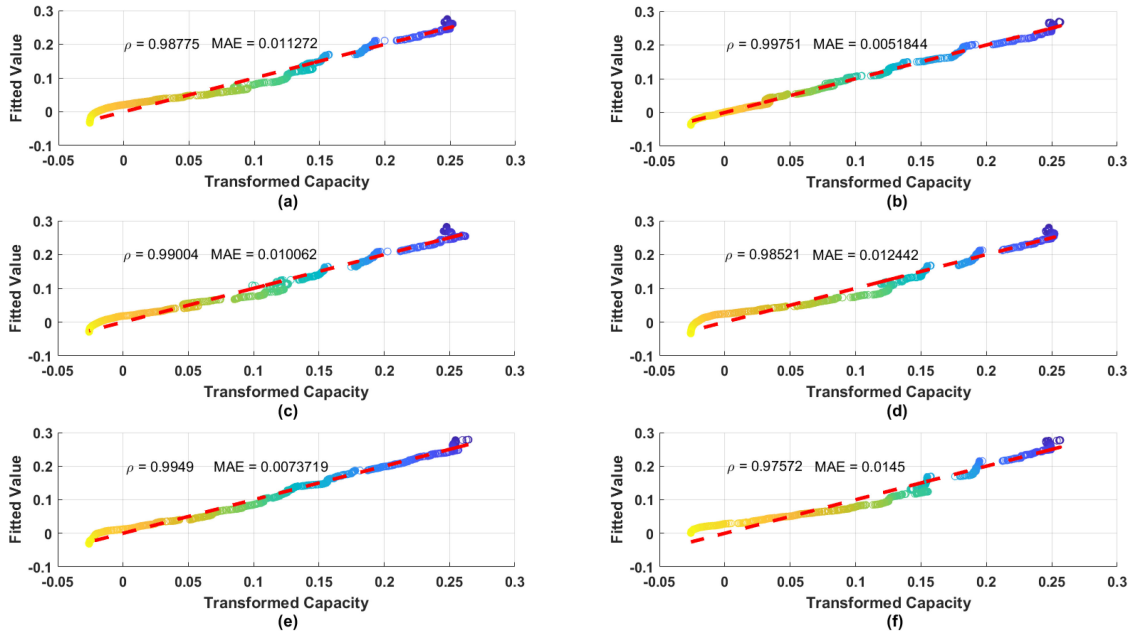

Fig. 9. Illustration of the linear regression model between transformed capacities and fitted values for cells 2 to 7.

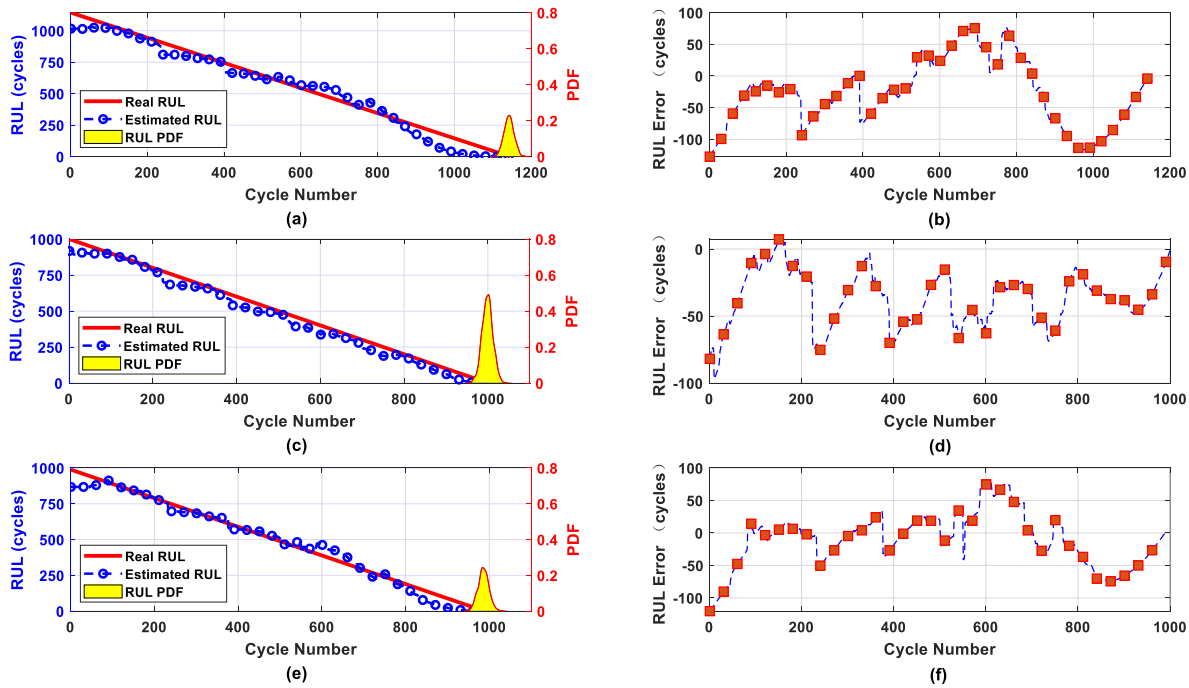

Fig. 10. The RUL prediction results and errors in the lifespan. (a)-(b) The results and errors of RUL prediction for cell 2; (c)-(d) The results and errors of RUL prediction for cell 3; (e)-(f) The results and errors of RUL prediction for Cell 4.

relationship between the transformed capacities and cycles. To sum up, the prediction error is within a reasonable range, validating that the proposed method can accurately predict the battery RUL within the whole cycle life based on the estimated capacity.

The linear model is extrapolated to make multiple-step ahead predictions, and when the predicted is lower than the threshold, an EOL is reported. The predicted EOL and true EOL are shown in Table VIII. The prediction error of EOL for cells 2 to 4 are respectively 26, 36 and 9 cycles, highlighting that the developed method can accurately predict the EOL of battery. There is one EOL prediction for each MC simulation, and the simulation is repeated 1000 times in this study. The prediction PDF and 95% confidence interval are obtained based on the MC simulation. It can be seen from Fig. 10(a), (c) and (e) that the PDF distribution of cells 2 to 4 is relatively concentrated, and the 95% confidence interval are [1100, 1137], [956, 972] and [961, 997], respectively. We can find that the interval spans are 37,

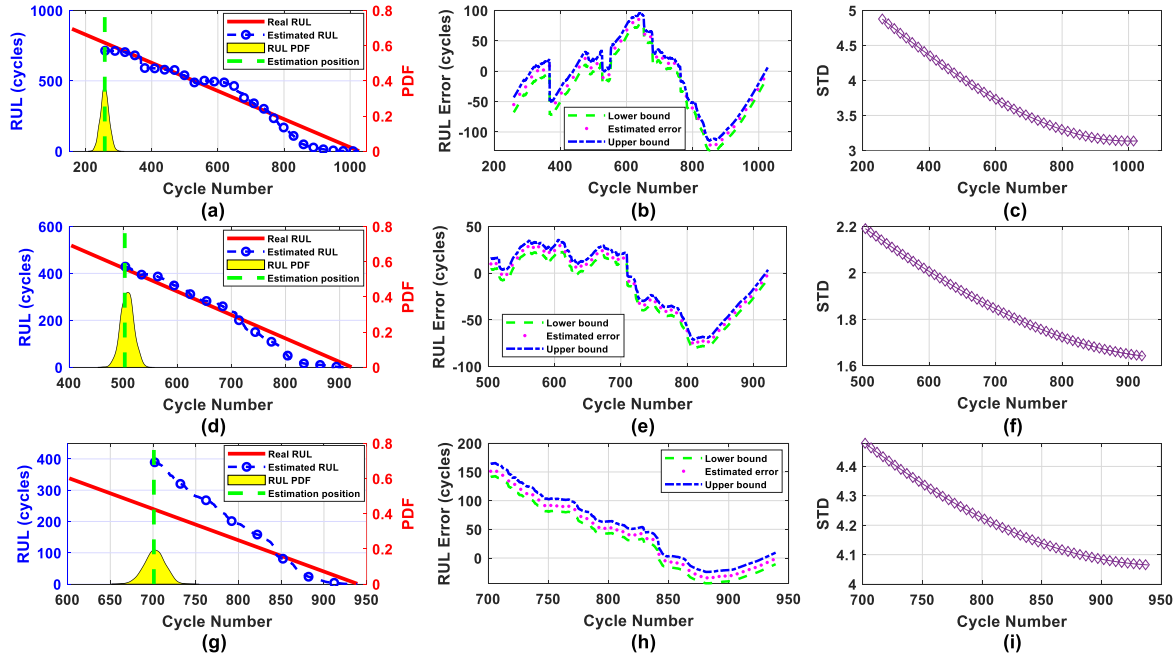

Fig. 11. The RUL prediction results and errors at different start position. (a)–(c) the results and errors of RUL prediction for cell 5; (d)–(f) the results and errors of RUL prediction for cell 6; (g)–(i) the results and errors of RUL prediction for cell 7.

TABLE VIII
THE EOL PREDICTION RESULTS AND ERRORS OF CELLS 2 TO 4

Battery Number	Evaluation Indicator			
	Real EOL	Estimated EOL	Error	95% confidence interval
Cell 2	1144	1118	26	[1100,1137]
Cell 3	1000	964	36	[956,972]
Cell 4	988	979	9	[961,997]

16 and 36 cycles, showing that the proposed prediction method has high credibility. To sum up, the prediction results justify that the proposed method can predict the battery RUL and EOL with preferable accuracy and high reliability.

C. RUL Prediction at Different Cycle Life

To analyze the robustness and stability of proposed algorithm, the developed method is applied for predicting the battery RUL at different starting position based on the estimated capacity data of cells 5 to 7. Note that the larger the cycle number corresponding to the prediction starting position, the less capacity data used to model for the RUL prediction will be, and therefore the higher uncertainty of RUL prediction will emerge. Fig. 11 shows the RUL prediction results and errors of cells 5 to 7. The starting positions of prediction are 30%, 60% and 80% of the whole cycle life, corresponding to 259, 504 and 702 cycles, respectively; and the capacity data employed to construct the linear model for cells 5 to 7 are respectively 70%, 40% and 20% of the whole data. As can be seen from Fig. 11(a) and (d), the prediction results of cells 5 and 6 can better track the actual RUL trajectories, whereas the predicted RUL of cell 7 deviates from the real RUL curve. The prediction results can also reflect the viability of linear relationship between the transformed capacities and the cycles, and the RUL prediction accuracy relies on this relationship. It can be seen from Fig. 9 that the fitted correlation coefficient ρ of linear model for cells 5 to 7 is 0.9852, 0.9949 and 0.9757, respectively. The ρ of cell 7 is the smallest and that of cell 6 is the largest, indicating that the linear relationship of transformed capacities via cycle number for cell 6 is the strongest and that to cell 7 is the weakest. The RUL prediction results are consistent with the results of linear model fitting. As can be seen form Fig. 11(b), (e) and (h) that the prediction error is mostly less than 50 for cell 6, except few individual points with a slight oscillation. The prediction error for cell 7 is the largest with the maximum error of 151 cycles; however, as the cycle number increases, the prediction error gradually decreases, and finally the prediction error of EOL is only 16.

The blue and green dot-dashed lines in Fig. 11(b), (e) and (h) are the upper and lower bounds of the 95% confidence interval of the prediction error calculated based on the MC simulation. In each case, the 95% confidence boundary is furnished based on all RUL prediction errors at the specified cycles. As can be seen that the RUL prediction errors are within $[-50, 150]$, and the 95% confidence bounds are within 40 cycles. This error margin of RUL predictions indicates a high prediction stability of the developed method. Fig. 11(c), (f) and (i) show the standard deviation (STD) of the RUL prediction at the specified cycles. The STD is obtained by the MC simulation. Since the STDs are all within 5 cycles, which shows that the RUL prediction of all cells is precise. The STD is a monotonically decreasing function, suggesting a more precise RUL prediction as the cycle number increases.

To further evaluate the prediction performance of the developed method, the prediction MAE is calculated by (21) for cells 5 to 7, as shown in Table IX. The prediction MAE of cells 5 to 7 are respectively 44.48, 29.73 and 56.17 cycles. The prediction MAE of cell 7 is the largest, which is in line with the above results in which the linear relationship of transformed capacities

TABLE IX
THE EOL PREDICTION RESULTS AND ERRORS OF CELLS 5 TO 7

Battery Number	Evaluation Indicator				
	Real EOL	Starting cycle	Estimated EOL	MAE	Mean STD
Cell 5	1027	259	1002	44.48	3.98
Cell 6	922	504	906	29.73	1.88
Cell 7	939	702	923	56.17	3.95

with cycle number for cell 7 is relatively weak. Based on the above discussion, we can conclude that the developed method can predict the battery EOL from different starting cycle position with high accuracy, and the maximum error and MAE of RUL prediction are 151 and 56.17 cycles. To sum up, the prediction results validate that the proposed method can predict the battery RUL with high accuracy, stability and strong robustness.

VI. CONCLUSION

In this paper, a RUL prediction method based on capacity estimation and BCT is proposed for lithium-ion batteries. In the developed method, the internal resistance, the average temperature of each specified cycle and the absolute value of discharge incremental capacity peak are considered as the aging features. The RFR with the aging features as model input and the corresponding capacity as output is then employed to estimate the battery capacity. The BCT is exploited to transform the estimated capacity data and to construct a linear model of transformed capacities via cycles. In addition, the RRA is employed to identify the linear model parameters. The battery RUL is predicted based on the extrapolation of the linear model, and the prediction uncertainties are generated using the MC simulation. To evaluate the prediction performance of the proposed method, the aging experimental data involving 7 cells are employed to test and validate the algorithm. The capacity estimation results validate that when only one battery data is used for training, the capacity estimation error of other cells is less than 2%. In the whole cycle life, the experimental results show that the RUL prediction maximum error is 127 cycles, and the prediction error of EOL can reach a maximum value of 36 cycles. The maximum spans of 95% confidence interval is 37 cycles, indicating that the proposed prediction method shows high accuracy and credibility. Moreover, the developed method can also be performed for RUL prediction at different starting cycle position. The prediction results show that the RUL prediction errors are restricted with [-50, 150], and 95% confidence boundary is kept within 40 cycles. The experimental results illustrate that the proposed method can predict the battery RUL with preferable accuracy and certain robustness. It can also be indicated that the proposed method shows certain potential for real applications.

The RUL prediction is conducted based on single cell in this research. In our next step research, the RUL prediction of battery packs will be conducted, and more precise RUL estimation will be investigated with the consideration of different operating temperatures and real-time operation data.

REFERENCES

[1] Z. Chen, M. Sun, X. Shu, R. Xiao, and J. Shen, "Online state of health estimation for lithium-ion batteries based on support vector machine," *Appl. Sci.*, vol. 8, no. 6, pp. 1–13, 2018.

[2] X. Shu, G. Li, J. Shen, W. Yan, Z. Chen, and Y. Liu, "An adaptive fusion estimation algorithm for state of charge of lithium-ion batteries considering wide operating temperature and degradation," *J. Power Sources*, vol. 462, pp. 1–13, 2020.

[3] X. Hu, S. E. Li, and Y. Yang, "Advanced machine learning approach for lithium-ion battery state estimation in electric vehicles," *IEEE Trans. Transp. Electrific.*, vol. 2, no. 2, pp. 140–149, Jun. 2016.

[4] X. Hu, H. Yuan, C. Zou, Z. Li, and L. Zhang, "Co-estimation of state of charge and state of health for lithium-ion batteries based on fractional-order calculus," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10319–10329, Nov. 2018.

[5] Z. Chen, Q. Xue, R. Xiao, Y. Liu, and J. Shen, "State of health estimation for lithium-ion batteries based on fusion of autoregressive moving average model and Elman neural network," *IEEE Access*, vol. 7, pp. 102662–102678, 2019.

[6] L. F. Wu, X. H. Fu, and Y. Guan, "Review of the remaining useful life prognostics of vehicle lithium-ion batteries using data-driven methodologies," *Appl. Sci.*, vol. 6, no. 6, 2016, Art. no. 11.

[7] C. Lyu, Q. Z. Lai, T. F. Ge, H. H. Yu, L. X. Wang, and N. Ma, "A lead-acid battery's remaining useful life prediction by using electrochemical model in the particle filtering, framework," *Energy*, vol. 120, pp. 975–984, 2017.

[8] S. S. Y. Ng, Y. J. Xing, and K. L. Tsui, "A naive Bayes model for robust remaining useful life prediction of lithium-ion battery," *Appl. Energy*, vol. 118, pp. 114–123, 2014.

[9] B. Saha, K. Goebel, S. Poll, and J. Christophersen, "Prognostics methods for battery health monitoring using a Bayesian framework," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 2, pp. 291–296, Feb. 2009.

[10] D. Wang, F. F. Yang, K. L. Tsui, Q. Zhou, and S. J. Bae, "Remaining useful life prediction of lithium-ion batteries based on spherical cubature particle filter," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 6, pp. 1282–1291, Jun. 2016.

[11] D. Wang, Q. Miao, and M. Pecht, "Prognostics of lithium-ion batteries based on relevance vectors and a conditional three-parameter capacity degradation model," *J. Power Sources*, vol. 239, pp. 253–264, 2013.

[12] F. F. Yang, X. B. Song, G. Z. Dong, and K. L. Tsui, "A coulombic efficiency-based model for prognostics and health estimation of lithium-ion batteries," *Energy*, vol. 171, pp. 1173–1182, 2019.

[13] Y. Z. Zhang, R. Xiong, H. W. He, and M. G. Pecht, "Lithium-ion battery remaining useful life prediction with box-cox transformation and Monte Carlo simulation," *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1585–1597, Feb. 2019.

[14] Y. Li *et al.*, "Random forest regression for online capacity estimation of lithium-ion batteries," *Appl. Energy*, vol. 232, pp. 197–210, 2018.

[15] X. Y. Li, L. Zhang, Z. P. Wang, and P. Dong, "Remaining useful life prediction for lithium-ion batteries based on a hybrid model combining the long short-term memory and Elman neural networks," *J. Energy Storage*, vol. 21, pp. 510–518, 2019.

[16] X. Y. Li, X. Shu, J. W. Shen, R. X. Xiao, W. S. Yan, and Z. Chen, "An on-board remaining useful life estimation algorithm for lithium-ion batteries of electric vehicles," *Energies*, vol. 10, no. 5, 2017, Art. no. 15.

[17] J. Liu and Z. Q. Chen, "Remaining useful life prediction of lithium-ion batteries based on health indicator and Gaussian process regression model," *IEEE Access*, vol. 7, pp. 39474–39484, 2019.

[18] M. A. Patil *et al.*, "A novel multistage support vector machine based approach for li ion battery remaining useful life estimation," *Appl. Energy*, vol. 159, pp. 285–297, 2015.

[19] P. Y. Guo, Z. Cheng, and L. Yang, "A data-driven remaining capacity estimation approach for lithium-ion batteries based on charging health feature extraction," *J. Power Sources*, vol. 412, pp. 442–450, 2019.

[20] D. Zhou, H. T. Yin, W. Xie, P. Fu, and W. B. Lu, "Research on online capacity estimation of power battery based on EKF-GPR model," *J. Chem.*, vol. 2019, 2019, Art. no. 9.

[21] X. Y. Li, Z. P. Wang, and J. Y. Yan, "Prognostic health condition for lithium battery using the partial incremental capacity and Gaussian process regression," *J. Power Sources*, vol. 421, pp. 56–67, 2019.

[22] Y. Z. Zhang, R. Xiong, H. W. He, and M. G. Pecht, "Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 5695–5705, Jul. 2018.

[23] Y. P. Zhou and M. H. Huang, "Lithium-ion batteries remaining useful life prediction based on a mixture of empirical mode decomposition and ARIMA model," *Microelectronics Rel.*, vol. 65, pp. 265–273, 2016.
[24] L. Ren, L. Zhao, S. Hong, S. Q. Zhao, H. Wang, and L. Zhang, "Remaining useful life prediction for lithium-ion battery: A deep learning approach," *IEEE Access*, vol. 6, pp. 50587–50598, 2018.
[25] G. J. Ma, Y. Zhang, C. Cheng, B. T. Zhou, P. C. Hu, and Y. Yuan, "Remaining useful life prediction of lithium-ion batteries based on false nearest neighbors and a hybrid neural network," *Appl. Energy*, vol. 253, pp. 11, 2019.
[26] B. Long, W. M. Xian, L. Jiang, and Z. Liu, "An improved autoregressive model by particle swarm optimization for prognostics of lithium-ion batteries," *Microelectron. Rel.*, vol. 53, no. 6, pp. 821–831, 2013.
[27] Y. P. Zhou, M. H. Huang, and M. Pecht, "Remaining useful life estimation of lithium-ion cells based on k-nearest neighbor regression with differential evolution optimization," *J. Cleaner Prod.*, vol. 249, p. 12, Mar. 2020.
[28] W. He, N. Williard, M. Osterman, and M. Pecht, "Prognostics of lithium-ion batteries based on Dempster-Shafer theory and the Bayesian Monte Carlo method," *J. Power Sources*, vol. 196, no. 23, pp. 10314–10321, 2011.
[29] Y. Xing, E. W. M. Ma, K. L. Tsui, and M. Pecht, "An ensemble model for predicting the remaining useful performance of lithium-ion batteries," *Microelectron. Rel.*, vol. 53, no. 6, pp. 811–820, 2013.
[30] X. Han *et al.*, "A review on the key issues of the lithium ion battery degradation among the whole life cycle," *eTransportation*, vol. 1, pp. 1–21, 2019.
[31] K. A. Severson *et al.*, "Data-driven prediction of battery cycle life before capacity degradation," *Nat. Energy*, vol. 4, no. 5, pp. 383–391, 2019.
[32] Y. Wu, Q. Xue, J. Shen, Z. Lei, Z. Chen, and Y. Liu, "State of health estimation for lithium-ion batteries based on healthy features and long short-term memory," *IEEE Access*, vol. 8, pp. 28533–28547, 2020.
[33] 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.
[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[35] B. Leo, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.
[36] L. L. Kupper, "Extending the box–cox transformation to the linear mixed model," *J. Roy. Stat. Soc.*, vol. 169, no. 2, pp. 273–288, 2005.
[37] D. R. Cox, "An analysis of transformations," *J. Roy. Stat. Soc.*, vol. 78, no. 2, 1964.
[38] P. D. Allison, S. Horizons, and P. Haverford, "Handling missing data by maximum likelihood," in *SAS Global Forum 2012*, Statistical Horizons: Harrisburg, PA, USA, pp. 312–2012, 2012.
[39] G. C. McDonald, "Ridge regression," *Wiley Interdisciplinary Reviews Computational Statistics*, vol. 1, no. 1, pp. 93–100, Jul. 2009.
[40] C. Borges and J. Dias, "A model to represent correlated time series in reliability evaluation by non-sequential Monte Carlo simulation," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1511–1519, Mar. 2016.
[41] M. S. H. Lipu *et al.*, "A review of state of health and remaining useful life estimation methods for lithium-ion battery in electric vehicles: Challenges and recommendations," *J. Cleaner Prod.*, vol. 205, pp. 115–133, Dec. 2018.

Qiao Xue received the B.S. degree in traffic engineering in 2018 from the Kunming University of Science and Technology, Kunming, China, where he is currently working toward the M.S. degree in transportation engineering. His research interests include state of health monitoring for battery and battery management system.

Shiquan Shen received the Ph.D. degree in power machinery and engineering from Tianjin University, Tianjin, China, in 2020. He was a joint Ph.D. Student with Argonne National Laboratory, Lemont, IL, USA, from 2018 to 2019. He is currently a Lecturer with the Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming, China. His research interests include energy management of hybrid electric vehicles, model predictive control, multi-phase flow, etc.

Guang Li (Member, IEEE) received the Ph.D. degree in electrical and electronics engineering, specialized in control systems, from the University of Manchester, Manchester, U.K., in 2007. He is currently a Senior Lecturer in dynamics modeling and control with the Queen Mary University of London, London, U.K. His research interests include constrained optimal control, model predictive control, adaptive robust control, and control applications including renewable energies, energy storage, etc.

Yuanjian Zhang (Member, IEEE) received the M.S. degree in automotive engineering from the Coventry University, Coventry, U.K., in 2013, and the Ph.D. degree in automotive engineering from Jilin University, Changchun, China, in 2018. In 2018, he joined the University of Surrey, Guildford, U.K., as a Research Fellow in advanced vehicle control. He currently works with Sir William Wright Technology Centre, Queen's University Belfast, Belfast, U.K. His current research interests include advanced control on electric vehicle powertrains, vehicle-environment-driver cooperative control, vehicle dynamic control, and intelligent control for driving assist system.

Zheng Chen (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering and the Ph.D. degree in control science engineering from Northwestern Polytechnical University, Xi'an, China, in 2004, 2007, and 2012, respectively. He was a Post-Doctoral Fellow and a Research Scholar with the University of Michigan, Dearborn, MI, USA, from 2008 to 2014. He is currently a Professor with the Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming, China, and also a Marie-Curie Research Fellow with the School of Engineering and Materials Science, Queen Mary University of London, London, U.K. He has conducted more than 30 projects and has authored/coauthored more than 80 peer-reviewed journal papers and conference proceedings. His research interests include battery management system, battery status estimation, and energy management of hybrid electric vehicles. He is a Fellow of the Institution of Engineering and Technology. In addition, he was a recipient of the Yunnan Oversea High Talent Project, China, and the second place of IEEE VTS Motor Vehicles Challenge in 2017 and 2018.

Yonggang Liu (Senior Member, IEEE) was born in Chongqing, China, in 1982. He received the B.S. and Ph.D. degrees in automotive engineering from Chongqing University, Chongqing, China, in 2004 and 2010, respectively. He was a joint Ph.D. candidate with the University of Michigan-Dearborn, Dearborn, MI, USA, from 2007 to 2009. He is currently a Professor and a Doctoral Supervisor, the Dean Assistant with the School of Automotive Engineering, Chongqing University. He has led more than 20 research projects, such as National Natural Science Foundation of China (both Youth Fund and General Program), Ph.D. Programs Foundation of Ministry of Education of China, and China Postdoctoral Science Foundation. More than 70 research papers have been published and ten patents have been awarded. His research interests mainly include optimization and control of intelligent electric vehicles (EV/HEV) power system, and integrated control of vehicle automatic transmissions. He was the Head of the Secretariat in the International Conference on Power Transmissions in 2016 and the Session Chairman of the International Symposium on Electric Vehicles in 2017, etc. He is also a Committeeman of Technical Committee on Vehicle Control and Intelligence of Chinese Association of Automation.