Journal of Power Sources 476 (2020) 228654

Contents lists available at ScienceDirect

Journal of Power Sources

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

Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model

Fangfang Yang [a], Dong Wang [b], Fan Xu [a,*], Zhelin Huang [a], Kwok-Leung Tsui [a]

[a] School of Data Science, City University of Hong Kong, Hong Kong
[b] The State Key Laboratory of Mechanical Systems and Vibration, Shanghai Jiao Tong University, Shanghai, 200240, PR China

HIGHLIGHTS

- Various features are extracted and explored for effective battery lifespan prediction.
- A GBRT model is proposed to directly predict battery lifespan from extracted features.
- The impact of model hyper-parameters on lifespan prediction is investigated.
- The relative importance of input features is analyzed and discussed.
- Its performance is compared with other state-of-art machine learning methods.

ARTICLE INFO

Keywords:
Lifespan prediction
Lithium-ion batteries
Machine learning
Gradient boosting regression tree
Feature extraction

ABSTRACT

Accurate battery lifespan prediction is critical for the quality evaluation and long-term planning of battery management systems. As battery degradation process is typically nonlinear, accurate early prediction of cycle life with significantly less degradation is extremely challenging. Approaches using machine learning techniques, which are mechanism-agnostic alternatives, to predict battery lifespan are therefore becoming more and more attractive. In this paper, a gradient boosting regression tree (GBRT) is proposed to model complex nonlinear battery dynamics and predict battery lifespan through various extracted battery features. Essentially, the GBRT works by constructing additional trees through minimizing the prediction residues from existing base models. It can identify complex and nonlinear end-to-end relationship and meanwhile provide relative importance for each input feature. Various potential features including voltage-related features, capacity-related features, and temperature-related features are constructed in both discharge time dimension and life cycle dimension and explored for effective lifespan prediction. Key hyper-parameters including learning rate, number of trees, and maximum number of splits are investigated for optimal GBRT prediction. Comparative studies confirm that the proposed method is significantly superior to other machine learning algorithms for battery lifespan prediction with limited data samples and various input features, with mean average percentage error around 7%.

1. Introduction

Featured by high energy density and high power density, lithium-ion batteries have been recognized as one of the most attractive energy storage devices and they are been broadly used in lots of scenarios, ranging from portable electronics to electric vehicles [1,2]. Despite the high safety standards used in manufacturing of lithium-ion batteries, there have been numerous cases of lithium-ion battery fires and explosions, resulting in property losses and bodily injuries. Therefore, accurate lifespan prediction of lithium-ion batteries is important to help assess battery quality in advance, improve the long-term battery planning and subsequently guarantee the safety and reliability of battery operations.

Existing studies for battery lifespan prediction mainly fall into three categories: mechanism methods, model-based methods, and machine learning methods. The mechanism methods focus on battery aging mechanisms and describe the battery chemical and physical reactions with arrays of equations. Dai et al. [3] developed a capacity fade model

* Corresponding author.
E-mail addresses: fangfyang2-c@my.cityu.edu.hk (F. Yang), dongwang4-c@sjtu.edu.cn (D. Wang), fanxu8@cityu.edu.hk (F. Xu), zhuang44@cityu.edu.hk (Z. Huang), kltsui@cityu.edu.hk (K.-L. Tsui).

https://doi.org/10.1016/j.jpowsour.2020.228654
Received 12 April 2020; Received in revised form 26 May 2020; Accepted 13 July 2020
Available online 17 August 2020

to depict the manganese dissolution of lithium-ion batteries. Ning et al. [4] focused on solvent reduction reaction and built a capacity fade model to describe the loss of active lithium ions. The battery lifespan is then evaluated by extrapolating underlying battery capacity model. While these chemistry-based or mechanism-specific models have shown to be effective, developing models that describe full cells cycled under relevant operating conditions remains challenging due to the complex inherent coupling of thermal and chemical heterogeneities within a cell. Additionally, the mechanism methods often require intensive computations, hence inappropriate for real-time scenarios.

In contrast, the model-based methods establish a mathematical model to capture battery degradation dynamics with minor consideration of physicochemical principles and then incorporates the model with advanced filtering techniques to perform online lifespan prediction, including Kalman filter [5,6], particle filter [7–9], etc. A number of mathematical models have been proposed for battery remaining useful life prediction, such as empirical models (a quadratic polynomial model [10], a two-term exponential model [11], a two-term logarithmic model [12], etc.) and semi-empirical models (a coulombic efficiency-based model [7], a square-root-of-time model [13], etc.). These models often presume simple mathematical relationships between battery capacity and aging cycle. The prediction performance relies largely on the accuracy of underlying battery models and are often subject to cell-to-cell variations. Moreover, these methods often yield poor prediction performance at early stages of battery life as there exists multi-stage capacity degradation in lithium-ion batteries [12].

The machine learning methods, which require no explicit battery models, regard the battery system as a black box and infer battery state-of-health (SOH, usually use capacity as a key indicator) or lifespan directly from extracted features. According to different feature types, the existing methods can be further classified into the following groups.

1) Features from directly measurable variables, such as aging cycle, charging time, and open circuit voltage. For example, Rezvani et al. [14] estimated battery SOH directly from capacity aging cycle through an adaptive neural network.
2) Features from original voltage-capacity/time curves. Yang et al. [15] extracted four features including charge and voltage durations, slope, and vertical slope from constant current charging curves and employed a Gaussian process regression (GPR) model to infer battery SOH. Meng et al. [16] extracted features from voltage responses under short-term current pulse test and used a support vector machine (SVM) to infer battery SOH. He et al. [17] estimated SOH from the terminal voltage in charge process by a dynamic Bayesian network.
3) Features from processed voltage curves, such as incremental capacity (IC) curve and differential voltage (DV) curve. Weng et al. [18] mapped battery capacity from IC peak intensity via a SVM. Berecibar et al. [19] extracted and mapped a set of features from IC and DV curves to battery SOH by a SVM. Richardson et al. [20] inferred battery SOH from filtered IC and DV curves through a GPR model.
4) Features from statistical metrics, such as parameters of linear capacity fit [21]. Hu et al. [22] extracted the sample entropy of voltage sequence and mapped it to battery capacity through a sparse Bayesian predictive modeling method. Pan et al. [23] estimated battery SOH from estimated internal and polarized internal resistances using an extreme learning machine.

The machine learning methods provide great flexibility and model-free characteristics and are gaining more and more attentions in recent years due to the rapid development of computer hardware and computing power [24]. Besides above-mentioned methods, neural network [25] and fuzzy logic [26] have also been used for battery SOH estimation. Most machine learning-based work focus mainly on estimating battery SOH for the same cell [27]. The lifespan prediction can then be further performed by modeling and extrapolating the indicator degradation using the other two methods.

Most of these works were evaluated with small battery samples (usually less than 10) due to the long time (usually more than one year at room temperature) needed for cells to age. It is necessary to further compare their performance when more subjects are available, as the prediction performance of machine learning methods relies heavily on the quantity and quality of training datasets. What's more, most above-mentioned work extracted only one or few features for prediction, more effective features still need to be explored for the black-box model to learn the intrinsic battery dynamics. For instance, the temperature variation [28] and voltage evolution [21] within battery charge-discharge process presented to provide abundant information for battery prognostics and diagnostics, but are rarely investigated in literature. Last, the accurate performance of current estimation or prediction methods requires continuously collecting cycle data under strict operating condition so as to well update real-time battery parameters, which is complicated and practically unavailable.

To address the feature insufficiency problem, in this paper, we increase the diversity of features by investigating and exploring a lot of voltage-related features, capacity-related features, and temperature-related features for battery lifespan prediction. These selected features either appeared in or are inspired by various existing battery literatures [18,19,21], serving diversified research orientations. We further identify them from the cycling voltage and temperature curves in both discharge time dimension (short term) and life cycle dimension (long term). Life cycle dimension is considered in this work in the hope of gaining better insight about the degradation behavior, reducing the unwanted errors caused by noises and outliers, and improving the robustness of the prediction. Most of these features are constructed based on raw battery data collected from certain cycle points, making on-board data collection much more convenient than before.

With large numbers of battery features considered in this work, another major challenge to employing machine learning in battery lifespan prognostics is the relatively small number of degraded cell samples available for machine training. Generally popular neural network model requires very large training datasets to avoid overfitting, while simple machine learning algorithms such as logistic regression, elastic net, Naïve Bayes, and SVM become ineffective with large numbers of features [29]. For example, with elastic net used in Ref. [21], prediction error obviously increased when feature number increases from 6 ('discharge' model) to 9 ('full' model), for results of both primary and secondary tests. Within the machine learning field, tree-based ensemble models, which are flexible, easy training, and have the ability to handle unnecessary or correlated features without overfitting, show relatively best performance for this particular battery application [30]. A gradient boosting regression tree (GBRT) model [31], which constructs additional trees through minimizing the prediction residues from existing base models, is thus proposed in our study to directly predict battery lifespan from extracted battery features. The boosting design enables the GBRT to address challenging cases through generating an optimal set of trees, and is helpful to handle cases in terms of limited battery degradation samples with many feature inputs, which can not only capture the complex and nonlinear characteristics of batteries dynamics with varying lifespan, but also improve the overall prediction performance. The proposed GBRT method can estimate battery lifespan based on extracted features in an accurate fashion and simultaneously rank the relative importance of each input feature for further deep analysis.

To demonstrate the effectiveness of proposed method, the MIT dataset [21] consisting of 124 battery degradation samples with cycle lives varying from 300 to 2300 cycles using 72 different fast-charging conditions are used for battery prognostics study. While in MIT work, only 3 feature set sizes (1, 6, and 9) were considered, in this work, a total of 72 features are constructed from raw battery data of first 250 cycles in both discharge time dimension (short term) and life cycle dimension (long term). For an optimal lifespan prediction, key hyper-parameters of

proposed GBRT model including learning rate, number of trees, and maximum number of splits are investigated. The relative importance of our input 72 features is systematically studied and analyzed based on the prediction results of determined optimal GBRT model. Instead of directly splitting data into three batches as in original MIT work [21], a 5-folder cross validation technique [29] is adopted in this paper to better evaluate the algorithm performance and prevent potential overfitting. The prediction performance of GBRT model is then compared with several other popular machine learning methods including decision tree, random forest, SVM, and GPR. Comparative studies show that the proposed GBRT method significantly improves the performance of battery lifespan prediction with limited data samples and various input features, with testing root mean square error around 83 and mean average percentage error around 7%.

The rest of this paper is outlined as follows. A prediction framework based on the GBRT model is proposed in Section 2. The experimental data and extracted features are introduced in Section 3. The experimental results are demonstrated in Section 4. Section 5 concludes this paper.

2. Gradient boosting model for lifespan prediction

In this section, a GBRT model is presented to predict battery lifespan from a set of pre-selected features. The GBRT is based on the decision tree model, which is a popular machine learning method for its ability to describe complex relationships between general input-output data and interpretability of input features.

2.1. Decision tree

The decision tree [29] splits the feature space into sub-regions according to classification/regression features and then constructs different (linear) models to fit each region. The split point is determined such that the residual sum is best reduced, where the residual is defined as the difference between the observed lifespan and the predicted lifespan. Recursively performing this process, a single tree-like structure that "best" depicts the underlying input-output relationships in a dataset can be obtained. Decision tree represents information in an intuitive and visualizable way and holds several other advantageous properties such as strong interpretability, robustness to outliers, and ease for implementation.

2.2. Gradient boosting regression tree

GBRT enhances the traditional decision tree approach by combining a statistical technique called boosting, of which the core idea is to aggregate a set of "weak" models to form a single "strong" consensus model [25], instead of building one optimized model. In GBRT, new decision trees are generated sequentially by minimizing the existing residuals. This sequential model building process is fundamentally a form of functional gradient descent, i.e. the prediction is optimized by adding a new tree at each step to minimize the loss function [32].

Assuming a training set $\{(x_i, y_i)\}_{i=1}^{N}$ is present, where x and y represent the input features and response lifespan, respectively. $L(y, F(x))$ is a pre-selected feasible loss function measuring the amount of the predicted lifespan $F(x)$ deviates from the true response lifespan y, which may be the squared error, absolute error, Huber error etc. Assuming M decision trees will be constructed, the GBRT framework starts with an initial model $F_0(x)$. For each iteration $m = 1, 2, \cdots, M$, compensating the residues is equivalent to optimizing the expansion coefficients ρ_m and α_m:

$$(\rho_m, \alpha_m) = \text{argmin}_{\rho,\alpha} \sum_{i=1}^{N} L[y_i, F_{m-1} + \rho h(x_i; \alpha)], \tag{1}$$

and get:

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; \alpha_m). \tag{2}$$

However, Eq. (1) is hard to be directly solved. Nonetheless, $\rho h(x_i; \alpha)$ can be regarded as increment along $h(x_i; \alpha)$, since the gradient boosting is an additive model. According to the idea of gradient descent, the optimal α_m can be solved using least square formulation,

$$\alpha_m = \text{argmin}_{\alpha,\beta} \sum_{i=1}^{N} [r_i - \beta h(x_i; \alpha)]^2, \tag{3}$$

where β is weight factor, r_i is the negative gradient evaluated using the previous model,

$$r_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \cdots, N. \tag{4}$$

The weight of the resulting decision tree or the gradient-descent step size can be further optimized as a one-dimensional optimization problem,

$$\rho_m = \text{argmin}_{\rho} \sum_{i=1}^{N} L[r_i, F_{m-1} + \rho h(x_i; \alpha_m)]^2. \tag{5}$$

Finally, the newly evaluated residue model will be added to the previous model following Eq. (2). The generic gradient boosting method is summarized in Table 1.

There are plenty of smooth loss functions that are feasible in the gradient boosting framework, such as AdaBoost, LogitBoost and L2Boosting [33]. In this work, the squared loss function is used for its simplicity and coherence for regression problem:

$$L(y, F_M(x)) = \sum_{i=1}^{N} (y_i - F_M(x_i))^2. \tag{6}$$

To prevent overfitting and increase the model's generalization ability, some sort of regularization techniques are often preferred during the training. Specifically, shrinkage, also known as learning rate, is often used in gradient boosting, which introduces a new variable ν_m to control the model update rate as

$$F_m(x) = F_{m-1}(x) + \nu_m \cdot \rho_m h(x; \alpha_m), \quad 0 < \nu_m < 1, \tag{7}$$

where the smaller ν_m is, the slower the model is updated. It has been reported that using small learning rates yields improvements in models' generalization ability over gradient boosting without shrinkage [29], which comes at the cost of increased computational time as more decision trees will be needed. Besides, several other parameters, such as the number of trees M and their depths (maximum number of splits), which closely relate to the final tree's structure and model complexity, also need to be fine-tuned to optimize model performance.

2.3. Relative importance of influential factors

In contrast to general modeling approaches such as neural network, autoregressive integrated moving average model, and SVM, the GBRT model can identify and rank the importance of predictor variables on

Table 1
The gradient boosting algorithm.

Input: Training set $\{(x_i, y_i)\}_{i=1}^{N}$, loss function $L(y, F(x))$, number of iterations M
Initialize: $F_0 = \text{argmin}_{\rho_0} \sum_{i=1}^{N} L(y_i, \rho_0)$
For $m = 1$ to M do:
$r_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \cdots, N$
$\alpha_m = \text{argmin}_{\alpha,\beta} \sum_{i=1}^{N} [r_i - \beta h(x_i; \alpha)]^2$
$\rho_m = \text{argmin}_{\rho} \sum_{i=1}^{N} L[r_i, F_{m-1} + \rho h(x_i; \alpha_m)]^2$
$F_m(x) = F_{m-1}(x) + \rho_m h(x; \alpha_m)$
end For
Output: The final regression function $F_m(x)$

response predictions at training stage. Understanding the relative importance of each variable can provide insights into inherent mechanism that bridges the input and output variables.

For a single decision tree T, the measures are based on the number of times a variable is selected for splitting. The relative importance of the predictor x_k in predicting the response is approximately calculated by Ref. [34]:

$$I_k^2(T) = \sum_{t=1}^{J-1} \hat{\tau}_t^2 I(x(t) = k), \tag{8}$$

where the summation is over the non-terminal node t of J-terminal node tree T, $x(t)$ is the splitting variable associated with node t, and $\hat{\tau}_t^2$ is the corresponding empirical improvement in squared error as a result of the split. For a collection of decision trees $\{T_m\}_1^M$, Eq. (8) is extended by averaging over all trees:

$$I_k^2 = \frac{1}{M} \sum_{m=1}^{M} I_k^2(T_m). \tag{9}$$

3. Experimental data

3.1. Dataset

The MIT lithium-ion battery cycle life test data [21] is used in this study to construct the training and testing dataset, which up till now is the largest available public dataset for battery long-term degradation study. In total, 124 randomly selected APR18650M1A lithium iron phosphate (LFP)/graphite batteries (with 1.1 A h nominal capacity, from A123 Systems) were cycled with various charging profiles and constant discharging profiles (4C–2.0 V) in a temperature chamber (30°C). Cells were charged from 0% to 80% state-of-charge (SOC) with one of 72

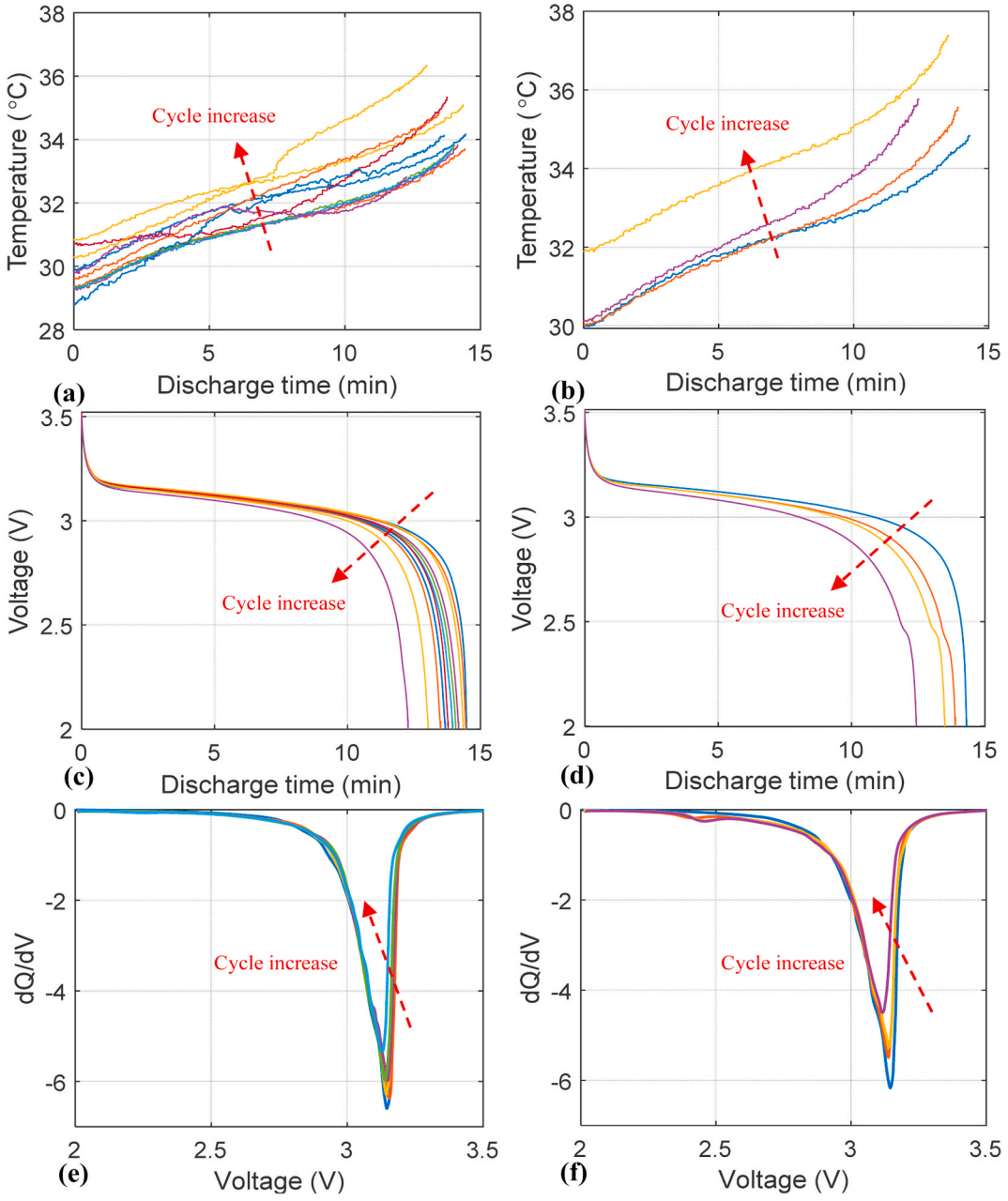

Fig. 1. Measured temperatures, voltages, and IC curves of two samples during the discharge process, left column: sample with 2039 cycle life; right column: sample with 691 cycle life.

different charging policies (with details described in Ref. [21]) and then charged from 80% to 100% SOC under 1C constant-current/constant-voltage mode. The cutoff voltages for the cycle test were at 3.6 V and 2 V, respectively. Voltage, current, and cell temperature were continuously measured and recorded during cycling. Fig. 1 shows the measured temperatures and voltages of two battery samples during the discharge processes, where the temperature and voltage curves are sampled every 200 cycles. The available maximum capacity at each cycle, quantified in ampere-hours (Ah), is calculated as the integral of current over discharge time during the discharge process. And the lifespan of a battery is defined as the number of charge-discharge cycles a battery can run before its available maximum capacity drops below 80% of its nominal value. Fig. 2 plots the capacity evolution of all test samples, where the lifespans of all samples range from 300 to 2500 cycles (one sample with lifespan less than 150 cycles is excluded here).

3.2. Feature construction

A set of features are constructed from raw current, voltage, and temperature data during discharge processes for battery lifespan prediction. These features are identified from the cycling voltage and temperature curves in both discharge time dimension (short term) and life cycle dimension (long term) and are categorized into voltage-related features, capacity-related features, and temperature-related features, as summarized in Table 2. The proposed features either appeared in or are inspired by various existing battery literatures [18,19,21], serving diversified research orientations.

3.2.1. Voltage-related features

Voltage curves, capturing the electrochemical evolution of individual cells during cycling, hold abundant information related to degradation diagnosis. For instance, the derivatives incremental capacity (IC) curves and the $Q(V)$ curves have been exhibiting good results not only for degradation mechanisms detection [1,35], but also for online SOH estimation [18,19,36].

Eight features are extracted from the discharge voltage curves, where V1 and V2 are extracted from the IC curve, which describes the relationship of a capacity change related with a voltage step ($\Delta Q/\Delta V$) during a discharge process [1]. Fig. 1(e–f) plots the IC curves with battery degradation for two cells, where one peak is observed in each discharge cycle. Generally, the peak intensity decreases as cycle increases while its voltage shifts to low voltage at the same time, which reveals important signatures about battery health state [1]. Here feature V1 and V2 are defined as the intensity decrease and voltage shift of two different cycles I and J, respectively.

Feature V3 to V8 are calculated as the summary statistics including minimum, mean, variance, skewness, and kurtosis of the $\Delta Q(V)$ curves. The calculation of these summary statistics refers to Ref. [37]. The $\Delta Q(V)$, as cycle-to-cycle evolution of $Q(V)$, depicts the change in discharge capacity-voltage curves between two cycles, where $Q(V)$ depicts the discharge capacity as a function of voltage. Each statistic is a scalar quantity that depicts the change of two cycles in voltage curves. For example, $\Delta Q_{100\text{-}30}(V) = Q_{100}(V) - Q_{30}(V)$ denotes the capacity change in discharge voltage curves between cycle 30 and 100, where the subscripts indicate cycle number.

For each feature, four cycle-to-cycle evolutions including [30, 100], [30, 150], [30, 200], [30, 250] are considered. For instance, feature V1-1 means the intensity decrease between cycle 30 and 100, feature V1-2 means the intensity decrease between cycle 30 and 150, so on and so forth. We start from cycle 30 instead of the first cycle, as it usually takes some time for the battery to establish internal balance. Moreover, as the degradation rate changes along the cycle life, different end cycle combinations are considered in the hope of gaining better insight about the degradation behavior within first 250 cycles, as well as reducing the unwanted errors caused by noises and outliers and improving the robustness of the prediction. In total, 32 features are identified based on discharge voltage curves.

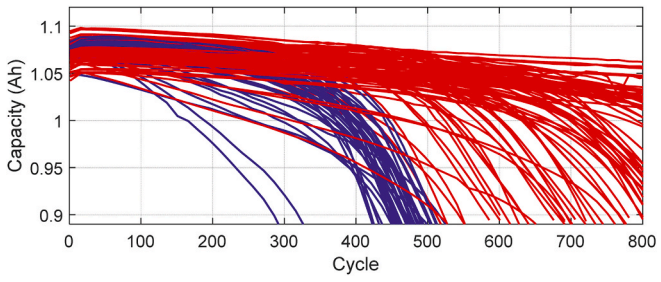

Fig. 2. Capacities v.s. cycle of all tested samples: red lines: batteries with lifespan larger than 500 cycles; blue lines: batteries with lifespan less than 500 cycles. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Extracted features for battery lifespan prediction. ($I = 30$; $J = 100$, 150, 200, and 250).

Type	Feature of interest
Voltage-related features	V1: Intensity decrease of IC peak from cycle I to J V2: Voltage shift of IC peak from cycle I to J V3: Minimum of $\Delta Q_{I-J}(V)$ V4: Maximum of $\Delta Q_{I-J}(V)$ V5: Mean of $\Delta Q_{I-J}(V)$ V6: Variance of $\Delta Q_{I-J}(V)$ V7: Skewness of $\Delta Q_{I-J}(V)$ V8: Kurtosis of $\Delta Q_{I-J}(V)$
Capacity-related features	C1: p_1 of model A, fit to capacity fade curve from cycle I to J C2: p_2 of model A, fit to capacity fade curve from cycle I to J C3: p_3 of model B, fit to capacity fade curve from cycle I to J C4: p_4 of model B, fit to capacity fade curve from cycle I to J C5: p_5 of model C, fit to capacity fade curve from cycle I to J C6: p_6 of model C, fit to capacity fade curve from cycle I to J C7: p_7 of model C, fit to capacity fade curve from cycle I to J
Temperature-related features	T1: average surface temperature of discharge process, difference between cycle I to J T2: maximum surface temperature of discharge process, difference between cycle I to J T3: minimum surface temperature of discharge process, difference between cycle I to J

3.2.2. Capacity-related features

The capacity evolution also holds in store numerous information regarding cell degradation. Quite some degradation models have been put forward for battery remaining useful life prediction, the model parameters are adopted as capacity-related features. Three commonly-used simple models including linear model (model A) [21], square-root-of-time model (model B) [13], and coulombic efficiency-based model (model C) [7] are considered in this work. Model B and C are both semi-empirical models and have been widely used to capture the capacity degradation of lithium-ion batteries. With a focus on the critical fade mechanism of LFP batteries: loss of lithium inventory, they construct simple functions through experimental observation and data analysis. The square-root-of-time model depicts lithium loss via modeling the thickening of solid electrolyte interface film following the finding that the film thickness increases proportional to the square root of time [38]. In comparison, the coulombic efficiency-based model derives relationship between coulombic

efficiency and battery degradation based on the finding that the coulombic efficiency indicates the battery degradation rate [1]. The details of three models are as following:

$$\begin{aligned} &\text{Model A:}\quad C_k = p_1 \cdot k + p_2 \\ &\text{Model B:}\quad C_k = p_3 \cdot \sqrt{k} + p_4, \\ &\text{Model C:}\quad C_k = p_5 \cdot p_6^k + p_7 \end{aligned} \tag{10}$$

where C_k is the discharge capacity at cycle k, p is the pre-determined model parameter. Specifically, p_6 in model C represents the columbic efficiency. Feature C1 to C7 are defined as the presented seven model parameters. These features are evaluated by initializing the above-mentioned models using data from four different data ranges as in voltage-related features. In total, 28 features are extracted here. For instance, feature C1-1 and C2-1 denote fitted parameter values p_1 and p_2 of the linear model using capacity data within 30th cycle and 100th cycle.

3.2.3. Temperature-related features

Temperature is another key factor that may influence the battery degradation. Usually, high temperature results in low cell lifespan. Although under constant operating temperature conditions, there exists temperature variations, which are useful for battery diagnostics. As in Fig. 1(c–d), the battery surface temperature increases with the depth of discharge increases. And as battery degradation, the surface temperature also increases, which is related to the decrease of both the electrical conductivity of the electrolyte and the lithium-ion diffusivity [28]. The temperature variations between different cells may present useful information of battery life prediction. Here feature T1, T2, T3 are defined as the difference of average, maximum, and minimum temperatures of two cycles. Same as in Section 3.2.1, four cycle-to-cycle evolutions are considered here to capture the temperature variances along the degradation process. In total, 12 features are identified based on temperature curves.

3.3. Training and testing

Fig. 3 shows the flowchart of proposed prognostic method for lifespan prediction of lithium-ion batteries. First, raw current, voltage, and temperature data are measured and collected. Then voltage-related, capacity-related, and temperature-related features are extracted from these raw data according to criteria in Section 3.2. The extracted data are then spitted into training dataset (2/3) and testing dataset (1/3). The selected features are fed into the GBRT model to train a well regression model, where 5-folder cross-validation is adopted to optimize the GBRT parameters. By using well-trained GBRT model, lifespan prediction can be performed on new testing data.

3.4. Metrics

The performance of GBRT model is evaluated by the following four statistical metrics:

Mean absolute error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{11}$$

Root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{12}$$

Mean absolute percentage error (MAPE):

$$\text{MAPE} = \frac{1}{n}\sum_{k=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100\%. \tag{13}$$

R^2 :

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{k=1}^{n}(y_i - \bar{y})^2}. \tag{14}$$

where n represents the number of battery samples, y_i and $\hat{y}_i$ represents the experimental values and the predicted values for sample i. The MAE measures how close estimates are to the corresponding outcomes. The RMSE, which characterizes the variation in errors, is more sensitive to large errors than the MAE. The MAPE expresses the error as a percentage and evaluates the performance in terms of relative error. For the three metrics, small value indicates good model performance and high values indicates poor model performance. The R^2 is percentage-based metric, and ideally the R^2 is as close as possible to 100% or 1.

4. Experimental results

4.1. Optimization of the model parameters

In this section, a GBRT model is trained and validated for battery lifespan prediction on the MIT dataset. The model takes extracted features as input and output the estimated cycle life for each battery sample. Hyper-parameters such as the number of trees, the learning rate, and the maximum number of splits are key to the performance of GBRT model. To determine the optimal hyper-parameters, the GBRT model is trained and validated with various number of trees (1–1000), learning rates (0.005–0.5), and maximum number of splits ($2^0 - 2^6$). To prevent overfitting, a 5-folder cross validation technique [29] is used to evaluate the prediction performance, where the data is randomly partitioned into 5 equal-sized subsets and in each round, a single subset is selected as the validation data while the remaining 4 subsets are used as the training data. The 5-folder results are then averaged over the rounds to produce a cross-validated result.

Fig. 4 shows the cross-validated prediction performance versus number of trees. In general, cross-validated errors reduce with more

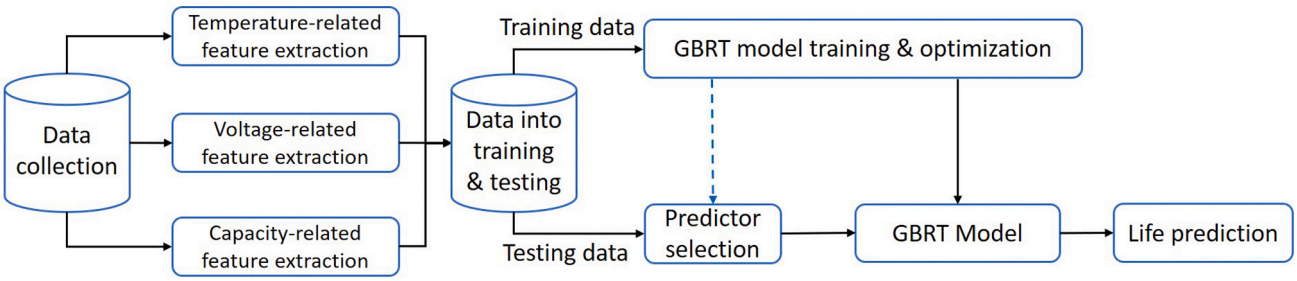

Fig. 3. Flowchart of proposed prognostic method for lifespan prediction of lithium-ion batteries.

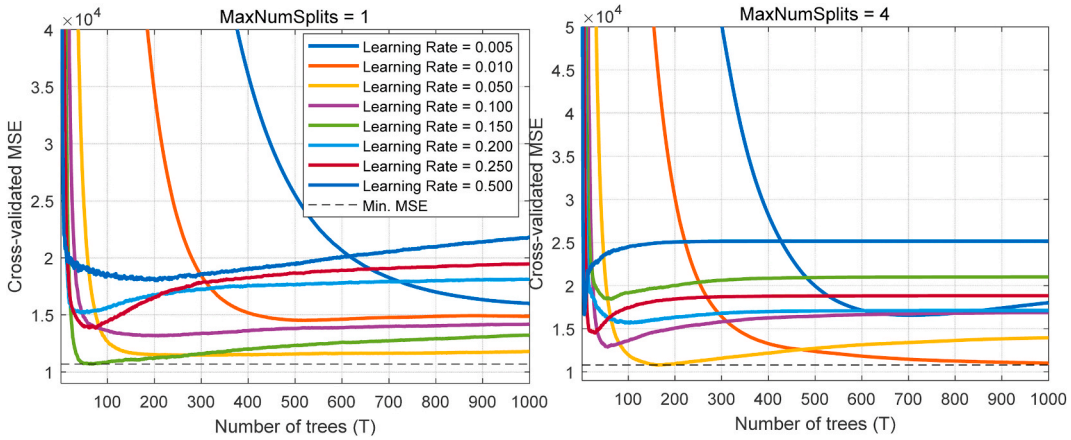

Fig. 4. Cross-validated MSE against tree number for models trained with various learning rates and different maximum number of splits.

base trees available. However, as more trees are used, over-fitting may occur, which affects the prediction performance on unforeseen validation data. Therefore, for a fixed learning rate, the cross-validated RMSE decreases and then slowly increases as overfit emerges. Subsequently, there is an optimal tree number that minimizes the cross-validated RMSE for a given learning rate and maximum number of splits. Moreover, as learning rate limits the contribution of base trees, the cross-validated RMSE reaches its minimal with less tree numbers at larger learning rate. For example, in Fig. 4(a), the line with learning rate 0.05 has relatively slower decreasing rate as compared with that of 0.5, which reaches its minimum error at tree number 10. Finally, when larger number of splits are allowed, as in Fig. 4(b) where the maximum number of splits is 4, best prediction performance is generally obtained with fewer tree numbers for the same learning rate. For instance, for learning rate 0.1, the minimal cross-validated RMSE occurs with tree number at 200 and 50 for splits 1 and 4, respectively. In general, an optimal number of trees exists which minimizes the cross-validation RMSE. The optimal tree number decreases with larger learning rate and generally decreases with a greater maximum number of splits allowed.

According to the above experiment results, in most cases there is a global minimum associated with tree number for a given learning rate and maximum number of splits combination. A line search [39] is thus performed to determine the optimal tree number for each combination, where the initial tree number is heuristically fixed at 1 and bounded up to 1000. Hyper-parameters associated with the smallest RMSE of all cases are then selected to construct the optimal model, which in this case, number of trees = 25, maximum number of splits = 2, learning rate = 0.25. Although these hyper-parameters vary according to different training and validation data combination, they are determined with the same strategy.

With determined hyper-parameters, the prediction performance of proposed method can now be evaluated. Fig. 5(a) shows the prediction results where the X and Y axis represent the ground-truth and the predicted lifespan, respectively. The blue-dot and the red-triangle represent the training results and the testing results, respectively. The experiment is repeated three other times with different training and testing dataset combinations, all the prediction results are summarized in Table 3. In all cases, the testing RMSEs are less than 92 and the MAEs are less than 62. The MAPE, or equivalently, the mean absolute percentage error is around 8%. In other words, the deviation between the predicted lifespan and the ground-truth are less than 62 cycles, or equivalently 8%, for each sample and in every case. In the MIT work [21], using 'discharge' model (best model), the RMSE was 91 and 173 for the primary and secondary test sets, respectively, while the MAPE was 13% and 8.6% for the primary and secondary test sets, respectively. By comparison, the proposed method provides much better prediction results. In summary, the proposed method can capture the inherent nonlinearities in battery dynamics and yield superior prediction accuracy.

Table 3
Testing results of 4 repeated experiments.

Experiment	RMSE	MAE	MAPE	R^2	Time
Case 1	82.79	56.34	0.07	0.94	0.024
Case 2	90.57	61.05	0.07	0.92	0.021
Case 3	91.47	62.16	0.08	0.92	0.0187
Case 4	74.79	54.62	0.08	0.96	0.0167

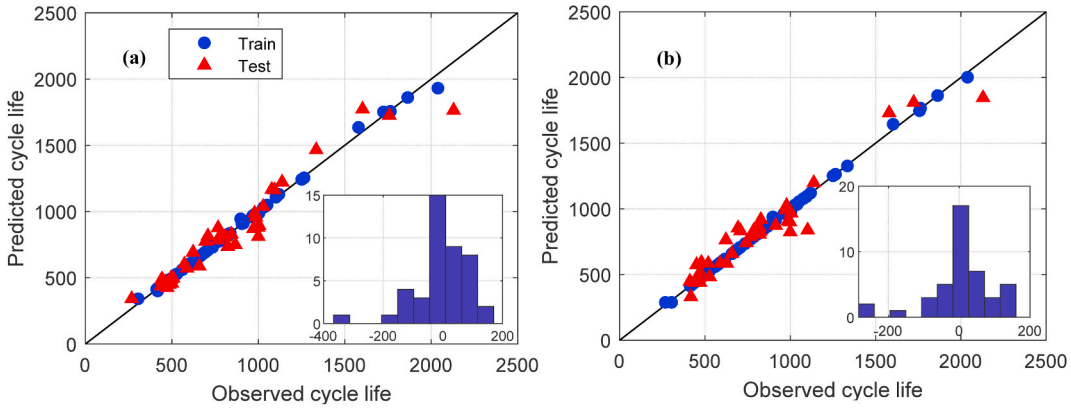

Fig. 5. Prediction Results: a) case 1; b) case 3.

4.2. Relative importance of input features

Fig. 6 plots the relative importance of top 20 features to the model output according to the proposed GBRT model. Normalization is performed such that the importance of all input features sums up to 1. From Fig. 6(a), it can be seen clearly that the importance of different features varies. Fig. 6(b) shows the relative importance of different feature groups. The voltage-related features contribute the most (up to 70%) to the prediction. In contrast, the contribution of temperature-related features is less than 1%. The capacity-related features, which lies in between, contributes around 29% to the results. If we use the three group features separately, the prediction RMSEs of voltage-related, capacity-related, and temperature-related features are 88, 200, and 312 (Table 4), respectively, which agrees with the relative importance of each feature group.

As to the importance of each feature alone, among all features, feature V6-4 (Variance of $\Delta Q_{250-30}(V)$) and V6-2 (Variance of $\Delta Q_{150-30}(V)$) contribute the most to the predicted battery lifespan, with relative importance of 9.9% and 9.8%, respectively. Interestingly, the top 5 features together account for approximately 46%, and all of them are voltage-related features. Among voltage-related features, variances across different cycle intervals V6-4, V6-2, V6-1, and V6-3 contribute the most, with a total contribution of 34.6%. Besides variances, mean V5 (V5-4, V5-3, V5-1) and minimum V3 (V3-4, V3-3, V3-2, V3-1) are also important features with contributions of 11.5% and 17.6%, respectively. Feature V1-4 (Intensity decrease of IC peak between cycle 30 and 250) is also an important feature with contribution of 5.9%. In contrast, the maximum, skewness, kurtosis metrics have insignificant impacts. Among capacity-related features, feature C7-4 (p_7 of CE-based model, fit to capacity fade curve from cycle 30 to 250), C7-2, and C7-3 are top 11 important features, with relative importance of 6.4%, 5.8%, and 4.8%, respectively. Besides C7, C5 (p_5 of model C, C5-3 (2.7%), C5-2 (0.7%), C5-4 (0.5%)), C1 (slope of model A, C1-4 (2.3%)), and C4 (p_4 of model B, C4-4 (1.9%)) are also among top 20 important features. Compared with linear model and square-root-of-time model, features extracted from coulombic efficiency-based model provide more importance for battery lifespan prediction.

As to the temperature-related features, feature T3-1 (minimum surface temperature of discharge process, difference between cycle 30 and 100), contributes the most to the prediction. Nonetheless, its relative importance is less than 0.5% and ranks 28 among all the features. The correlation between the proposed temperature-related features and battery lifespan prediction is not observed by the GBRT model on the sample dataset, this may because the extracted temperature range in this work is too narrow.

To see the effectiveness of the above importance ranking, battery lifespan predictions are re-conducted with selected features. Table 4 tabulates the prediction performance when different feature combinations are considered. When top 20 features are used for prediction, we can almost get the same prediction performance as when all the features are selected, with RMSE of 83 and R^2 of 94%.

Table 4
Prediction performance using different feature combinations.

Selected Features	RMSE	MAE	MAPE	R^2
All features	82.79	56.34	0.07	0.94
Voltage-related features	88.09	67.37	0.10	0.93
Capacity-related features	200.33	111.69	0.14	0.66
Temperature-related features	312.43	207.18	0.30	0.19
Top 2 features	87.48	65.65	0.09	0.93
Top 6 features	86.20	63.44	0.09	0.94
Top 12 features	86.08	63.31	0.09	0.93
Top 16 features	84.15	62.64	0.08	0.94
Top 20 features	83.30	58.44	0.08	0.94

4.3. Model comparison

In this section, the performance of GBRT method on battery lifespan prediction is compared with several other machine learning methods including decision tree, random forest (RF), SVM, and GPR. For SVM, the medium Gaussian function is selected as the kernel function. For GPR, the exponential kernel function is used. The same training and testing data are used to evaluate the performance of these methods, where 5 cross-validation technique is used during the training process for model parameter optimization.

Table 5 shows the prediction performance of all methods using all features and top-20 features as input factors, respectively. The proposed GBRT method presents the best prediction performance in both cases, in terms of RMSE, MAE, MAPE, R^2 values, and computation time of the whole prediction process. The computation costs of these feed-forward machine learning methods are comparable and all less than 0.03s on our lab computer, which are acceptable for onboard applications. As another ensemble learning method, RF yields slightly worse prediction results. On the contrary, SVM receives the worst performance for lifespan prediction of all cases. When selected features are used, there is a significant improvement on the prediction performance for SVM and GPR. For the other three tree-based methods, the prediction performance doesn't show much difference, which means the tree-based model can handle the complex input-output problem. After feature reduction, all the computation time significantly reduces. Fig. 7 plots the prediction results by using selected features. The GBRT method produces satisfying life predications for all battery samples. In contrast, the performance of other four models are satisfactory for batteries with short-to-normal cycle life and are undesirable for batteries with long cycle life. While this can be explained by the fact that there are very few

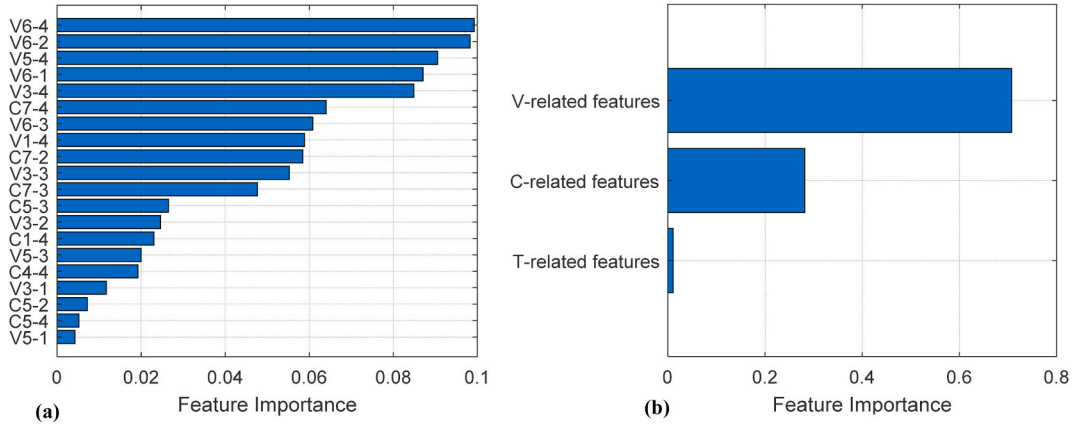

Fig. 6. Relative importance ranking for features used for battery lifespan prediction.

Table 5
Comparison with different models for battery lifespan prediction.

Model	Performance (all features)					Performance (selected features)				
	RMSE	MAE	MAPE	R^2	Time (s)	RMSE	MAE	MAPE	R^2	Time (s)
GBRT	82.8	56.3	0.07	0.94	0.024	84.5	58.9	0.08	0.93	0.008
Tree	117.0	75.0	0.10	0.89	0.026	117.7	77.1	0.11	0.89	0.006
SVM	153.7	88.2	0.13	0.83	0.026	137.1	79.7	0.11	0.89	0.005
RF	84.5	59.3	0.08	0.94	0.021	88.8	64.9	0.10	0.93	0.015
GPR	145.3	95.2	0.15	0.81	0.023	104.0	60.0	0.08	0.91	0.005

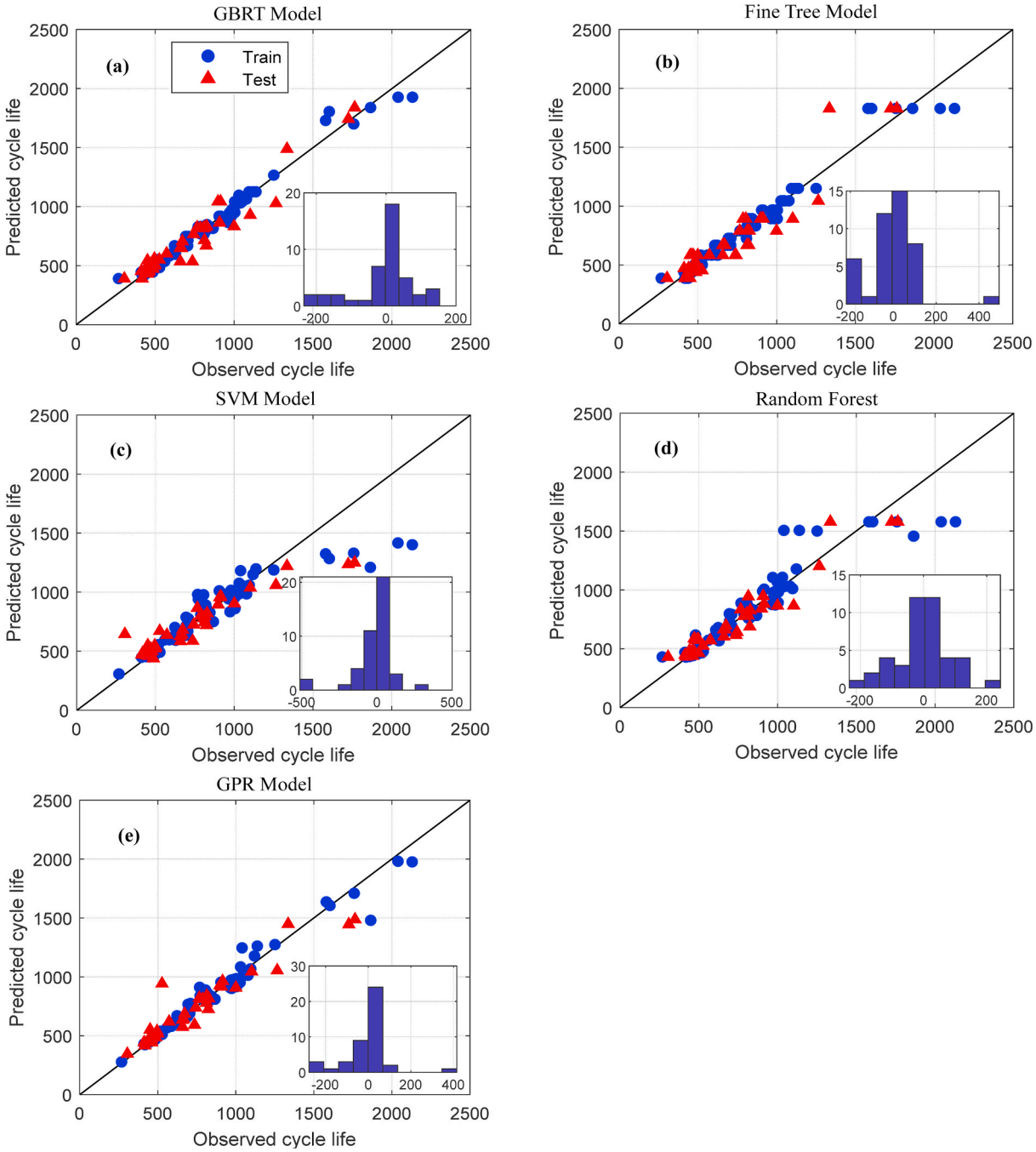

Fig. 7. Prediction results with selected features: a) GBRT model; (b) Tree model; (c) SVM model; (d) RF model; (e) GPR model.

battery samples with long cycle life, it also reflects the superiority of GBRT method in modeling complex relations between extracted battery features and battery lifespan.

5. Conclusion

Battery lifespan prediction is important for prognostics and

diagnostics of battery management systems. In this paper, a GBRT model was proposed to model the complex nonlinear battery dynamics and predict battery lifespan through various extracted battery features. The MIT dataset, which is the largest publicly available dataset consisting of 124 battery degradation data (cycle lives varying from 148 to 2300), was used to demonstrate the effectiveness of GBRT model. Various features including voltage-related features, capacity-related features, and temperature-related features were constructed and explored for effective lifespan prediction. Key hyper-parameters including learning rate, number of trees, and maximum number of splits were investigated for optimal GBRT prediction. Comparative studies confirm that, compared with decision tree, SVM, RF, and GPR models, the proposed GBRT model was significantly superior for battery lifespan prediction, with absolute mean average percentage error around 7%. In addition, it was found that the variance extracted from the $\Delta Q(V)$ curve was the greatest contributor to lifespan prediction with more than 34.6% relative importance. Other voltage-related features, such as mean and minimum of the $\Delta Q(V)$ curve were also found significant. Next to the voltage-related features, capacity-related features were weaker but still effective, of which the total importance was around 30%. In contrast, correlation between the proposed temperature-related features and battery lifespan prediction was not observed by the GBRT model on the sample dataset.

CRediT authorship contribution statement

Fangfang Yang: Conceptualization, Data curation, Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Dong Wang:** Investigation, Methodology, Software. **Fan Xu:** Software, Validation, Writing - review & editing. **Zhelin Huang:** Validation. **Kwok-Leung Tsui:** Project administration, Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was partly supported by the General Research Fund (Project No. CityU 11204419) and the General Research Fund (Project No. CityU 11216014).

References

[1] F. Yang, D. Wang, Y. Zhao, K.-L. Tsui, S.J. Bae, A study of the relationship between coulombic efficiency and capacity degradation of commercial lithium-ion batteries, Energy 145 (2018) 486–495.
[2] Y. Wang, G. Gao, X. Li, Z. Chen, A fractional-order model-based state estimation approach for lithium-ion battery and ultra-capacitor hybrid power source system considering load trajectory, J. Power Sources 449 (2020) 227543.
[3] Y. Dai, L. Cai, R.E. White, Capacity fade model for spinel LiMn2O4 electrode, J. Electrochem. Soc. 160 (1) (2013) A182–A190.
[4] G. Ning, B.N. Popov, Cycle life modeling of lithium-ion batteries, J. Electrochem. Soc. 151 (10) (2004) A1584–A1591.
[5] X. Zheng, H. Fang, An integrated unscented kalman filter and relevance vector regression approach for lithium-ion battery remaining useful life and short-term capacity prediction, Reliab. Eng. Syst. Saf. 144 (2015) 74–82.
[6] Z. Wei, K.J. Tseng, N. Wai, T.M. Lim, M. Skyllas-Kazacos, Adaptive estimation of state of charge and capacity with online identified battery model for vanadium redox flow battery, J. Power Sources 332 (2016) 389–398.
[7] F. Yang, X. Song, G. Dong, K.-L. Tsui, A Coulombic Efficiency-Based Model for Prognostics and Health Estimation of Lithium-Ion Batteries, Energy, 2019.
[8] Y. Wang, Z. Chen, A framework for state-of-charge and remaining discharge time prediction using unscented particle filter, Appl. Energy 260 (2020) 114324.
[9] D. Wang, F. Yang, K.-L. Tsui, Q. Zhou, S.J. Bae, Remaining useful life prediction of lithium-ion batteries based on spherical cubature particle filter, IEEE Trans. Instrum. Meas. 65 (6) (2016) 1282–1291.
[10] M.V. Micea, L. Ungurean, G.N. Carstoiu, V. Groza, Online state-of-health assessment for battery management systems, IEEE Trans. Instrum. Meas. 60 (6) (2011) 1997–2006.
[11] W. He, N. Williard, M. Osterman, M. Pecht, Prognostics of lithium-ion batteries based on Dempster–Shafer theory and the Bayesian Monte Carlo method, J. Power Sources 196 (23) (2011) 10314–10321.
[12] F. Yang, D. Wang, Y. Xing, K.-L. Tsui, Prognostics of Li (NiMnCo) O 2-based lithium-ion batteries using a novel battery degradation model, Microelectron. Reliab. 70 (2017) 70–78.
[13] X. Han, M. Ouyang, L. Lu, J. Li, A comparative study of commercial lithium ion battery cycle life in electric vehicle: capacity loss estimation, J. Power Sources 268 (2014) 658–669.
[14] M. Rezvani, S. Lee, J. Lee, A Comparative Analysis of Techniques for Electric Vehicle Battery Prognostics and Health Management (PHM), SAE Technical Paper, 2011.
[15] D. Yang, X. Zhang, R. Pan, Y. Wang, Z. Chen, A novel Gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve, J. Power Sources 384 (2018) 387–395.
[16] J. Meng, L. Cai, G. Luo, D.-I. Stroe, R. Teodorescu, Lithium-ion battery state of health estimation with short-term current pulse test and support vector machine, Microelectron. Reliab. 88 (2018) 1216–1220.
[17] Z. He, M. Gao, G. Ma, Y. Liu, S. Chen, Online state-of-health estimation of lithium-ion batteries using Dynamic Bayesian Networks, J. Power Sources 267 (2014) 576–583.
[18] C. Weng, Y. Cui, J. Sun, H. Peng, On-board state of health monitoring of lithium-ion batteries using incremental capacity analysis with support vector regression, J. Power Sources 235 (2013) 36–44.
[19] M. Berecibar, F. Devriendt, M. Dubarry, I. Villarreal, N. Omar, W. Verbeke, J. Van Mierlo, Online state of health estimation on NMC cells based on predictive analytics, J. Power Sources 320 (2016) 239–250.
[20] R.R. Richardson, C.R. Birkl, M.A. Osborne, D.A. Howey, Gaussian process regression for in situ capacity estimation of lithium-ion batteries, IEEE Trans. Ind. Inf. 15 (1) (2018) 127–138.
[21] K.A. Severson, P.M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M.H. Chen, M. Aykol, P.K. Herring, D. Fraggedakis, Data-driven prediction of battery cycle life before capacity degradation, Nat. Energy 4 (5) (2019) 383.
[22] X. Hu, J. Jiang, D. Cao, B. Egardt, Battery health prognosis for electric vehicles using sample entropy and sparse Bayesian predictive modeling, IEEE Trans. Ind. Electron. 63 (4) (2015) 2645–2656.
[23] H. Pan, Z. Lü, H. Wang, H. Wei, L. Chen, Novel battery state-of-health online estimation method using multiple health indicators and an extreme learning machine, Energy 160 (2018) 466–477.
[24] F. Yang, S. Zhang, W. Li, Q. Miao, State-of-charge estimation of lithium-ion batteries using LSTM and UKF, Energy (2020) 117664.
[25] H.-T. Lin, T.-J. Liang, S.-M. Chen, Estimation of battery state of health using probabilistic neural network, IEEE Trans. Ind. Inf. 9 (2) (2012) 679–685.
[26] A.J. Salkind, C. Fennie, P. Singh, T. Atwater, D.E. Reisner, Determination of state-of-charge and state-of-health of batteries by fuzzy logic methodology, J. Power Sources 80 (1–2) (1999) 293–300.
[27] Y. Zhang, R. Xiong, H. He, M.G. Pecht, Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries, IEEE Trans. Veh. Technol. 67 (7) (2018) 5695–5705.
[28] A. El Mejdoubi, A. Oukaour, H. Chaoui, H. Gualous, J. Sabor, Y. Slamani, State-of-charge and state-of-health lithium-ion batteries' diagnosis according to surface temperature variation, IEEE Trans. Ind. Electron. 63 (4) (2015) 2391–2402.
[29] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, Springer series in statistics, New York, 2001.
[30] J. Ko, S.N. Baldassano, P.-L. Loh, K. Kording, B. Litt, D. Issadore, Machine learning to detect signatures of disease in liquid biopsies–a user's guide, Lab Chip 18 (3) (2018) 395–405.
[31] R.E. Schapire, The Boosting Approach to Machine Learning: an Overview, Nonlinear Estimation and Classification, 2003, pp. 149–171. Springer.
[32] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Anim. Ecol. 77 (4) (2008) 802–813.
[33] P. Bühlmann, B. Yu, Boosting with the L 2 loss: regression and classification, J. Am. Stat. Assoc. 98 (462) (2003) 324–339.
[34] L. Breiman, Classification and Regression Trees, 2017. Routledge.
[35] M. Dubarry, B.Y. Liaw, Identify capacity fading mechanism in a commercial LiFePO 4 cell, J. Power Sources 194 (1) (2009) 541–549.
[36] C.P. Lin, J. Cabrera, Y. Denis, K.L. Tsui, SOH estimation and SOC recalibration of lithium-ion battery with incremental capacity analysis&cubic smoothing spline, J. Electrochem. Soc. 167 (9) (2020), 090537.
[37] Z. Wei, Y. Wang, S. He, J. Bao, A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection, Knowl. Base Syst. 116 (2017) 1–12.
[38] R. Spotnitz, Simulation of capacity fade in lithium-ion batteries, J. Power Sources 113 (1) (2003) 72–80.
[39] L. Grippo, F. Lampariello, S. Lucidi, A nonmonotone line search technique for Newton's method, SIAM J. Numer. Anal. 23 (4) (1986) 707–716.