Battery health prediction using fusion-based feature selection and machine learning

Xiaosong Hu, *Senior Member, IEEE*, Yunhong Che, Xianke Lin, *Member, IEEE*, and Simona Onori, *Senior Member, IEEE*

***Abstract*—State of health (SOH) is a key parameter to assess lithium-ion battery feasibility for secondary usage applications. SOH estimation based on machine learning has attracted great attention in recent years, and holds potentials for battery informatization and cloud battery management techniques. In this paper, a comprehensive study of the data-driven SOH estimation methods is conducted. A new classification for health indicators (HIs) is proposed where the HIs are divided into the measured variables and calculated variables. To illustrate the significance of data preprocessing, four noise reduction methods are assessed in the HIs extraction process; different feature selection methods, including filter-based method, wrapper-based method, and fusion-based method, are applied to select HIs subsets. The four widely used machine learning algorithms, including artificial neural network, support vector machine, relevance vector machine, and Gaussian process regression, are applied and compared. In order to evaluate the estimation performance in potential real usages under future big data era, the three HIs selection methods and four machine learning methods are evaluated using three public data sets and two estimation strategies. The results show that the combination of the fusion-based selection method and Gaussian process regression has an overall superior estimation performance in terms of both accuracy and computational efficiency.**

***Index Terms*—lithium-ion batteries, state of health, feature extraction, feature selection, machine learning, comprehensive comparison**

I. Introduction

Because of its powerful computing power and robustness, machine learning has been widely used in various fields and is an important tool in the era of big data [1]. Lithium-ion batteries are one of the key components in electric vehicles (EVs), smartphones, mobile energy storage equipment, and smart grids, due to their high power and energy density, as well as low self-discharge rate [2, 3]. A battery management system (BMS) is essential to ensure safety and reliability [4]. However, the battery's physical and chemical properties will change during storage and operation, resulting in a decrease in battery capacity and power [5]. The BMS is requested to estimate the state of health (SOH) accurately and reliably in order to monitor the battery aging conditions and provide guidance for the second use [6]. However, the contradiction between limited computing power and the need for more information has driven the battery informatization process and the establishment of cloud battery management technology. This makes the data-driven approaches play important roles in future health prognostic for batteries[7].

The methods for SOH estimation can be divided into two categories: model-based methods and data-driven methods. Model-based methods can be further divided into electrochemical model-based, equivalent circuit model-based, and empirical/semi-empirical model-based methods. In electrochemical model-based methods, the first-principles equations are established based on the battery internal electrochemical processes, and then the accurate state are calculated [8]. However, the computational cost associated with this approach is high, which makes it hard for online applications. In equivalent circuit model-based methods, electric models are established firstly, such as RC equivalent circuit models [9], fractional-order equivalent circuit models [10], and impedance spectrum growth models [11]. Then, filtering algorithms are used to update the model parameters for the SOH estimation [12]. This method can be used online due to low computational cost. Empirical/semi-empirical model-based methods are widely used in the battery SOH estimation. The basic idea of these methods is to fit the capacity loss [13] or internal resistance increase [14] with time or cycles. The methods for data fitting mainly include particle filter [13, 15], particle swarm optimization algorithm (PSO) [14], just to cite a few. According to the fitted model, the SOH could be obtained based on time or running cycles. The advantage of this type of methods is that it is simple and easy for online estimations. However, it is susceptible to noise, and its robustness and accuracy are not enough. Moreover, the fitted curve of one type of batteries is usually not suitable for other types of batteries, and the fitted models need a lot of data and labor.

Recently, data-driven methods have been developed rapidly and applied widely. This type of methods treats batteries as a "black box". The external variables such as voltage, current, and temperature are the inputs, and the estimated SOH can be obtained through a complex calculation of the "black box" [16, 17]. The input variables are called health indicators (HIs). The extraction and selection of HIs in data preprocessing are the foundation and key to the accuracy of SOH estimation [18]. The extraction methods of HIs can be divided into two categories: direct extraction methods based on measured variables and indirect extraction methods based on calculated variables.

The main battery variables that BMS can obtain online are current, voltage, and temperature [19]. Therefore, extracting

This work was in part supported by National Natural Science Foundation of China (Grant No. 51875054 and No. U1864212), Graduate scientific research and innovation foundation of Chongqing, China (Grant No. CYS20018), and Chongqing Natural Science Foundation for Distinguished Young Scholars (Grant No. cstc2019jcyjjq0010), Chongqing Science and Technology Bureau, China. (X. Hu and Y. Che contributed equally to this work), (Corresponding author: X. Hu and S. Onori)

X. Hu, and Y. Che are with the Department of Automotive Engineering, Chongqing University, Chongqing 400044, China. (e-mail: xiaosonghu@ieee.org; cyh@cqu.edu.cn)

X. Lin is with the Department of Automotive, Mechanical and Manufacturing Engineering, University of Ontario Institute of Technology, Oshawa, ON L1G 0C5, Canada. (e-mail: xianke.lin@uoit.ca)

S. Onori is with Department of Energy Resources Engineering, Stanford University, Stanford, CA 94305, USA (e-mail: sonori@stanford.edu)

HIs based-on battery voltage and temperature is the simplest, most direct, and effective way. The voltage and temperature curves change noticeably during charge and discharge, due to the capacity loss and internal resistance increase. For instance, the voltage slope will increase due to the increase of internal resistance and decrease of capacity during the charging process [1, 20]. The capacity loss will shorten the charge and discharge process, thereby reducing the time for charging and discharging [21]. Consequently, the initial voltages during charge and discharge also change during the aging process [22]. The temperature will rise and fall repeatedly during the charge-discharge cycles. However, the maximum temperature value and its appearance time [23], as well as the mean temperature value [24] will change at different aging stages. Therefore, the characteristics of the voltage curve and temperature curve as they change significantly with the aging of the battery can be used as HIs. In addition, time is recorded in BMS, and many widely used HIs are designed as time-dependent HIs. The constant current charge and discharge time [24] [25], and constant voltage charge time [21], are examples of time-related HIs. Also, the combination of time with voltages such as time elapsed at the same voltage change, or the voltage variation during the same time period are also useful HIs [26].

The battery characteristics that are reflected by direct measurement data are useful but somehow limited. In recent years, researchers have also proposed HIs based on indirectly calculated variables. The most widely studied HIs are based on incremental analysis and differential analysis [16]. Incremental capacity (IC) is the change of battery capacity with voltage in a short time, and the IC curve is the curve of IC with voltage [27]. According to the electrochemical reaction process, there are multiple peaks in the IC curve. Each peak is a reflection of the phase equilibrium position on the battery voltage curve [28]. During the battery aging process, the peak value of the IC curve [29], the peak position [30], and the peak area [31] also show changes. Therefore, they are extracted and used as HIs. Differential voltage (DV) can be seen as the opposite of IC, which is the change of battery voltage with capacity over a short time, and the DV curve is the curve of DV with battery capacity [32]. In contrast to the IC curve, the peak in the DV curve represents the phase transition position [33]. Similarly, features such as the valley value [34], and the valley position [35] in the DV curve are also considered as good HIs. Besides, the differential temperature (DT) curve has also been proposed in existing studies [36]. Based on the DT curve, health factors similar to the DV curve can also be extracted. The IC, DV, and DT curves often have a lot of noise, so filtering is required. Common filtering methods include moving average filtering [37], differential filtering [38], wavelet transform [30], and Gaussian filtering [39].

Different HIs have different degrees of correlation with battery SOH. Therefore, the feature selection is required to select the most effective HIs, which would reduce the time of data preprocessing and increase the estimation performance of machine learning algorithms. The existing studies mainly focus on correlation analysis, such as gray correlation [38], Pearson correlation coefficient [39], et al., to analyze the correlation between HIs and battery SOH, and then select a few most relevant features as the final feature set. Another key to data-driven battery SOH estimation is the selection of machine learning algorithms. The algorithms currently used mainly include artificial neural networks (ANN) [40], Support vector machines (SVM) [41], Relevance vector machines (RVM) [42], and Gaussian processes regression (GPR) [20].

Although data-driven battery SOH estimation has been used widely, there are still some research gaps that need to be addressed. The existing studies extracted many HIs based on both measured variables and calculated variables. However, which HIs lead to better SOH estimation? HI selection is an important part of data-driven methods, and the existing studies do not provide enough deep understanding. The HI selection process in the existing studies is mainly based on the correlation coefficients, where HIs with low correlation coefficients are deleted. Nevertheless, feature selection should not only remove irrelevant features but also eliminate redundant features to obtain a more effective feature set for machine learning. Moreover, the accuracy and robustness of the results obtained by different machine learning algorithms are greatly different. The above research gaps reveal the need of a comprehensive evaluation for the data preprocessing and machine learning algorithms in battery SOH estimation to guide applications in the future big data era.

This paper provides a comprehensive study of data-driven battery SOH estimation methods. The methodologies from the data processing to the machine learning algorithms are elaborated in detail. The main contributions are summarized as follows. 1) A classification method for HIs is proposed according to the different HI extraction methods. HIs are divided into direct extraction based on measured variables and indirect extraction based on calculated variables. The specific extraction methods of different HIs are described in detail, and HIs are extracted using three different battery aging data sets. According to IC and DV curves, the noise reduction performance of different filtering algorithms is analyzed. 2) The feature selection methods are reviewed and applied to elaborate the significance of data preprocessing in data-driven approaches. Besides filtering methods (correlation analysis) in the existing papers, the sequence forward search-based wrapper method is used. By combining the filtering and wrapper methods, a fusion method is proposed. The different HI subsets are used as inputs for SOH estimation, and the different estimation performances are evaluated. 3) Four advanced machine learning algorithms, including ANN, SVM, RVM, and GPR, are reviewed and implemented for SOH estimation. The accuracy and computational efficiency are compared and assessed. 4) Three public battery data sets are used for the comparison study. The comprehensive results provide useful guidance for data preprocessing and machine learning selection in potential real-world applications.

The rest of this paper is structured as follows. The three experimental data sets are introduced in Section II. Then in Section III, a new classification of HIs is proposed, and the extraction methods are described in detail, and a comparison for different filtering methods is proposed. Next, the different feature selection methods are reviewed, and a fusion method is introduced in Section IV. In Section V, four used machine algorithms are described and the comprehensive comparison results are given and evaluated in Section VI. Finally, the main conclusions are summarized in Section VII.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTE.2020.3017090, IEEE Transactions on Transportation Electrification

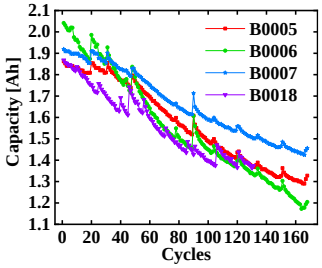
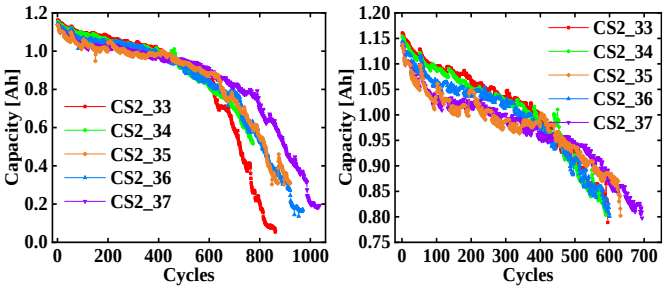


Fig. 1. NASA battery capacity decreases over cycling.



(a) Overall capacity decreases (b) Capacity decreases before 30% loss

Fig. 2. CALCE battery capacity decreases over cycling.

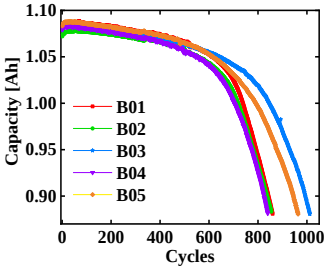


Fig. 3. A123 battery capacity decreases over cycling.

Table. I. Charge protocols of the five battery cells.

Battery	Charge policy	Cycle life	Barcode
B01	5.4C(60%)-3.6C	862	EL150800460653
B02	5.4C(60%)-3.6C	857	EL150800460522
B03	6C(30%)-3.6C	1014	EL150800460477
B04	6C(40%)-3.6C	842	EL150800463198
B05	8C(15%)-3.6C	966	EL150800463229

II. Battery Data Sets

Aging experiments are used to evaluate and verify the methods for battery SOH estimation. Calendar aging and cycling aging are two types of aging experiments, both of them usually take a large amount of time. In this paper, three publicly available experimental data sets are used for the comparative study of SOH estimation methods. Due to different battery types and different experimental conditions in each data set, the extracted HIs are shown to have different effects on SOH estimation.

A. NASA data set

NASA battery data set has been used for the battery SOH estimation [39, 43]. NASA provides six experimental data sets with different temperatures, different discharge rates, and different depths of discharge[44]. Among them, B0005-B0007 and B0018 are the most widely used datasets, which are also selected for the SOH estimation verification in this paper. During cyclic aging, each battery went through three test conditions: constant current-constant voltage charge (CC-CV), CC discharge, and impedance spectrum testing. In charge process, the batteries were first charged at a constant current of 1.5 A until the voltage reaches 4.2 V, then charged at constant voltage till the current drops below 20 mA. In the discharge process, the batteries were discharged at a constant current of 2 A until the voltage reached 2.7 V, 2.5 V, 2.2 V, and 2.5 V for the four batteries, respectively. During the impedance test, the impedance data is obtained by sweeping the frequency with an EIS of 0.1 Hz-5 kHz. Repeat the above three working conditions until the battery reaches the end of its life (30% capacity loss). The capacity curves with the cycle numbers are shown in Fig. 1.

B. CALCE data set

CALCE data set is another widely used set for SOH estimation research [45, 46]. The CS2 dataset includes 6 sub data set. The cycle aging conditions are CC-CV charge and CC discharge. In the charging process, the battery is charged at a constant current of 0.5 C until the voltage reaches 4.2 V, then charged at a constant voltage of 4.2 V till the current drops below 0.05 A. In the discharge process, the battery is discharged at different constant currents until the voltage drops below 2.7 V [47]. In this paper, five batteries, namely CS2_33, CS2_34, CS2_35, CS2_36, and CS2_37 are selected for comparison. The constant discharge current of CS2_33 and CS2_34 is 0.5 C, while the constant discharge current of the other three is 1 C. All the batteries were tested with the Arbin Battery Tester. The capacity loss with cycle numbers of each battery is shown in Fig. 2(a) and the experimental data before the 30% capacity loss is used for SOH estimation in this paper, as shown Fig. 2(b).

C. A123 System data set

The above two data sets use the same charging protocol but different current rates. The third data set from the Massachusetts Institute of Technology and Stanford University [48] is used in this paper for the comparative study. The third dataset was collected on the A123 system, whose protocol is different from the first two datasets. Five battery cells (lithium-ion phosphate (LFP)/graphite) are selected and labeled as B01-B05. The cells have a nominal capacity of 1.1 Ah, and the upper-cutoff and lower-cutoff voltage are 3.6 V and 2.0 V, respectively. The batteries were charged with a two-step fast-charging protocol. This protocol has the format “C1(Q1)-C2”, in which C1 and C2 are the first and second constant-current steps, respectively, and Q1 is the state-of-charge (SOC, %) at which the currents switch. The second current step ended at 80% SOC, after which the cells were charged at 1C CC-CV. The specific charge protocols of these five battery cells are listed in Table I. The 4 C constant current is used to discharge the batteries until it reaches the lower-cutoff voltage. All the batteries are tested in the chamber at a temperature of 30 °C.

III. Health Indicators Extraction

HI extraction is a very important preprocessing step for data-driven estimation, which largely determines the estimation performance. In this paper, a new classification of the HI extraction for batteries is presented. As shown in Fig. 4, the HIs

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTE.2020.3017090, IEEE Transactions on Transportation Electrification

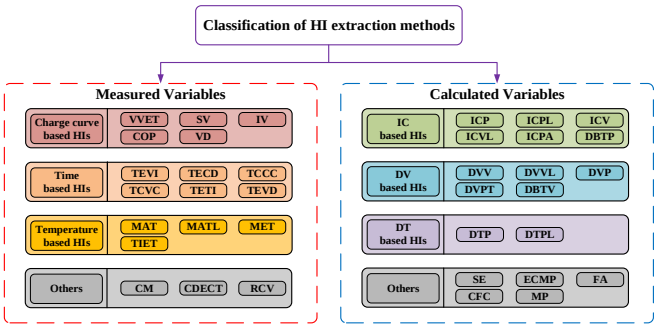


Fig. 4. Classification of HI extraction methods.

are divided into measured variables and calculated variables. In the following subsections, the detailed process of these HI extractions is elaborated.

A. *Measured variables*

The variables that can be measured by BMS include current, voltage, temperature, and time. Therefore, a simple method is to extract the HIs from these measured variables, which has been widely studied by researchers. The measured variable-based HIs can be divided into four categories: voltage curve-based, time-based, temperature-based, and others. The extraction process of each method will be detailed in the following subsections.

1) Voltage curve-based HIs

The internal physical and chemical reactions change gradually with aging during storage and operation, resulting in internal ohm resistance increase and capacity loss [19]. This will cause noticeable changes in voltage curves during the charge and discharge cycles. Some characteristic HIs, such as voltage variation during the equal time (VVET), slope of the voltage curve (SV), initial voltage (IV), cut-off points (COP), and voltage distance (VD) from the initial cycle, show noticeable changes during cycles. The HIs are extracted from the charging process and discharge process.

Generally, the charging protocol remains the same throughout the entire cycle life, and the charging curves will change over cycling due to battery degradation. Therefore, extracting HIs from the charging curves has attracted wide attention. According to the existing studies, the HIs extracted from voltage measurements mainly include: voltage increase during equal charge time (VIECT) [48], slope of charge voltage curve (SCV) [20], upper cut-off voltage point (UVP) [20], and initial charge voltage (ICHV) [22]. Here VIECT means the voltage increase values after a constant charge time from the preset voltage. Due to the capacity loss and resistance increase, VIECT usually gets larger over cycling. The voltage slope will increase during the aging process due to the increase of internal resistance. Therefore, the SCV at the same point on the charging voltage curve can be extracted as HI. In addition, the UVP will gradually shift to an earlier point, and the ICHV will increase gradually as a result of the resistance increase and capacity loss. Therefore, these two variables can be used as HIs.

When the battery undergoes a constant discharge current experiment, the voltage curve shows noticeable changes with cycling. Some traits from the discharge voltage curve can be used as HIs. The HIs in the existing studies mainly include: the voltage decrease during equal discharge time (VDEDT) [49], initial discharge voltage (IDV) [50], voltage curve distance (VD) [51], as well as lower cut-off point (LVP) [41]. The VDEDT indicates the voltage decrement in a fixed discharge time interval. IDV means the initial voltage when the discharge current is loaded. The capacity loss and internal resistance increase will shorten the discharge time, thereby moving the LVP to the left and increasing VD.

2) Time-based HIs

The battery charge/discharge time is closely related to the charge/discharge rate, charge/discharge depth, and aging conditions. In the aging experiments, the rate and depth are always kept constant, therefore the charging and discharging time decrease with the aging cycles. As the battery undergoes degradation, the charge/discharge time decreases due to capacity loss, and voltage rises/falls faster because of the increase of internal resistance during constant current charge/discharge process. Therefore, the constant voltage charge time usually increases. Time-based HIs can reflect the battery health conditions and are used widely for SOH estimation. HIs such as the time interval during equal voltage increase (TEVI) in constant current charge process [25] and the time interval of equal voltage decrease (TEVD) in the discharge process [52] are widely used. These HIs represent the time intervals until the voltage reaches a preset value from certain starting voltage. Similarly, the time interval during the equal current decrease (TECD) in the constant voltage charge process [24] is also extracted by researchers. Another two widely used time-based HIs are the time of constant current charge (TCCC) process [23], and the time of the constant voltage charge (TCVC) process [41]. These two HIs directly reflect the charge capacities in CC and CV process. Because the temperature is also measured by the BMS, the time interval of equal temperature increase (TETI) [53] is also a meaningful time-based HI. It represents the time interval until the temperature reaches a preset value from a starting point. These six HIs are all extracted in this paper for comparative study. In addition, it is worth noting that the capacity-based HIs mentioned by some researchers can also be classified into time-based HIs, because the capacity is calculated by integrating current over time.

3) Temperature-based HIs

Similar to voltage, the temperature is also recorded online, and temperature changes are easy to detect. Therefore, temperature-based HIs are very useful for data-driven SOH estimation. In the charge and discharge aging cycles, the temperature will rise and fall due to the internal chemical and physical reactions. During the aging process, the maximum temperature (MAT) and its location (MATL) [23, 42], as well as the mean temperature (MET) [24] will change. Due to the increase in internal resistance, the temperature will increase under the same load current according to Joule's law. Therefore, MAT and MET will increase as the number of cycles increases. The capacity will decrease, and MATL will change due to the change in charge/discharge time. The temperature increase during the equal time (TIET) [54] is another useful HI. It represents the time interval until the temperature reaches a preset temperature from the initial temperature.

4) Others

In addition to the aforementioned HIs, the cycle number (CM) [55] and the current decrease during equal charge time (CDECT) [48] in the constant voltage charge process are also widely used.

Moreover, the HI that represents the ratio of CC time to CV time (RCV) [54] can also be calculated from TCCC and TCVC.

B. Calculated variables

The health information reflected by measured variables is limited. The calculated variables are used to extract HIs that reflect more information about the SOH. In this type of methods, the measured variables are transformed, and then the HIs are extracted from the transformed curves. The mainstream transformations include IC analysis, DV analysis, and DT analysis [16, 18]. Although the transformations introduce more computational cost, usually more useful HIs can be extracted. With powerful cloud computing, these HIs can also be extracted online in the future. In the rest of this subsection, the calculated variables are described in detail, and a comparison of the filtering methods is conducted.

1) IC-based HIs

IC curve analysis is a powerful non-invasive electrochemical analysis method that can detect subtle changes in electrochemical processes due to capacity loss [27, 56]. Different from EIS or SEI measurement, IC curve can be obtained based on the voltage curve, which is easier to obtained from the BMS [31]. The platform of the voltage curve can be more clearly represented by the peak value, while the rising can be more clearly reflected by the valley in the transformed IC curve. The IC is obtained by the capacity difference to a small voltage interval, and the IC curve shows the IC values versus voltage [30, 38]. Features, such as IC peak value (ICP) and IC valley value (ICV) [27], IC peak location (ICPL) and IC valley location (ICVL) [37], and IC peak area (ICPA) [43] are widely used as HIs. As the charge/discharge process becomes shorter due to capacity loss, the voltage curve starts to have larger slopes. Therefore, the platforms and risings in the voltage curve will decrease, which results in the reduction of ICP and ICV and the early appearance of ICPL and ICVL. These parameters can be extracted as useful HIs to reflect the aging conditions, which shows more information about the internal electrochemical changes. The ICPA is calculated from the area under the IC peak, and due to the decrement of IC, it will also decrease. In addition, some batteries have more than one voltage platform, which will cause multi peaks on the IC curve. The distance between two peaks (DBTP) [57] is also extracted and used for SOH estimation.

2) DV-based HIs

The DV curve is also widely used for SOH estimation, in that it also reflects the electrochemical process. In contrast to the IC curve, the peaks in the DV curve represent the phase transition process while the valleys represent the phase equilibrium positions. The DV is obtained by the voltage difference to a small capacity interval, and the DV curve shows the DV values versus capacities [33]. Based on the DV curve, the peak (DVP) [58]and valley (DVV) [35], the peak location (DVPL) [32, 59], and the valley location (DVVL) [34, 60] are extracted as HIs. The DV curve can be considered as the opposite of the IC curve. Therefore, the extraction methods of these HIs are similar to the IC curve. Also, some batteries have more than one valley/peak. Therefore, the distance (DBTV) can also be used as HIs [58].

3) DT-based HIs

Similar to the voltage curve, the temperature curve can also be transformed into the DT curve to reflect the temperature transition and equilibrium position more clearly. The DT value is obtained by the temperature difference to a small capacity interval, and the DT curve shows the DT values versus capacities [61]. The peak (DTP) and its location (DTPL) are also used for SOH estimation [36]. The DTP is the peak value on the DT curve, which reflects the temperature quick rise stage in the DT curve, and the DTPL is the location or time of the DTP. Similar to the DVP and DVPL in the DV curve, these two parameters will also show changes due to the resistance increase and capacity loss.

4) Others

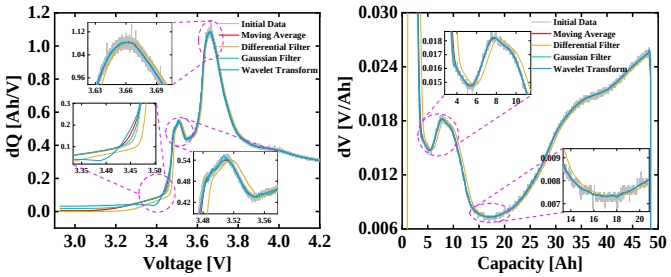
There are some other calculated variables that have been proposed in the literature in addition to the above three kinds. Sample entropy (SE) can be used to evaluate the predictability of the time series and to quantify the regularity of data series [62]. The parameters of the equivalent circuit model (ECMP) under different aging conditions can effectively reflect the changes in the internal resistance of the battery [63]. Recent studies [64, 65] on the use of Fisher information and Cramer-Rao bound reveals the importance of optimizing the current profile to accurately estimate battery parameters, and therefore can improve the accuracy of the ECMP-based HIs. The analysis of the curve fitting coefficients (CFC) can be used to reflect the SOH of the battery, such as OCV curve fitting coefficients [66]. Frequency analysis (FA) is used to extract features from frequency-domain curves for SOH estimation [67]. In addition, the method of mechanical parameter (MP) analysis is also used to estimate battery SOH [68].

C. Comparison of filtering methods

The IC, DV, and DT curves obtained through curve transformation usually are subject to non-negligible noise, which requires the curves to be filtered. Currently, the most widely used methods are: moving average filtering [37], differential filtering [38], wavelet transforming [30], and Gaussian filtering [39].

In this paper, these four commonly used filtering methods are compared. The 1/3 C constant current charging data of an experimental data set are used. The nominal capacity is 50 Ah, the upper cut-off and lower cut-off voltage are 4.2 V and 2.75 V, respectively. The voltage interval of the IC curve is 80 mV, and the capacity interval of the DV curve is 1 Ah. All filtering methods adopt the same window size of 200, and the number of layers decomposed by WT is 8. The filtering results of the four methods are shown in Fig. 5. It can be found that the method based on DF has the maximum deviation from the original signal processing and presents drifting. Furthermore, the other three filtering methods are compared near the inflection point. The method based on WT has smaller fluctuations, and the deviation is larger at the first inflection point of the IC curve. Compared with the MA method, GF can be closer to the original data at the point where IC and DV are about to enter the peak and valley (the first enlarged figure in the figure), and the concave and convex effect is more obvious at the peak and valley values. Therefore, the Gaussian filtering is selected as it provides better signal noise reduction before the HI extraction.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTE.2020.3017090, IEEE Transactions on Transportation Electrification



(a) IC curve filtering (b) DV curve filtering

Fig. 5. Noise reduction using different filtering methods.

IV. HEALTH INDICATORS SELECTION

In general, the extracted HIs are multi-dimensional. Some of these parameters have poor correlations with the output, or there may be strong correlations between the parameters. If they are all used as inputs, it will affect the estimation accuracy and easily lead to overfitting [69]. Selecting appropriate features can reduce the dimensions of data required, consequently, improve the efficiency and reliability of SOH estimation [70]. Feature selection is a data preprocess which eliminates subsets that are not closely related [71]. Feature selection methods can be generally divided into three categories: filter-based methods, wrapper-based methods, and fusion-based methods [72]. Next, in Section 4.1-4.2 the first two feature selection methods are introduced, whereas in Section 4.3 the fusion method based on filter and wrapper is presented.

A. Filter-based method

The filter-based method is a feature screening process separated from model training algorithms. In general, the correlation between each feature and the target is evaluated by scoring each feature, removing the features with a lower correlation than a certain threshold, and keeping the features with a higher correlation as the input of the model [73]. The advantage of the filter method is that it is separate from model training and does not affect the subsequent application after filtering. This makes it easy to apply to high-dimensional data with a relatively small computational cost [72, 74]. However, because the correlation of each feature is calculated separately, the correlation between features is ignored, which may lead to the poor performance of the selected feature set [72]. The correlation coefficient method is widely used in the selection of battery HIs, mainly including gray correlation analysis [24, 38], Pearson correlation coefficient analysis [39, 49], and Spearman correlation coefficient analysis [48, 49]. Among them, the Pearson correlation coefficient has some advantages, such as easy to calculate and express the linear relationship between the input and the target. Therefore, the Pearson correlation coefficient is selected in this paper for the filtering process, which is denoted as follows [39]:

$$Pearson = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \quad (1)$$

where x and y represent the feature and the target at time k, respectively. As for the HIs selection process, all the correlation coefficients between HIs and capacities are evaluated based on equation (1), those less than 0.8 are removed and the rest is used for model training and prediction.

B. Wrapper-based method

The wrapper-based models use specific learning algorithms to evaluate the quality of the selected attributes [73]. The model prediction performance is used to evaluate the feature subset [74]. Typically, a search process is predefined in the space of possible feature subsets, and various feature subsets are generated and evaluated [72]. The general process is: select a subset, evaluate the subset according to the prediction performance, select a new subset, and continue to evaluate until the expected quality is reached [75]. By using a cross-validation strategy in wrapper-based models, the model accuracy can be improved, but this results in a large amount of computation and easy overfitting [72]. The sequence backward search (SBS), which has high efficiency and therefore is widely used, contains all the features firstly, and it removes one feature per iteration until the results reach a preset threshold. In this paper, the SBS is used for the subset searching in the wrapper-based method for each machine learning algorithms.

C. Fusion-based method

From what stated above, it can be found that wrapper and filter complement each other. The filter-based method can search in the feature space efficiently and quickly, but the evaluation deviation of the subsequent learning task is large, while the wrapper-based method has good accuracy, but the search speed is slow. Therefore, the fusion of filter and wrapper method has been widely used to ensure accuracy and reduce computational complexity [76]. For battery SOH estimations, some of the features are not highly correlated with SOH, while others are strongly correlated with each other, which may easily lead to overfitting. The fusion method can filter out the unimportant features based on the filter process, and remove the redundant features based on the wrapper process, to establish a SOH estimator with high accuracy and low computing cost. In this paper, the filter method (Pearson correlation coefficients) is first used to remove HIs with low correlation, and then use the SBS wrapper method to remove the redundant HIs.

V. MACHINE LEARNING ALGORITHMS

After the data preprocessing (feature extraction and selection), the features are inputted to the machine learning models for the training or estimation. Machine learning has many superiorities such as model-free and good robustness. Machine learning models map the input HIs to the output estimation, and then predict the SOH using the later data. The selection of machine learning algorithms also leads to different estimation performances. There are four widely used methods, namely artificial neural network (ANN), support vector machine (SVM), relevance vector machine (RVM), and Gaussian process regression (GPR) for the SOH estimation. In this section, the methods of each machine learning algorithm are introduced in detail, and the estimation performance of each method are compared and evaluated in the next section.

A. Artificial neural network

Inspired by biological systems, especially the human brain, ANNs are designed to mathematically imitate this process in the nonlinear problems [77, 78]. Specifically, the ANN consists of an input layer, an output layer, and multi hidden layers [77].

The input layer receives the pre-processed data and actives as a window between the hidden layers and the features. Next, in the hidden layers, each neuron contains a mathematical model that determines its output based on the input and can be represented by a weighted linear combination encapsulated in the activation function. The total value is converted to the activation value of the node by the activation function. It becomes the input to the next level node until the output activation value is finally determined. The accuracy of the output is determined by the hidden layer number, the neurons in each layer, the weights of each neuron, as well as the activation function. Generally, the more layer number and neuron number, the more accurate the model is, but riskier for overfitting. When the neuron is more sensitive, its weight is generally bigger. Therefore, it is needed to adjust these parameters for a better prediction performance.

The ANN used in battery SOH estimation can be generally divided into the feed-forward neural network (FFNN) and recurrent neural network (RNN) [40]. Among these methods, the back-propagation FFNN (BPFFNN) is one of the most widely used methods. Forward signal propagation is used to pre-train the model, and backward error propagation is conducted to revise the weights in order to minimize the loss function in BPFFNN [79]. The BPFFNN was used to establish the observation equation of the battery, and then the UKF was used to estimate the remaining capacity of the battery [80]. In another study, the "importance sampling" was used to select the feature, and BPFFNN was adapted to estimate the RUL of the batteries [40]. The BPFFNN is used for evaluation.

B. Support vector machine

SVM is another widely used machine learning method for nonlinear systems. It is also a mature approach to SOH estimation [81]. This method maps data to a high-dimensional space and constructs an optimal separating hyperplane in this space. The key to the data transformation is a kernel function. Using the constraints of the Karush-Kuhn-Tucker condition, only a small part of the training data known as support vectors is retained and used to establish a classification or regression prediction model. The SVM model is defined as follows [81]:

$$f(\boldsymbol{x}) = \boldsymbol{w}\varphi(\boldsymbol{x}) + b, \boldsymbol{x} \in R^m, b \in R, \tag{2}$$

where $\boldsymbol{x}$ is the input matrix with m features, $\varphi(\boldsymbol{x})$ is a nonlinear mapping function, $\boldsymbol{w}$ and b represent the weight matrix and intercept of the hyperplane. Then, an insensitive loss function is introduced in order to solve nonlinear regression problems,

$$L(f(\boldsymbol{x}), y) = \begin{cases} 0, & |y - f(\boldsymbol{x})| \le \varepsilon \\ |y - f(\boldsymbol{x})| - \varepsilon, & |y - f(\boldsymbol{x})| > \varepsilon \end{cases}, \tag{3}$$

where ε is the allowable error between the real value and the estimated value. The problems of regression optimization using standard SVM can be summarized as follows [81]:

$$\min_{w} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right)$$
$$s.t. \begin{cases} y_i - \boldsymbol{w}\varphi(\boldsymbol{x}_i) - b \le \varepsilon + \xi_i \\ -y_i + \boldsymbol{w}\varphi(\boldsymbol{x}_i) + b \le \varepsilon + \xi_i^*, i = 1, 2, ..., l \\ \xi_i, \xi_i^* \ge 0 \end{cases}, \tag{4}$$

where ξ_i, ξ_i^* are the slack variables, and C is a non-negative penalty coefficient, l is for the number of training samples. As for SOH estimation, the capacity and resistance were estimated by the SVM, and then the SOH was estimated [82]. The IC peak was tracked by the SVM to achieve the on-board battery SOH estimation [27]. Ten HIs were extracted from the charging and discharging process, and the SVM was applied to map the regression model for the SOH estimation [41].

C. Relevance vector machine

The above two methods are non-probabilistic. However, the uncertainty from the measurements and the preprocessing needs to be quantified. The probabilistic predictions provide uncertainty and their correlation with data for the results [83]. RVM has the same equations as SVM, but provides probabilistic predictions based on the Bayesian framework [84]. The high sparsity of the RVM makes a large number of weights zero, which improves the computational efficiency compared to the SVM. The regression problem based on RVM can be expressed as follows [84]:

$$p\left(y_n \mid x_n, w, \sigma^2\right) = N\left(\mu\left(x_n; w\right), \sigma^2\right), \tag{5}$$

where x is the input matrix, $\mu\left(x_n; w\right)$ is the regression model without noise, w is the regression coefficient, y_n is the probabilistic output of a normal distribution with a mean of μ squared and variance of σ^2. Many studies have applied the RVM to the battery SOH estimation problem. The empirical model of the capacity degradation is fitted by the RVM for capacity estimation [85]. A wavelet denoising approach is used to reduce the uncertainty of the RVM, and the mean entropy helped select the embedding dimension, then the SOH estimation is obtained [86]. Given the presence of hyper-parameters in the SVM, and RVM needs to be optimized. The PSO is applied for the hyper-parameter optimization in this paper. The detailed description of PSO can be found in [87].

D. Gaussian process regression

The GPR, as another probabilistic prediction tool, has become popular in the field of battery SOH estimation due to its flexible, nonparametric, and probabilistic properties. It can model the behavior of any system through the appropriate combination of the gaussian process and achieve the prediction based on a Bayesian framework combined with prior knowledge [88]. It contains finite variable sets, and each set is jointly Gaussian distributed [89]. By extending the multivariate Gaussian distribution to infinite dimension, the Gaussian process f(x) can be obtained, which is constructed by means of the mean function m(x) and the covariance function $k(x_i, x_j)$ [90]:

$$m(x) = E\left(f(x)\right), \tag{6}$$

$$k_f\left(x_i, x_j\right) = E\left[\left(f(x_i) - m(x_i)\right)\left(f\left(x_j\right) - m\left(x_j\right)\right)\right]. \tag{7}$$

The covariance function, also known as the kernel function, is used to capture the similarity between different inputs, which is highly sensitive to the predicted performance of the GPR. The hyper-parameters, usually optimized by the maximum likelihood estimation of the edge probability [91]. According to the training data and test data, the prior distribution is constructed, and the predicted posterior distribution can be obtained by using Bayesian theory [92]:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTE.2020.3017090, IEEE Transactions on Transportation Electrification

$$p\left(y_m \mid x_t, y_t, x_m\right) = N\left(\mu_m, \sigma_m^2\right), \tag{8}$$

where x_t, y_t are the input and output of the training sample, x_m, y_m are the input and forecast output of the test sample, m is the predicted mean, σ_m^2 is the predicted variance. A data-driven diagnostic technique based on GPR for *in situ* capacity estimation is conducted, which estimates capacity over a short period of galvanostatic operation [92]. The features are extracted from the measured voltage curve [20] and calculated IC curve [30, 93], and GPR is adopted for regression tracking.

E. SOH estimation process

The flowchart of the SOH estimation is shown in Fig. 6. It contains four steps, including data acquisition, feature extraction and selection, model training, and SOH estimation. Some recorded data such as voltage and current are inputs for the measured variables-based HIs and calculated variables-based HIs extraction. Then, three different feature selection methods are adopted to select each subset. After that, different machine learning algorithms are used for model training, including self-model and mutual model. Finally, the SOH is estimated based on the different models, and the accuracy, robust, and computational efficiency are evaluated to demonstrate the estimation performance of each method.

VI. Results and Discussion

The SOH estimation results are presented and evaluated in this section. The model training strategies for SOH estimation can be divided into two categories. One uses historical data to train the model, and estimates SOH for the remaining cycle life. The other trains a regression model using one battery's data, and then estimates the other batteries' SOH using the same model. In this paper, these two training strategies are conducted to compare the different HI selection methods and different machine learning algorithms. HIs selected by each method are listed in the *APPENDIX*.

A. Machine learning algorithms

In this paper, publicly available toolboxes are used to implement the four machine learning algorithms. The toolbox used for ANN is the Neural Net Fitting app in MATLAB; the toolbox used for SVM is libsvm-3.23 from [94]; the toolbox used for RVM is from [95], and the toolbox used for GPR is gpml-v4.2 [96]. There are some hyper-parameters in the algorithms, which have significant impacts on the estimation effects. Therefore, the optimization of the hyper-parameters is necessary before the model training and prediction. The PSO algorithm is used for the hyper-parameter optimization of SVM and RVM. The number of evolutionary generations is 100, and the population is 20. The number of hidden layers is 20 for ANN, and the evolutionary generation is 100 for GPR. The RMSE and MAE are used to evaluate the estimation accuracy and the calculation time is used to assess the computational efficiency, respectively. The calculation time contains model training and estimation time except for parameter optimization time because the optimization process is different, and the parameter number is also different.

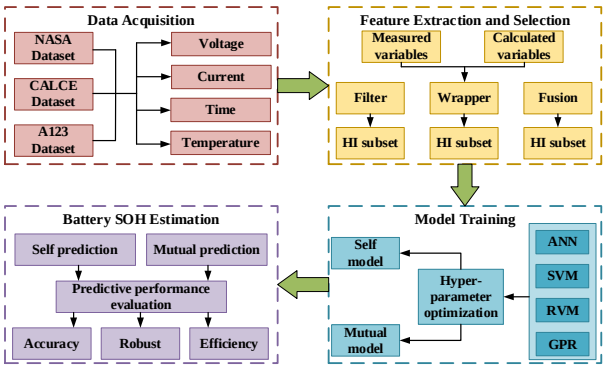


Fig. 6. Flowchart of the battery SOH estimation.

B. Evaluations for the first training strategy

In this section, the first training strategy is examined based on the NASA data set. The early experimental data is used for model training, and the remaining data is used for the SOH estimation. First, the comparisons of different machine learning algorithms are carried out where 50% of the experimental data is used for model training, and the rest is used for SOH estimation. The comparison results are listed in Table. II, where the time is the average value of the four batteries containing the training and estimation process. It shows different performances for these four algorithms in both accuracy and computational efficiency. The accuracy of RVM is generally worse than the other three, while the calculation time (mean of the four batteries) of ANN is much larger than the other three. It suggests that the GPR and SVM are performs better. Although the evolutionary generations are 100 for both GPR and SVM, the optimization speed of GPR (0.08 seconds) is faster than that of SVM (24.86 seconds). And the results of GPR are probabilistic with confidence intervals. Also, the computational efficiency is higher. Therefore, the GPR is more suitable for real applications.

Next, the comparison of different HI selection methods is conducted. First, the 50% data set is used to train the regression model, and the remaining data is used to estimate the SOH using the GPR algorithm. The estimated error results for each battery based on different HI selection methods are listed in Table. III. And the results of B0006 are depicted in Fig. 7, where (a) to (d) represents the results based on Initial HIs (InH), filter selected HIs (FiSH), wrapper selected HIs (WSH), and fusion selected HIs (FuSH), respectively. It shows that all the results are satisfying. However, it can obviously be seen that Fig. 7(b) has the widest confidence interval, and the MAE is also the largest. The estimated errors of Fig. 7(c) seem to be the smallest, while the confidence interval of Fig. 7(d) seems to be the narrowest. Table. III shows that the RMSE and MAE of WSH are overall smaller except for B0005 whose RMSE and MAE are larger than those of FuSH. This may be caused by the local optimal of the wrapper methods in the selection process, and that can be avoided by the fusion method. FuSH also provides satisfying estimation results, which are slightly bigger than wrapper but much smaller than filter. Moreover, the HIs selected by fusion method has the smallest dimension, which contributes to creating a SOH estimator with shorter computational time and less storage memory.

Next, the data set is reduced to 10% for training the regression models to see the robustness and the adaptability of each HIs subset. The error results are listed in Table. IV, and the estimation results of B0005 are shown in Fig. 8. Apparently, when less data is used for model training, all the results get worse. Among the four plots in Fig. 8, the first two results are divergent, and the last two results are still convergent, but all the confidence intervals are increased. It seems the estimation can only track the standard value in the first few cycles and lose its accuracy in the following cycles in Fig. 8(a) and Fig. 8(b), which means the HIs are not sufficient. On the contrary, the estimation in Fig. 8(c) and Fig. 8(d) can follow the standard value entire the cycle life. In addition, HIs selected by fusion-based method are much less than that of wrapper-based method, and the confidence interval is also narrower in the early cycles. The RMSEs are less than 1.5%, while the MAEs are less than 4.0% for B0005-B0007 and less than 7.5% for B0018, which means the results support the accuracy requirements on EVs.

C. Evaluations for the second training strategy

In this section, the second estimation strategy where the data of one battery is used for model training and the trained model is used for predictions on other batteries. First, the estimation results of the CALCE data set are assessed, and then the estimations of the A123 system data set are evaluated.

1) Comparisons and evaluations using CALCE data set

The data of CALCE is first used for comparisons based on the GPR algorithm, where the aging data of CS2_35 is used for model training. First, the estimation results of CS2_36 and CS2_37, whose constant discharge current rates are the same as that of CS2_35, are analyzed. The data dimensions of each HI selected method and the estimated errors are listed in Table. V. And the estimated results for CS2_36 are shown in Fig. 9. The results don't show the obvious difference, and all the HI sets could provide accurate estimations. The RMSEs are close, but the dimension of FuSH is the smallest, means the smallest computation cost is needed, but satisfying results could be obtained. There is a large error in Fig. 9(b) that causes the MAE larger than the others. The estimation of CS2_37 shows the same phenomenon. But the MAEs obtained by InH and FiSH are obviously larger than of WSH and FuSH. That means the wrapper and fusion select the more optimal combinations of HIs. From the results of CS2_36 and CS2_37, it can be concluded that when the test batteries go through the same aging protocol with the training one, the estimations are accurate and reliable whatever the subset. However, in this case, the computational efficiency needs to be considered for potential usage. It is shown that the fusion-based method removes the less correlated and redundant HIs, and create a more refined subset.

Next, the regression model trained by the experimental data of CS2_35 is used to estimate the CS2_33 and CS2_34, whose discharge current rates are different from CS2_35. These results reflect the estimation robustness and adaptability of each HI subsets. The error results are listed in Table. VI, and the estimation results of CS2_33 are shown in Fig. 10. The results are quite different from those from CS2_36 and CS2_37. When the batteries go through different discharge conditions, the estimation from the regression model gets worse. The estimation results in Fig. 10(a) and Fig. 10(b) are poor and their estimation performance is poor too. However, when wrapper and fusion based methods are used to find the optimal subset, the estimation results are still good enough for the real applications. As for CS2_33, the estimation obtained by FiSH has big deviations in the last few cycles, which causes a large MAE. MAE and RMSE are less than 3.0% and 2.0% for both WSH and FuSH. And the errors of WSH are slightly less than that of FuSH. But the HI number of FuSH is 5 while that of WSH is 9. As for the estimation results of CS2_34, it shows that the first two HI sets lose their estimation effect, while the last two HI sets could still give satisfactory estimations. However, the MAE and RMSE of FuSH are less than that of WSH. This indicates that some HIs selected by the wrapper method are not suitable for the estimation of this battery. The MAE and RMSE obtained by the FuSH are less than 1.1% and 3.7%, which are good enough for the BMS requirements. Here, the HIs needed are less than the other three, and the accuracy is satisfactory. Therefore, the HIs selected by fusion method could guarantee the accuracy and robustness with a low computational cost.

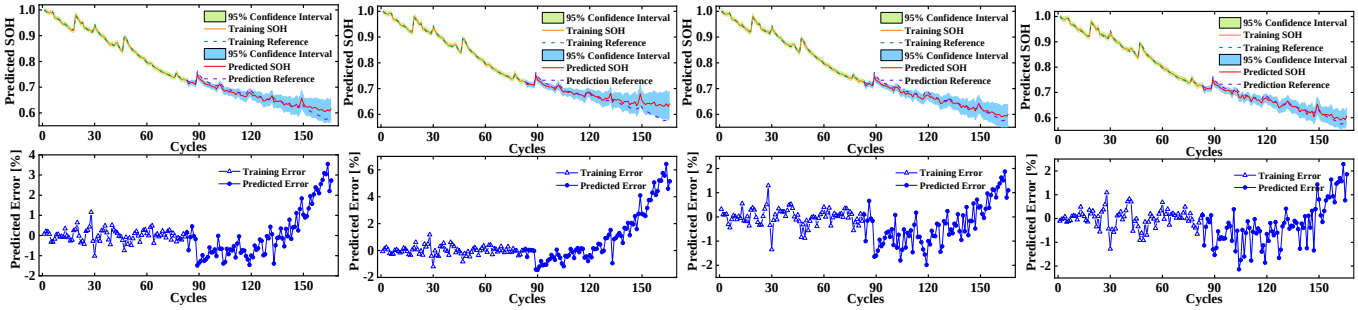
2) Comparisons and evaluations using A123 system data set

After the comparison based on CALCE data, the data set provided by Ref.[51] is used. The cathode material of the batteries is LPF, different from the above two types of batteries. The B01 is used for the regression model training, and the other batteries are used for estimation. The estimation results of B02 and B04 are drawn in Fig. 11 and Fig. 12, respectively. The dimension of each feature subset and the estimation errors are listed in Table. VII. It is worth noting that the B02 went through the same aging protocol with the trained battery. Therefore, the estimation results shown in Fig. 11 are perfect. All the feature sets can provide accurate and reliable estimations. The RMSE and MAE of each method don't have significant differences, all the RMSEs are less than 0.12%, and all the MAEs are less than 0.60%. However, the feature dimension of FuSH is much less than the other three, less than half of the InH, which would largely reduce the computational cost and storage memory of the HI extraction process, and still guarantee the estimation accuracy. The SOH estimations of the other three batteries are carried out, whose charging protocol is different from B01, but the discharging protocol is the same as B01. The RMSE and MAE of each battery cell are listed in Table. VII. It shows that if the charging protocol is different, the accuracy will get slightly worse. The estimation results of the B04 are shown in Fig. 12. It shows that Fig. 12(d) has the best estimation performance, and the errors are closer to 0 over the all cycles. The confidence interval of Fig. 12(b) is narrow, but doesn't cover the standard value, which means the reliability is low. It shows the WSH is not better than FuSH, the possible reason is that the selection process fell into local minima. On the other hand, there are 25 HIs in WSH. Some may not be good and suitable for the other batteries. That is the drawback of wrapper.

Table. II. Estimated errors (%) and calculate time (ms) of different algorithms.

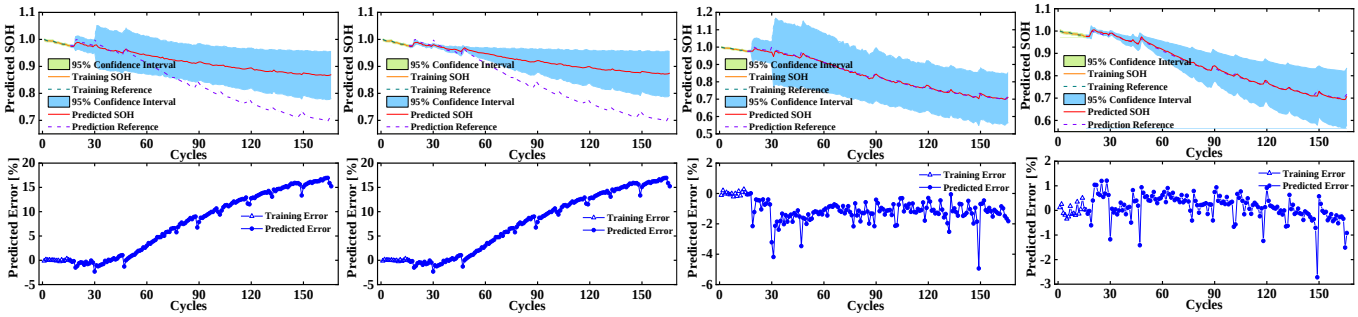
Battery	Index	ANN	SVM	RVM	GPR
B0005	RMSE	1.12	0.60	1.64	0.55
	MAE	3.15	1.23	3.01	2.07
B0006	RMSE	2.71	1.98	7.36	1.27
	MAE	6.57	6.04	13.13	3.55
B0007	RMSE	1.76	1.05	3.16	0.39
	MAE	3.65	2.25	4.69	1.03
B0018	RMSE	2.12	1.40	8.41	1.63
	MAE	6.24	2.91	15.32	5.14
Time	(ms)	256.31	13.43	12.21	10.48

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTE.2020.3017090, IEEE Transactions on Transportation Electrification



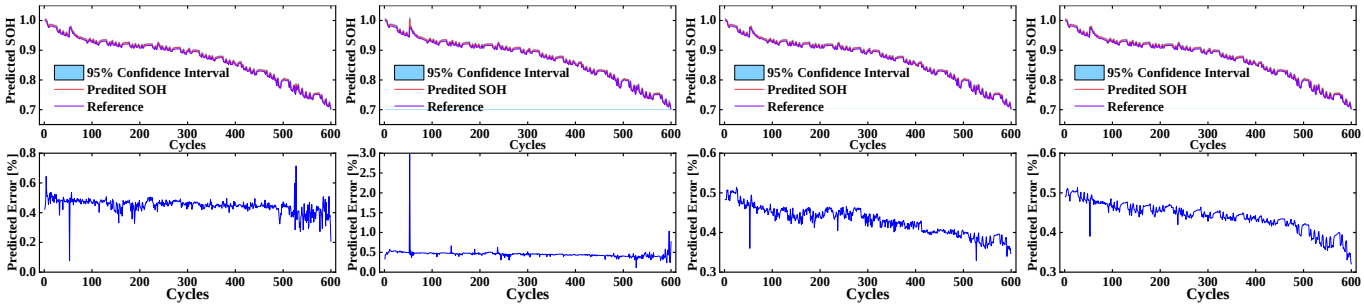
(a) Estimation results using InH (b) Estimation results using FiSH (c) Estimation results using WSH (d) Estimation results using FuSH

Fig. 7. SOH estimation results of B0006 using 50% data for model training.



(a) Estimation results using InH (b) Estimation results using FiSH (c) Estimation results using WSH (d) Estimation results using FuSH

Fig. 8. SOH estimation results of B0005 using 10% data for model training.



(a) Estimation results using InH (b) Estimation results using FiSH (c) Estimation results using WSH (d) Estimation results using FuSH

Fig. 9. SOH estimation results of CS2_36.

Table. III. SOH estimated errors based on different HIs with a 50% training set.

HIs	Index	B0005	B0006	B0007	B0018
InH	Dimension	30	30	30	30
	RMSE	0.55	1.27	0.39	1.63
	MAE	2.07	0.39	1.03	5.14
FiSH	Dimension	18	20	16	16
	RMSE	1.25	2.27	1.41	2.89
	MAE	2.91	6.43	3.33	6.82
WSH	Dimension	20	22	23	21
	RMSE	0.28	0.93	0.30	1.38
	MAE	1.29	1.98	1.00	4.07
FuSH	Dimension	13	11	9	13
	RMSE	0.24	0.96	0.35	1.01
	MAE	0.82	2.29	1.44	2.54

Table. IV. SOH estimated errors based on different HIs with a 10% training set.

HIs	Index	B0005	B0006	B0007	B0018
InH	Dimension	30	30	30	30
	RMSE	10.16	5.78	8.82	5.55
	MAE	16.94	12.02	14.81	8.39
FiSH	Dimension	18	20	16	16
	RMSE	10.57	6.17	1.32	11.49
	MAE	17.49	12.98	3.43	23.81
WSH	Dimension	22	23	21	23
	RMSE	0.75	1.22	1.04	1.49
	MAE	4.42	3.34	3.14	7.07
FuSH	Dimension	13	11	13	13
	RMSE	0.56	1.50	1.30	1.61
	MAE	2.72	3.69	3.39	7.44

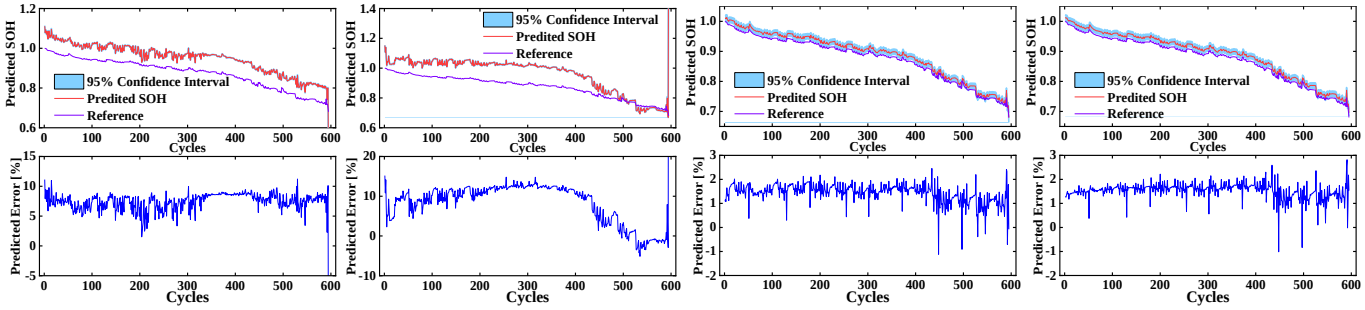
Table. V. SOH estimated errors for CS2_36 and CS2_37.

Feature set	Feature dimension	SOH predicted error for CS2_36		SOH predicted error for CS2_37	
		RMSE	MAE	RMSE	MAE
InH	19	0.45	0.71	0.28	1.29
FiSH	12	0.47	2.97	0.27	1.28
WSH	14	0.43	0.51	0.29	0.48
FuSH	9	0.44	0.51	0.27	0.44

Table. VI. SOH estimated errors for CS2_33 and CS2_34.

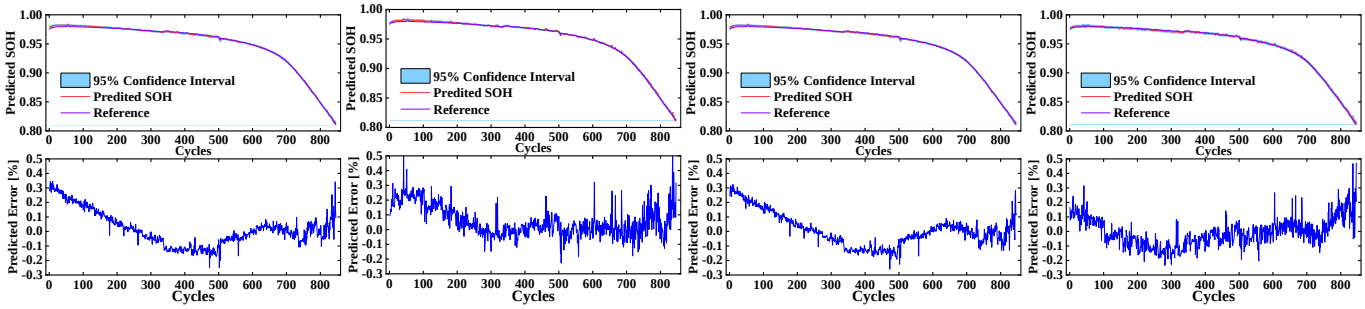
Feature set	Feature dimension	SOH predicted error for CS2_33		SOH predicted error for CS2_34	
		RMSE	MAE	RMSE	MAE (%)
InH	14	7.71	35.24	114.7	494.1
FiSH	7	11.79	169.36	729.0	4070
WSH	9	1.49	2.45	2.83	5.91
FuSH	5	1.56	2.82	1.03	3.70

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTE.2020.3017090, IEEE Transactions on Transportation Electrification



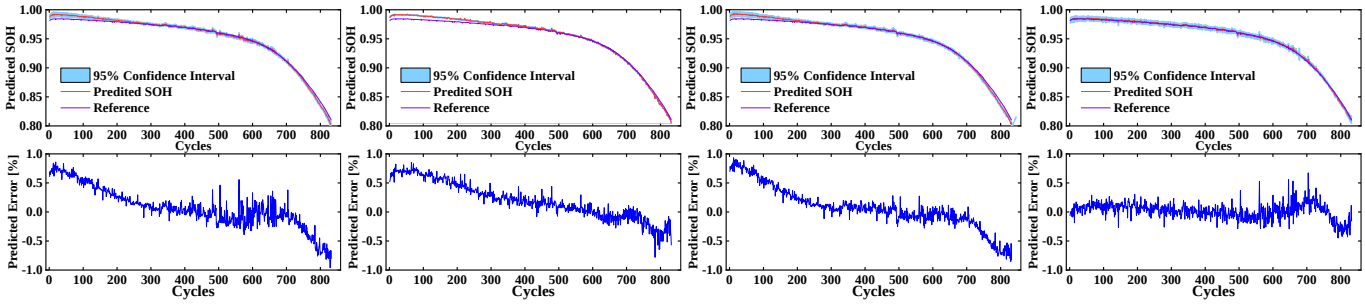
(a) Estimation results using InH (b) Estimation results using FiSH (c) Estimation results using WSH (d) Estimation results using FuSH

Fig. 10. SOH estimation results of CS2_33.



(a) Estimation results using InH (b) Estimation results using FiSH (c) Estimation results using WSH (d) Estimation results using FuSH

Fig. 11. SOH estimation results of B02.



(a) Estimation results using InH (b) Estimation results using FiSH (c) Estimation results using WSH (d) Estimation results using FuSH

Fig. 12. SOH estimation results of B04.

Compared to the estimation results of the CALCE data set, the A123 system data set has more accurate estimations. It may be because the features are larger than the former one, and the experiments are carried out more strictly, which makes HIs more effective. On the other hand, when the test batteries went through the same aging protocol with the training battery, both the two estimations are accurate. The CALCE uses the same charging protocol, while the A123 system uses the same discharging protocol. The results show the errors of the A123 system are less than the CACLE, indicating that discharging affects accuracy more than charging. However, in real world, it is mostly closer to the later one and sometimes confirmed to the former one. Thus, combining the two results, the fusion and wrapper-based methods provide more accurate estimations, and fusion methods could reduce the computational cost and storage memory by reducing the feature dimensions, and avoid falling into local optimality to some extent.

Finally, the computational efficiency of the four algorithms is further compared using the A123 system data set. The InH and the FuSH are used for the comparison, whose results are listed in Table. VIII and Table. IX, respectively. It shows all the algorithms have an overall more accurate estimation when the FuSH is used. Similar to the comparison using the first strategy, the RVM has worse accuracy. The GPR is overall more accurate than ANN and SVM whose accuracy is close to each other. According to the calculation time, the computational cost of ANN is larger than the others, and the GPR shows the fastest computational speed. Compared to the InH, computational efficiency is better when the FuSH is used. In addition, the calculation time of the optimization process for the hyper-parameters is listed in Table. X. The optimization process of GPR is much faster than that of SVM and RVM and the optimization processes are faster when the FuSH is used.

D. Discussion and recommendation

Feature selection in data preprocessing and machine learning algorithms are the two keys to the data-driven SOH estimation. The comprehensive evaluations of both are carried out based on three publicly available battery data sets using two common training strategies. In the first strategy, there are no significant differences of the estimation performance when enough data is used for model training. However, when the training data is reduced, the wrapper-based method and fusion-based method still provide satisfying estimations based on more efficient HIs.

Table. VII. SOH estimated errors for A123 batteries.

Feature set	D	B02		B03		B04		B05	
		R	M	R	M	R	M	R	M
InH	30	0.12	0.35	0.75	3.63	0.35	0.96	0.55	2.60
FiSH	17	0.12	0.59	0.75	3.28	0.37	0.85	0.95	1.86
WSH	25	0.09	0.32	0.49	1.08	0.36	0.93	0.56	4.25
FuSH	13	0.10	0.47	0.30	1.08	0.14	0.67	0.34	1.37

Table. VIII. Estimated errors and calculate time using InH.

Battery	Index	ANN	SVM	RVM	GPR
B02	RMSE	0.52	0.53	1.37	0.12
	MAE	0.85	1.07	3.53	0.35
B03	RMSE	1.29	1.04	18.76	0.75
	MAE	1.63	6.95	77.83	3.63
B04	RMSE	0.26	0.27	2.44	0.35
	MAE	0.62	0.64	4.18	0.96
B05	RMSE	0.33	0.43	1.78	0.55
	MAE	1.31	1.29	5.46	2.60
Time	(ms)	297.75	284.81	235.23	101.59

Table. IX. Estimated errors and calculate time using FuSH.

Battery	Index	ANN	SVM	RVM	GPR
B02	RMSE	0.54	0.52	0.65	0.10
	MAE	0.84	1.06	1.36	0.47
B03	RMSE	0.29	0.27	0.66	0.30
	MAE	1.05	0.97	1.56	1.08
B04	RMSE	0.35	0.34	0.54	0.14
	MAE	0.83	0.76	1.69	0.67
B05	RMSE	0.24	0.26	0.47	0.34
	MAE	1.15	0.93	3.01	1.37

Table. X. Calculation time (s) of the parameters' optimization.

	SVM	RVM	GPR
InH	1822.03	934.81	7.22
FuSH	1056.41	805.23	6.41

The fusion-based method has the potentials to avoid the local optimization, which may lead to the accuracy decrease of the wrapper-based method. As for machine learning methods, the computational cost of ANN is much larger than the others in this estimation strategy, and the accuracy of RVM is lower than the others. The accuracy of SVM and GPR are similar, but the results of GPR are probabilistic, and the uncertainty is provided. Therefore, the combination of the fusion-based HI selection method and GPR based estimation method has more advantages based on accuracy and computational efficiency. In the second estimation strategy, when the testing battery goes through the same aging protocol with the training battery, all the HI sets provide a good estimation. On the contrary, when the tested battery has a different aging protocol, the results show obvious differences. InH and FiSH get rapid deterioration even lose their estimation performance, while the WSH and FuSH still estimate the SOH accurately and reliably. The WSH set is more accurate than that of FuSH for some testing batteries but can also be worse for a few batteries. One possible reason for that is that the feature dimension of WSH is larger than that of FuSH, while some HIs are not efficient for testing batteries. Similar to that of the first estimation strategy, the computational cost of ANN is the largest, and the accuracy of RVM is the worst. In this strategy, the calculation time of GPR is much shorter than the others in the estimation process, and the optimization time for hyper-parameters is also the smallest. Another interesting result is that the results from the CALCE data set are worse than the A123 system data set as the tested batteries go through different test protocols. It suggests that the difference in the discharge process may have more significant impacts compared to the difference in the charging process. However, EVs always go through the different discharging processes. The results from CALCE batteries show a more significant difference between the four selection methods. Therefore, the WSH and the FuSH have more practical potentials for real usages. In terms of the evaluation of data preprocessing and machine learnings, the combination of the fusion-based selection method and the GPR learning method shows great performance. This method reduces the data dimension, extracts features faster, and retains estimation performance. In future potential applications, this work provides valuable improvements for the battery health prediction when the cloud technology is combined with onboard BMS. And the method does not rely on one specific HI but all available HIs that can be extracted onboard.

VII. Conclusion

Machine learning is widely used in battery state estimation. Data-driven battery SOH estimation draws a significant role in the BMS for health management and secondary utilization guidance and will be widely applied in future EVs under the big data era. By making the optimal combination of data preprocessing methods and machine learning algorithms for real applications, a comprehensive study is carried out in this paper. A new classification of the HI extracted methods is proposed, and the detailed process for each method is reviewed and summarized. The noise reduction performance of the calculated parameters based on four filter methods are assessed. Three feature selection methods, including a proposed fusion method, are used for the HI subsets selection before model training. Four widely used machine learning algorithms are adopted for the SOH estimation. Comparisons and evaluations for the feature selection methods and machine learning algorithms are carried out, and the accuracy and computational efficiency are assessed using three public data sets. The results show the combination of FuSH and GPR is the most recommended method for the potential practical usages.

Appendix

Table. I. The HIs in each selected subset using 50% data for model training. (I, Fi, W, and Fu represent the InH, FiSH, WSH, and FuSH, respectively)

HI	B0005				B0006			
	I	F	W	Fu	I	Fi	W	Fu
VIECT	√	√	√	√	√		√	
SCV	√	√	√	√	√		√	
UVP	√	√	√	√	√	√	√	
ICHV	√		√		√	√		√
VDEDT	√	√	√	√	√	√	√	√
IDV	√		√		√	√	√	
VD	√	√		√	√	√	√	√
LVP	√	√		√	√	√		
TEVI	√	√	√		√	√	√	
TEVD	√	√	√		√	√	√	√
TECD	√	√	√	√	√	√	√	√
TCCC	√	√	√	√	√	√	√	
TCVC	√	√		√	√	√	√	√
TETI	√				√	√		
MAT	√		√		√	√	√	√
MATL	√	√	√		√	√	√	√
MET	√				√		√	
TIET	√				√		√	
CDECT	√				√			
ICP	√	√	√	√	√	√	√	√
ICV	√		√		√		√	√
ICPL	√	√	√	√	√	√	√	

ICVL	√				√			
ICPA	√	√	√	√	√		√	
DVP	√		√		√	√		√
DVV	√	√	√	√	√	√	√	
DVPL	√	√	√		√			
DVVL	√		√		√	√	√	
DTP	√	√			√			
DTPL	√				√	√	√	
Total	30	18	20	13	30	20	22	11
HI	B0007				B0018			
	I	Fi	W	Fu	I	Fi	W	Fu
VIECT	√		√		√			
SCV	√	√	√		√	√	√	√
UVP	√	√	√		√	√	√	√
ICHV	√	√	√		√		√	
VDEDT	√	√	√	√	√	√	√	√
IDV	√		√		√	√	√	√
VD	√	√	√		√	√	√	√
LVP	√	√			√	√	√	√
TEVI	√	√	√	√	√	√	√	
TEVD	√	√	√	√	√	√		√
TECD	√		√		√		√	
TCCC	√	√	√		√	√	√	√
TCVC	√	√		√	√		√	
TETI	√				√		√	
MAT	√		√		√		√	
MATL	√	√	√	√	√	√	√	√
MET	√		√		√		√	
TIET	√				√			
CDECT	√		√		√		√	
ICP	√	√	√	√	√	√	√	√
ICV	√		√		√			
ICPL	√	√	√	√	√		√	
ICVL	√				√			
ICPA	√	√	√	√	√	√	√	√
DVP	√		√		√	√		√
DVV	√	√	√	√	√	√	√	
DVPL	√		√		√		√	
DVVL	√				√	√	√	
DTP	√		√		√			
DTPL	√	√		√	√	√	√	√
Total	30	16	21	13	30		23	12

Table. II. The HIs in each selected subset using 10% data for model training

HI	B0005				B0006			
	I	Fi	W	Fu	I	Fi	W	Fu
VIECT	√	√	√	√	√		√	
SCV	√	√	√	√	√		√	
UVP	√	√	√	√	√	√	√	
ICHV	√		√		√	√	√	√
VDEDT	√	√	√	√	√	√	√	√
IDV	√		√		√	√	√	
VD	√	√		√	√	√	√	√
LVP	√	√		√	√	√		
TEVI	√	√	√		√	√	√	
TEVD	√	√	√		√	√	√	√
TECD	√	√	√	√	√	√	√	√
TCCC	√	√	√	√	√	√	√	
TCVC	√	√		√	√	√	√	√
TETI	√				√	√		
MAT	√		√		√	√	√	√
MATL	√	√	√		√	√	√	√
MET	√				√		√	
TIET	√				√		√	
CDECT	√		√		√			
ICP	√	√	√	√	√	√	√	√
ICV	√		√		√		√	√
ICPL	√	√	√	√	√	√	√	
ICVL	√				√			
ICPA	√	√	√	√	√		√	
DVP	√		√		√	√		√
DVV	√	√	√	√	√	√	√	
DVPL	√	√	√		√			
DVVL	√		√		√	√	√	
DTP	√	√			√			
DTPL	√		√		√	√	√	
Total	30	18	22	13	30	20	23	11
HI	B0007				B0018			
	I	Fi	W	Fu	I	Fi	W	Fu
VIECT	√		√		√			
SCV	√	√	√		√	√	√	√
UVP	√	√	√	√	√	√	√	√
ICHV	√	√	√	√	√		√	
VDEDT	√	√	√	√	√	√	√	√
IDV	√		√		√	√	√	√
VD	√	√	√	√	√	√	√	√
LVP	√	√			√	√	√	√
TEVI	√	√	√	√	√	√	√	
TEVD	√	√	√	√	√	√		√
TECD	√		√		√		√	
TCCC	√	√	√		√	√	√	√
TCVC	√	√		√	√		√	
TETI	√				√		√	
MAT	√		√		√		√	
MATL	√	√	√	√	√	√	√	√
MET	√				√		√	
TIET	√				√			
CDECT	√		√		√		√	
ICP	√	√	√	√	√	√	√	√
ICV	√		√		√			
ICPL	√	√	√	√	√		√	
ICVL	√				√			
ICPA	√	√	√	√	√	√	√	√
DVP	√		√		√	√		√
DVV	√	√	√	√	√	√	√	
DVPL	√		√		√		√	
DVVL	√				√	√	√	
DTP	√		√		√			
DTPL	√	√		√	√	√	√	√
Total	30	16	21	13	30		23	12

Table. III. The HIs in each selected subset for model training of CS2_35.

HI	CS2_36 and CS2_37				CS2_33 and CS2_34			
	I	Fi	W	Fu	I	Fi	W	Fu
VIECT	√		√		√		√	
SCV	√		√		√		√	
UVP	√	√		√	√	√		√
ICHV	√				√			
VDEDT	√	√	√	√				
IDV	√	√	√					
VD	√	√	√	√				
LVP	√	√	√	√				
TEVI	√	√		√	√	√		
TEVD	√	√	√					
TECD	√		√		√		√	
TCCC	√	√	√	√	√	√	√	√
TCVC	√	√	√	√	√	√	√	√
TETI								
MAT								
MATL								
MET								
TIET								
CDECT	√		√		√		√	
ICP	√	√	√	√	√	√	√	√
ICV								
ICPL	√				√			
ICVL								
ICPA	√	√	√		√	√	√	
DVP								
DVV	√	√	√	√	√	√	√	√
DVPL								
DVVL	√				√			
DTP								
DTPL								
Total	19	12	14	9	14	7	9	5

Table. IV. The HIs in each selected subset for model training of B01.

HI	Initial	Filter	Wrapper	Fusion
VIECT	√	√	√	
SCV	√	√	√	
UVP	√	√	√	√
ICHV	√	√	√	√
VDEDT	√	√	√	√
IDV	√		√	
VD	√		√	
LVP	√	√	√	√
TEVI	√	√	√	
TEVD	√	√	√	√
TECD	√		√	
TCCC	√	√	√	√
TCVC	√			
TETI	√			
MAT	√	√	√	√
MATL	√	√	√	
MET	√		√	
TIET	√		√	
CDECT	√		√	
ICP	√	√	√	√
ICV	√		√	
ICPL	√	√	√	√
ICVL	√			
ICPA	√	√		√
DVP	√			
DVV	√	√	√	√
DVPL	√		√	
DVVL	√	√	√	√
DTP	√		√	
DTPL	√	√	√	√
Total	30	17	25	13

REFERENCES

[1] X. Hu, Y. Che, X. Lin, and Z. Deng, "Health Prognosis for Electric Vehicle Battery Packs: A Data-Driven Approach," *IEEE/ASME Transactions on Mechatronics*, pp. 1-1.2020.

[2] X. Hu, F. Feng, K. Liu, L. Zhang, J. Xie, and B. Liu, "State estimation for advanced battery management: Key challenges and future trends," *Renewable and Sustainable Energy Reviews,* vol. 114.2019.

[3] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y. H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon, and W. C. Chueh, "Closed-loop optimization of fast-charging protocols for batteries with machine learning," *Nature,* vol. 578, no. 7795, pp. 397-402, Feb.2020.

[4] B. Jiang, H. Dai, X. Wei, and T. Xu, "Joint estimation of lithium-ion battery state of charge and capacity within an adaptive variable multi-timescale framework considering current measurement offset," *Applied Energy,* vol. 253, November.2019.

[5] K. L. Tsui, N. Chen, Q. Zhou, Y. Hai, and W. Wang, "Prognostics and Health Management: A Review on Data Driven Approaches," *Mathematical Problems in Engineering,* vol. 2015, pp. 1-17.2015.

[6] M. S. H. Lipu, M. A. Hannan, A. Hussain, M. M. Hoque, P. J. Ker, M. H. M. Saad, and A. Ayob, "A review of state of health and remaining useful life estimation methods for lithium-ion battery in electric vehicles: Challenges and recommendations," *Journal of Cleaner Production,* vol. 205, pp. 115-133, December.2018.

[7] F. Yang, Y. Xie, Y. Deng, and C. Yuan, "Predictive modeling of battery degradation and greenhouse gas emissions from U.S. state-level electric vehicle operation," *Nature Communications,* vol. 9, no. 1.2018.

[8] X. Lin, and W. Lu, "A battery model that enables consideration of realistic anisotropic environment surrounding an active material particle and its application," *Journal of Power Sources,* vol. 357, pp. 220-229, July.2017.

[9] J. Kim, S. Lee, and B. H. Cho, "Complementary Cooperation Algorithm Based on DEKF Combined With Pattern Recognition for SOC/Capacity Estimation and SOH Prediction," *IEEE Transactions on Power Electronics,* vol. 27, no. 1, pp. 436-451, January.2012.

[10] A. Guha, and A. Patra, "Online estimation of the electrochemical impedance spectrum and remaining useful life of lithium-ion batteries," *IEEE Transactions on Instrumentation Measurement,* vol. 67, no. 8, pp. 1836-1849.2018.

[11] B. Saha, K. Goebel, S. Poll, and J. Christophersen, "Prognostics Methods for Battery Health Monitoring Using a Bayesian Framework," *IEEE Transactions on Instrumentation and Measurement,* vol. 58, no. 2, pp. 291-296.2009.

[12] J. Bi, T. Zhang, H. Yu, and Y. Kang, "State-of-health estimation of lithium-ion battery packs in electric vehicles based on genetic resampling particle filter," *Applied Energy,* vol. 182, pp. 558-568, November.2016.

[13] Q. Miao, L. Xie, H. Cui, W. Liang, and M. Pecht, "Remaining useful life prediction of lithium-ion battery with unscented particle filter technique," *Microelectronics Reliability,* vol. 53, no. 6, pp. 805-810, June.2013.

[14] K.-H. Tseng, J.-W. Liang, W. Chang, and S.-C. Huang, "Regression Models Using Fully Discharged Voltage and Internal Resistance for State of Health Estimation of Lithium-Ion Batteries," *Energies,* vol. 8, no. 4, pp. 2889-2907.2015.

[15] L. Zhang, Z. Mu, and C. Sun, "Remaining Useful Life Prediction for Lithium-Ion Batteries Based on Exponential Model and Particle Filter," *IEEE Access,* vol. 6, pp. 17729-17740, April.2018.

[16] Y. Li, K. Liu, A. M. Foley, A. Zülke, M. Berecibar, E. Nanini-Maury, J. Van Mierlo, and H. E. Hoster, "Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review," *Renewable and Sustainable Energy Reviews,* vol. 113.2019.

[17] X.-S. Si, W. Wang, C.-H. Hu, and D.-H. Zhou, "Remaining useful life estimation – A review on the statistical data driven approaches," *European Journal of Operational Research,* vol. 213, no. 1, pp. 1-14.2011.

[18] R. Xiong, L. Li, and J. Tian, "Towards a smarter battery management system: A critical review on battery state of health monitoring methods," *Journal of Power Sources,* vol. 405, pp. 18-29.2018.

[19] L. Lu, X. Han, J. Li, J. Hua, and M. Ouyang, "A review on the key issues for lithium-ion battery management in electric vehicles," *Journal of Power Sources,* vol. 226, pp. 272-288, March.2013.

[20] 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," *Journal of Power Sources,* vol. 384, pp. 387-395.2018.

[21] J. Yang, B. Xia, W. Huang, Y. Fu, and C. Mi, "Online state-of-health estimation for lithium-ion batteries using constant-voltage charging current analysis," *Applied Energy,* vol. 212, pp. 1589-1600, February.2018.

[22] C. Hu, G. Jain, P. Zhang, C. Schmidt, P. Gomadam, and T. Gorka, "Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery," *Applied Energy,* vol. 129, pp. 49-55.2014.

[23] L. Ren, L. Zhao, S. Hong, S. 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.

[24] Y. Deng, H. Ying, J. E, H. Zhu, K. Wei, J. Chen, F. Zhang, and G. Liao, "Feature parameter extraction and intelligent estimation of the State-of-Health of lithium-ion batteries," *Energy,* vol. 176, pp. 91-102.2019.

[25] Q. Zhao, X. Qin, H. Zhao, and W. Feng, "A novel prediction method based on the support vector regression for the remaining useful life of lithium-ion batteries," *Microelectronics Reliability,* vol. 85, pp. 99-108.2018.

[26] L. Datong, X. Wei, L. Haitao, and P. Yu, "An Integrated Probabilistic Approach to Lithium-Ion Battery Remaining Useful Life Estimation," *IEEE Transactions on Instrumentation and Measurement,* vol. 64, no. 3, pp. 660-670.2015.

[27] C. Weng, Y. Cui, J. Sun, and H. Peng, "On-board state of health monitoring of lithium-ion batteries using incremental capacity analysis with support vector regression," *Journal of Power Sources,* vol. 235, pp. 36-44.2013.

[28] M. Dubarry, and B. Y. Liaw, "Identify capacity fading mechanism in a commercial LiFePO4 cell," *Journal of Power Sources,* vol. 194, no. 1, pp. 541-549.2009.

[29] C. Weng, X. Feng, J. Sun, and H. Peng, "State-of-health monitoring of lithium-ion battery modules and packs via incremental capacity peak tracking," *Applied Energy,* vol. 180, pp. 360-368, October 2016.

[30] Z. Wang, J. Ma, and L. Zhang, "State-of-Health Estimation for Lithium-Ion Batteries Based on the Multi-Island Genetic Algorithm and the Gaussian Process Regression," *IEEE Access,* vol. 5, pp. 21286-21295, October.2017.

[31] D. Ansean, V. M. Garcia, M. Gonzalez, C. Blanco-Viejo, J. C. Viera, Y. F. Pulido, and L. Sanchez, "Lithium-Ion Battery Degradation Indicators

Via Incremental Capacity Analysis," *IEEE Transactions on Industry Applications,* vol. 55, no. 3, pp. 2992-3002.2019.

[32] Y. Gao, J. Jiang, C. Zhang, W. Zhang, and Y. Jiang, "Aging mechanisms under different state-of-charge ranges and the multi-indicators system of state-of-health for lithium-ion battery with Li(NiMnCo)O2 cathode," *Journal of Power Sources,* vol. 400, pp. 641-651.2018.

[33] I. Bloom, L. K. Walker, J. K. Basco, D. P. Abraham, J. P. Christophersen, and C. D. Ho, "Differential voltage analyses of high-power lithium-ion cells. 4. Cells containing NMC," *Journal of Power Sources,* vol. 195, no. 3, pp. 877-882.2010.

[34] T. Goh, M. Park, M. Seo, J. G. Kim, and S. W. Kim, "Capacity estimation algorithm with a second-order differential voltage curve for Li-ion batteries with NMC cathodes," *Energy,* vol. 135, pp. 257-268.2017.

[35] X. Feng, J. Li, M. Ouyang, L. Lu, J. Li, and X. He, "Using probability density function to evaluate the state of health of lithium-ion batteries," *Journal of Power Sources,* vol. 232, pp. 209-218.2013.

[36] T. Shibagaki, Y. Merla, and G. J. Offer, "Tracking degradation in lithium iron phosphate batteries using differential thermal voltammetry," *Journal of Power Sources,* vol. 374, pp. 188-195.2018.

[37] Y. Li, M. Abdel-Monem, R. Gopalakrishnan, M. Berecibar, E. Nanini-Maury, N. Omar, P. van den Bossche, and J. Van Mierlo, "A quick on-line state of health estimation method for Li-ion battery with incremental capacity curves processed by Gaussian filter," *Journal of Power Sources,* vol. 373, pp. 40-53, January.2018.

[38] X. Li, Z. Wang, L. Zhang, C. Zou, and D. D. Dorrell, "State-of-health estimation for Li-ion batteries by combing the incremental capacity analysis method with grey relational analysis," *Journal of Power Sources,* vol. 410-411, pp. 106-114.2019.

[39] X. Li, Z. Wang, and J. Yan, "Prognostic health condition for lithium battery using the partial incremental capacity and Gaussian process regressio," *Journal of Power Sources,* vol. 421, pp. 56-67, March.2019.

[40] J. Wu, C. Zhang, and Z. Chen, "An online method for lithium-ion battery remaining useful life estimation using importance sampling and neural networks," *Applied Energy,* vol. 173, pp. 134-140, July.2016.

[41] D. Liu, Y. Song, L. Li, H. Liao, and Y. Peng, "On-line life cycle health assessment for lithium-ion battery in electric vehicles," *Journal of Cleaner Production,* vol. 199, pp. 1050-1065.2018.

[42] Y. Zhang, and B. Guo, "Online Capacity Estimation of Lithium-Ion Batteries Based on Novel Feature Extraction and Adaptive Multi-Kernel Relevance Vector Machine," *Energies,* vol. 8, no. 11, pp. 12439-12457.2015.

[43] X. Tang, C. Zou, K. Yao, G. Chen, B. Liu, Z. He, and F. Gao, "A fast estimation algorithm for lithium-ion battery state of health," *Journal of Power Sources,* vol. 396, pp. 453-458.2018.

[44] B. Saha, and K. Goebel. " "Battery Data Set", NASA Ames Prognostics Data Repository , NASA Ames Research Center, Moffett Field, CA," http://ti.arc.nasa.gov/project/prognostic-data-repository.

[45] 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," Journal of Power Sources, vol. 196, no. 23, pp. 10314-10321.2011.

[46] 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," Microelectronics Reliability, vol. 53, no. 6, pp. 811-820, June.2013.

[47] "CALCE Battery Data," https://web.calce.umd.edu/batteries/data.htm.

[48] J. Liu, and Z. 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.

[49] Y. Zhou, M. Huang, Y. Chen, and Y. Tao, "A novel health indicator for on-line lithium-ion batteries remaining useful life prediction," *Journal of Power Sources,* vol. 321, pp. 1-10.2016.

[50] R. Razavi-Far, M. Farajzadeh-Zanjani, S. Chakrabarti, and M. Saif, "Data-driven prognostic techniques for estimation of the remaining useful life of Lithium-ion batteries," *IEEE International Conference on Prognostics and Health Management (ICPHM)*.2016.

[51] K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh, and R. D. Braatz, "Data-driven prediction of battery cycle life before capacity degradation," *Nature Energy,* vol. 4, no. 5, pp. 383-391, March.2019.

[52] D. Zhou, L. Xue, Y. Song, and J. Chen, "On-Line Remaining Useful Life Prediction of Lithium-Ion Batteries Based on the Optimized Gray Model GM(1,1)," *Batteries,* vol. 3, no. 4.2017.

[53] P. Tagade, K. S. Hariharan, S. Ramachandran, A. Khandelwal, A. Naha, S. M. Kolake, and S. H. Han, "Deep Gaussian process regression for lithium-ion battery health prognosis and degradation mode diagnosis," *Journal of Power Sources,* vol. 445.2020.

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

[55] R. R. Richardson, M. A. Osborne, and D. A. Howey, "Gaussian process regression for forecasting battery state of health," *Journal of Power Sources,* vol. 357, pp. 209-219, July.2017.

[56] M. Dubarry, B. Y. Liaw, M.-S. Chen, S.-S. Chyan, K.-C. Han, W.-T. Sie, and S.-H. Wu, "Identifying battery aging mechanisms in large format Li ion cells," *Journal of Power Sources,* vol. 196, no. 7, pp. 3420-3425.2011.

[57] C. Weng, J. Sun, and H. Peng, "A unified open-circuit-voltage model of lithium-ion batteries for state-of-charge estimation and state-of-health monitoring," *Journal of Power Sources,* vol. 258, pp. 228-237.2014.

[58] M. Berecibar, M. Garmendia, I. Gandiaga, J. Crego, and I. Villarreal, "State of health estimation algorithm of LiFePO4 battery packs based on differential voltage curves for battery management system application," *Energy,* vol. 103, pp. 784-796, May.2016.

[59] X. Li, X. Shu, J. Shen, R. Xiao, W. 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.

[60] L. Wang, C. Pan, L. Liu, Y. Cheng, and X. Zhao, "On-board state of health estimation of LiFePO4 battery pack through differential voltage analysis," *Applied Energy,* vol. 168, pp. 465-472, April.2016.

[61] Y. Merla, B. Wu, V. Yufit, N. P. Brandon, R. F. Martinez-Botas, and G. J. Offer, "Novel application of differential thermal voltammetry as an in-depth state-of-health diagnosis method for lithium-ion batteries," *Journal of Power Sources,* vol. 307, pp. 308-319, March.2016.

[62] X. Hu, J. Jiang, D. Cao, and B. Egardt, "Battery Health Prognosis for Electric Vehicles Using Sample Entropy and Sparse Bayesian Predictive Modeling," *IEEE Transactions on Industrial Electronics*, pp. 2645-2656.2015.

[63] H. Pan, Z. Lü, H. Wang, H. Wei, and L. Chen, "Novel battery state-of-health online estimation method using multiple health indicators and an extreme learning machine," *Energy,* vol. 160, pp. 466-477.2018.

[64] Z. Song, X. Wu, X. Li, J. Sun, H. F. Hofmann, and J. Hou, "Current Profile Optimization for Combined State of Charge and State of Health Estimation of Lithium Ion Battery Based on Cramer–Rao Bound Analysis," *IEEE Transactions on Power Electronics,* vol. 34, no. 7, pp. 7067-7078.2019.

[65] Z. Song, J. Hou, H. F. Hofmann, X. Lin, and J. Sun, "Parameter Identification and Maximum Power Estimation of Battery/Supercapacitor Hybrid Energy Storage System Based on Cramer–Rao Bound Analysis," *IEEE Transactions on Power Electronics,* vol. 34, no. 5, pp. 4831-4843.2019.

[66] Z. Wang, S. Zeng, J. Guo, and T. Qin, "State of health estimation of lithium-ion batteries based on the constant voltage charging curve," *Energy,* vol. 167, pp. 661-669.2019.

[67] S. Khaleghi, Y. Firouz, J. Van Mierlo, and P. Van den Bossche, "Developing a real-time data-driven battery health diagnosis method, using time and frequency domain condition indicators," *Applied Energy,* vol. 255.2019.

[68] A. Barai, R. Tangirala, K. Uddin, J. Chevalier, Y. Guo, A. McGordon, and P. Jennings, "The effect of external compressive loads on the cycle lifetime of lithium-ion pouch cells," *Journal of Energy Storage,* vol. 13, pp. 211-219.2017.

[69] H. Liu, and H. Motoda, *Computational Methods of Feature Selection*, Boca Raton: Chapman & Hall/CRC, 2008.

[70] M. Sebban, and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recognition,* vol. 35, no. 4, pp. 835-846, Apr.2002.

[71] W. Daelemans, V. e. Hoste, F. D. Meulder, and B. Naudts, "Combined Optimization of Feature Selection and Algorithm Parameters in Machine Learning of Language," *European Conference on Machine Learning. Springer, Berlin, Heidelberg*, pp. 84-95.2003.

[72] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics,* vol. 23, no. 19, pp. 2507-2517, Oct 1.2007.

[73] S. Beniwal, and J. Arora, "Classification and feature selection techniques in data mining," *International Journal of Engineering Research & Technology,* vol. 1, no. 6, pp. 1-6.2012.

[74] R. Kohavi, and G. H. John, "Wrappers for feature subset selection,"

Artificial Intelligence, vol. 97, no. 1-2, pp. 273-324, Dec.1997.

[75] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," *Data classification: Algorithms and applications*, no. 37.2014.

[76] Z. X. Zhu, Y. S. Ong, and M. Dash, "Wrapper-filter feature selection algorithm using a memetic framework," *Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics,* vol. 37, no. 1, pp. 70-76, Feb.2007.

[77] M. H. Hassoun, *Fundamentals of Artificial Neural Networks*: MIT press, 1995.

[78] X. Hu, L. Xu, X. Lin, and M. Pecht, "Battery Lifetime Prognostics," *Joule,* vol. 4, no. 2, pp. 310-346.2020.

[79] M. W. Gardner, and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment,* vol. 32, no. 14-15, pp. 2627-2636.1998.

[80] W. He, N. Williard, C. Chen, and M. Pecht, "State of charge estimation for Li-ion batteries using neural network modeling and unscented Kalman filter-based error cancellation," *International Journal of Electrical Power & Energy Systems,* vol. 62, pp. 783-791.2014.

[81] X. Li, C. Yuan, and Z. Wang, "State of health estimation for Li-ion battery via partial incremental capacity analysis based on support vector regression," *Energy,* vol. 203.2020.

[82] V. Klass, M. Behm, and G. Lindbergh, "A support vector machine-based state-of-health estimation method for lithium-ion batteries under electric vehicle operation," *Journal of Power Sources,* vol. 270, pp. 262-272.2014.

[83] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature,* vol. 521, no. 7553, pp. 452-9, May 28.2015.

[84] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of machine learning research,* vol. 1, pp. 211-244.2001.

[85] D. Wang, Q. Miao, and M. Pecht, "Prognostics of lithium-ion batteries based on relevance vectors and a conditional three-parameter capacity degradation model," *Journal of Power Sources,* vol. 239, pp. 253-264.2013.

[86] H. Li, D. Pan, and C. L. P. Chen, "Intelligent Prognostics for Battery Health Monitoring Using the Mean Entropy and Relevance Vector Machine," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 44, no. 7, pp. 851-862.2014.

[87] J. Kennedy, and R. Eberhart, "Particle swarm optimization." pp. 1942-1948.

[88] M. Liu, G. Chowdhary, B. C. Da Silva, S.-Y. Liu, and J. P. How, "Gaussian processes for learning and control: A tutorial with examples," *IEEE Control Systems Magazine,* vol. 38, no. 5, pp. 53-86.2018.

[89] C. E. Rasmussen, *Gaussian processes in machine learning*: Springer, 2003.

[90] X. Li, C. Yuan, X. Li, and Z. Wang, "State of health estimation for Li-Ion battery using incremental capacity analysis and Gaussian process regression," *Energy,* vol. 190.2020.

[91] K. V. Mardia, and R. J. Marshall, "Maximum likelihood estimation of models for residual covariance in spatial regression," vol. 71, no. 1, pp. 135-146, April.1984.

[92] R. R. Richardson, C. R. Birkl, M. A. Osborne, and D. A. Howey, "Gaussian Process Regression for In Situ Capacity Estimation of Lithium-Ion Batteries," *IEEE Transactions on Industrial Informatics,* vol. 15, no. 1, pp. 127-138, January.2019.

[93] X. Li, C. Yuan, and Z. Wang, "Multi-time-scale framework for prognostic health condition of lithium battery using modified Gaussian process regression and nonlinear regression," *Journal of Power Sources,* vol. 467.2020.

[94] C.-C. Chang, and C.-J. Lin. "LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology," http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[95] K. Qiu. "Prediction based on Relevance Vector Machine (RVM) using SB2_Release_200 toolbox," http://www.miketipping.com/sparsebayes.htm.

[96] C. E. Rasmussen, and H. Nickisch. "GAUSSIAN PROCESS REGRESSION AND CLASSIFICATION Toolbox version 4.2 for GNU Octave 3.2.x and Matlab 7.x and higher," http://www.gaussianprocess.org/gpml/code/oldcode.html.

Xiaosong Hu (SM'16) received the Ph.D. degree in automotive engineering from the Beijing Institute of Technology, Beijing, China, in 2012. He did scientific research and completed the Ph.D. dissertation in Automotive Research Center at the University of Michigan, Ann Arbor, MI, USA, between 2010 and 2012.

He is currently a Professor with the State Key Laboratory of Mechanical Transmissions and at the Department of Automotive Engineering, Chongqing University, Chongqing, China. He was a Postdoctoral Researcher with the Department of Civil and Environmental Engineering, University of California, Berkeley, CA, USA, between 2014 and 2015, as well as at the Swedish Hybrid Vehicle Center and the Department of Signals and Systems at Chalmers University of Technology, Gothenburg, Sweden, between 2012 and 2014. He was also a Visiting Postdoctoral Researcher with the Institute for Dynamic Systems and Control at Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, in 2014. His research interests include modeling and control of alternative powertrains and energy storage systems.

Dr. Hu was the recipient of several prestigious awards/honors, including Emerging Sustainability Leaders Award in 2016, EU Marie Currie Fellowship in 2015, ASME DSCD Energy Systems Best Paper Award in 2015, and Beijing Best Ph.D. Dissertation Award in 2013.

Yunhong Che received the B.E. degree in Automotive Engineering from Chongqing University, Chongqing, China, in 2019. He is currently pursuing the M.S. degree in state estimation and life prediction of Li-ion batteries at the Vehicle Power System Lab, Department of Automotive Engineering, Chongqing University.

Xianke Lin (M'19) received his Ph.D. in Mechanical Engineering from the University of Michigan, Ann Arbor, in 2014. He previously finished his B.S. from Zhejiang University, China, in 2009.

He has extensive industrial experience at Fiat Chrysler Automobiles, Mercedes-Benz R&D North America, and General Motor of Canada. Currently, he is an Assistant Professor in the Department of Automotive, Mechanical and Manufacturing Engineering at the University of Ontario Institute of Technology, Oshawa, Canada. His research activities have concentrated on hybrid powertrain design and control strategy optimization, Multiscale/Multiphysics modeling and optimization of energy storage systems.

Simona Onori (SM'15) received the Laurea degree(summa cum laude) computer science engineering from the University of Rome Tor Vergata, Rome, Italy, in 2003, the M.S. degree in electrical and computer engineering from the University of New Mexico, Albuquerque, NM, USA, in 2004, and the Ph.D. degree in control engineering from the University of Rome Tor Vergata, in 2007. Since 2017, she has been an Assistant Professor with the Energy Resources Engineering Department, Stanford University, Stanford, CA, USA. Dr. Onori was a recipient of the SAE Ralph R. Teetor Educational Award, the 2017 NSF CAREER Award, the 2017 Clemson University College of Engineering and Science Dean's Faculty Fellows Award, the 2017 Clemson University Esin Gulari Leadership and Service Award, the 2016 Energy Leadership Award in the category Emerging Leader (for the Carolinas), and the 2015 Innovision Award (SC, USA). Award in the category Emerging Leader (for the Carolinas), the 2015 Innovision Award (South Carolina), and 2012 Lumley Interdisciplinary Research, 2011 Outstanding Technology Team Award, TechColumbus. She was Chair of the IEEE CSS Technical Committee of Automotive Controls from 2015-2017, she is vice-chair of the IFAC TC on Automotive Control TC7.1 since 2015, and associate editor of the SAE International Journal of Alternative Powertrains since 2012 and IEEE Intelligent Vehicle Transactions since 2019 . She has co-authored a book, 2 book chapters and more than 120 peer-reviewed papers on hybrid electric vehicles simulation, optimization and control, estimation and control of electrochemical processes and catalytic conversion devices, such as batteries and after-treatment devices.