Prediction and evaluation of health state for power battery based on Ridge linear regression model

Bixiong Huang1, Haiyu Liao1, Yiquan Wang1, Xintian Liu1 and Xiao Yan1

1School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai, China

Abstract

The state of health (SOH) of power battery reflects the difference between the current performance of the battery and the time it left the factory. Accurate prediction of it is the key to improving battery cycle efficiency. This paper studies the application of data-driven algorithms in power battery health estimation. Firstly, Using the data of actual operating vehicles which are monitoring in the data platform as the research objects. The charging event segmentation algorithm is designed for the full amount of data, and the K-means clustering model is used to extract slow charging events. Secondly, feature engineering is performed on the data, including the use of Pearson and Spearman coefficients analysis for numerical features, the use of one-hot encoding for category features to determine the final input features of SOH model. Eventually, using the Ridge linear regression model to predict the health status of the power battery. The research shows that the MAE is less than 5%, which meets the needs of practical use. In addition, this paper comparing Ridge with three other models named Linear Regression, Lasso, and Elastic Net. The result showed that the linear regression model with L2 regularization is more applicable in low-dimensional feature application scenarios without cell data in prediction of SOH.

Keywords

Data-driven, power battery health status, feature engineering, regularization

The state of health (SOH) of power battery is an important parameter of the battery management system (BMS), which can reflect the age of the battery, and its value usually
directly determines whether the device needs to replace the battery module or pack. In current research, SOH is generally defined as the ratio of the maximum capacity that the battery can currently charge to the rated capacity when it left the factory. The formula is:

$$
SOH = \frac{C_{\text{lef}}}{Q_N} \times 100\% \quad (1)
$$

In the formula, $C_{\text{lef}}$ represents the current total capacity of the battery; $Q_N$ represents the rated capacity when left the factory. The process of predicting battery health status is very complicated. Even batteries of the same brand and same batch model have different health status when they left the factory. Generally, the health status of the battery can be accurately obtained through experiments in a laboratory environment. In actual engineering applications, due to the influence of comprehensive factors such as working conditions, temperature, and usage habits, the predicted value can only provide a rough reference to the technicians.

At present, the commonly used traditional model-based power battery state estimation models are: equivalent circuit model, electrochemical model and finite element model. However, the internal chemical reactions of lithium batteries cannot be directly observed or measured, and there is a certain amount of noise. Therefore, the above model-based method has some deviations, and it is impossible to establish a model that can be reasonably and accurately applied to actual conditions.

In response to this problem, this paper studies the application of data-driven algorithms in estimating the state of health of power batteries. Comparing with the traditional model-based modeling method, this method considers the actual usage scenarios and habits of users, such as the average charging current and the temperature which have a greater impact on the actual SOH, and only needs to analyze a large amount of historical usage data in the data platform, and the estimation accuracy of the algorithm can be verified through the corresponding experimental data.

**Data preprocessing**

The data used in this article is provided by the Shanghai New Energy Vehicle Public Data Collection and Monitoring Research Center. The data comes from the Shanghai New Energy Vehicle Operation Data collected by the data center, all of which have been desensitized. The data includes the operating data of 25 pure electric vehicles, the data collection interval is 10 s, and the time span is 180 days, including vehicle operating data, battery extreme value data, and static information. Part of the sample data is shown in Table 1.

**Data segmentation**

Since the data used in this paper are all the original driving data and do not contain the label of the model: SOH, the most important thing is estimating the real SOH of vehicles from the raw data. As the definition of SOH which is showed in formula (1), the maximum rechargeable capacity in the current state needs to be calculated, so each charging event needs to be filtered from all the raw data at first. In this paper, the charging
Table 1. Sample data of new energy vehicles.

| datetime  | sequence | status | C丝毫不懂 | rotating speed | torque | temperature | controller voltage | controller current |
|-----------|----------|--------|-----------|---------------|--------|-------------|-------------------|-------------------|
| 2018/1/6 9:15 | 1        | 3      | 11        | 0             | 0      | 10          | 719               | 0.7               |
| 2018/1/6 9:15 | 2        | 3      | 13        | 0             | 0      | 49          | 715               | 0.7               |
| 2018/1/6 9:15 | 1        | 1      | 12        | 1380          | 0      | 10          | 714               | 0.6               |
| 2018/1/6 9:15 | 2        | 1      | 13        | 1178          | 0.1    | 49          | 710               | 0.6               |
| 2018/1/6 9:16 | 1        | 1      | 12        | 2055          | 26.9   | 10          | 700               | 11                |
| 2018/1/6 9:16 | 2        | 1      | 14        | 1745          | 31.7   | 49          | 696               | 11                |
| 2018/1/6 9:16 | 1        | 1      | 12        | 1186          | 40.7   | 10          | 704               | 10                |
| 2018/1/6 9:16 | 2        | 1      | 13        | 1018          | 47.8   | 49          | 701               | 10                |
| 2018/1/6 9:17 | 1        | 1      | 12        | 743           | 4.8    | 10          | 710               | 4                 |
| 2018/1/6 9:17 | 2        | 1      | 13        | 614           | 1.5    | 48          | 706               | 4                 |
| 2018/1/6 9:17 | 1        | 3      | 12        | 1547          | 0.5    | 10          | 710               | 0.9               |
| 2018/1/6 9:17 | 2        | 3      | 13        | 1317          | 0.6    | 48          | 708               | 0.9               |
| 2018/1/6 9:18 | 1        | 1      | 12        | 47            | 6.1    | 10          | 713               | 0.8               |
| 2018/1/6 9:18 | 2        | 1      | 13        | 41            | 7.2    | 48          | 711               | 0.8               |
| 2018/1/6 9:19 | 1        | 3      | 12        | 0             | 0      | 10          | 714               | 0.4               |
| 2018/1/6 9:19 | 2        | 3      | 13        | 0             | 0      | 48          | 711               | 0.4               |
| 2018/1/6 9:19 | 1        | 1      | 12        | 893           | 18     | 10          | 711               | 2.6               |
| 2018/1/6 9:19 | 2        | 1      | 13        | 761           | 18.6   | 48          | 709               | 2.1               |
| 2018/1/6 9:20 | 1        | 1      | 13        | 2747          | 29.2   | 10          | 697               | 15.1              |
| 2018/1/6 9:20 | 2        | 1      | 14        | 2343          | 34.3   | 48          | 692               | 15.1              |
| 2018/1/6 9:20 | 1        | 3      | 12        | 2463          | 0      | 10          | 711               | 0.5               |
| 2018/1/6 9:20 | 2        | 3      | 14        | 2094          | 0      | 48          | 707               | 0.5               |
event segmentation algorithm considers the following two possible abnormal situations: The first is that when the vehicle is charging in an underground parking garage or other places with poor network signal, the data may not be uploaded in time, resulting in the loss of the next piece of data in this charging process; The second is that the charging pile is abnormal and the output current is unstable. When the maximum or minimum current is collected by the SOC algorithm in BMS, it will generate a program bug. As a result, the SOC of the data of the latter frame during charging is smaller than that of the previous frame, and the normal SOC should be an increasing process during charging. Considering the above possible situations in reality, in order to reduce the misclassification of charging events, the following segmentation algorithm is designed which can ensure the accuracy of subsequent labels:

Case1: When the time interval between the previous piece of data and the next piece of data is ≤10 min and the SOC of previous piece ≤ the SOC of next piece, it is considered the same charging event;

Case2: When the time interval between the previous data and the next data is > 10 min and <30 min and the SOC of previous data < the SOC of next data, it is considered to be the same charging event;

Case3: Otherwise, it is considered as the next charging event.

According to formula (1), $Q_N$ is a fixed, known quantity. $C_{lef}$ can be converted to estimate the maximum charging capacity of the vehicle from actual operating data, so it can be changed to formula (2):

$$
\text{SOH} = \frac{\int_{t_{start}}^{t_{end}} t \, dt}{C_N}.
$$

According to the above charging event segmentation algorithm, 25 running vehicles were divided into a total of 6029 charging events, with an average of 241 times of charging per vehicle in half a year.

**K-means distinguishes fast and slow charging**

In the segmented charging events, selecting the charging events that can more accurately evaluate the SOH, and the conditions are as follows:

1. If the time of a charging event is too short, that is, the value of $SOC_{end} - SOC_{start}$ is too small, it will cause the numerator of formula (2) to tend to infinity, and eventually cause the estimated SOH to be abnormal. The first condition in this paper is that the value of $SOC_{end} - SOC_{start}$ is >30%.

2. There are two problems in the fast charging of vehicles: Firstly, for safety and prolonging the service life of lithium batteries, the charging pile will add a lower current operation during fast charging, that is, in the low SOC range is high current, and in the high SOC range, it will switch from constant current charging to constant voltage charging, and gradually switch to low current to protect the battery. If data loss occurs during this period, it will cause the calculated
ampere-hour integral error, and then cause SOH error; secondly, the current is too large during fast charging, part of the current is actually converted into heat loss, and the capacity calculated by the Ampere-hour integral is too high, resulting in the estimated SOH is too large.\textsuperscript{16}

Therefore, in all charging events, only the slow charging events are reserved for SOH estimation, which is to prepare labels for the training of the model. The condition for distinguishing between fast charging and slow charging is the current.\textsuperscript{17} As shown in Figure 1, it can be seen that there are two obvious peaks in the current distribution, namely, fast charging and slow charging. Since the current range of fast and slow charging cannot be artificially given, so it belongs to the unlabeled situation and is suitable for unsupervised learning models. In this paper, all charging current data are inputted into the K-means clustering model for training. Since there are only two types of fast charging and slow charging, $K = 2$ is used to classify the fast and slow charging events. The clustering result showed that the cluster center of slow charging events is $-13.7$ A, and the cluster center of fast charging events is $-76.4$ A, which is consistent with the result shown in the distribution diagram in Figure 1.

**Screening of charging events**

According to the algorithm in 1.1 and 1.2, 1468 events which meeting the conditions were screened out from all 6029 charging events, and the SOH of 25 vehicles was

![Figure 1](image.png)

*Figure 1.* Charging current distribution diagram.
estimated by formula (2) which is already different from the SOH calculated on board because of the SOH in the vehicle-mounted BMS on the market is basically obtained by multiplying the coefficient obtained by the cyclic test experiment by the rated capacity, or obtained by the look-up table method which does not consider the different actual use conditions of each vehicle. Since all 25 vehicles need to be flattened to the same dimension, considering the exact cumulative driving time cannot be known, the “absolute value” of cumulative mileage is selected as the abscissa. As can be seen from Figure 2, due to the different charging currents, different vehicle models and other characteristics, the SOH is also quite different under the same mileage, which is in line with the actual situation, but the SOH estimated based on 1468 charging events shows a gradual decline as the accumulated mileage increases. Secondly, using the principle of least squares to fit a straight line as shown in the green dashed line in Figure 2, and then calculating the 95% confidence interval based on the regression line, as shown in Figure 3, the red dashed internal interval in Figure 3, points outside the range of the interval are regarded as abnormal SOH values and deleted which further ensure the accuracy of subsequent labels. There are 1387 charging events rest These charging events can be used as labels for subsequent model training.

**Feature engineering**

The health status of the power battery is affected by many factors such as the number of cycles, use time, battery consistency, user driving habits, etc. This paper identifies 13 characteristics that affect the battery health status, as shown in Table 2.

Among them, the number of times the highest/lowest temperature probe code changes during this charge reflects the stability/disorder of the internal temperature distribution of

![Figure 2. 95% Confidence interval screening of charging events.](image)
the battery pack; the number of times the highest voltage monomer code changes during this charge reflects the consistency of the battery pack, when the battery consistency is better, the code of the highest voltage cell will change alternately. On the contrary, if the highest and lowest voltage always appear on two cells, it means that the battery pack has poor consistency; the initial SOC of this charge reflects the depth of discharge (DOD) of the battery pack, also reflects the user’s habits from the side. For the feature of month, this article uses One-Hot encoding for this feature and converts it into a classification feature. Drawing the distribution diagram between all the features in pairs as

![Distribution diagram](image)

**Figure 3.** Charging events after screening.

**Table 2.** Feature selection of SOH model.

| Feature |
|---------|
| Accumulated mileage |
| Accumulated charging time |
| Average temperature of this charge |
| Average temperature difference of this charge |
| The highest temperature probe code change times of this charge |
| The minimum temperature probe code change times of this charge |
| Average voltage difference of this charge |
| The number of times of change of the highest voltage monomer code of this charge |
| The total voltage at the end of this charge |
| Average current of this charge |
| Starting SOC of this charge |
| SOC at the end of this charge |
| Month |

| symbol | unit |
|--------|------|
| summileage | km |
| sumcharge_time | h |
| avg_temp | °C |
| avg_delta_temp | °C |
| max_temp_sn | times |
| min_temp_sn | times |
| avg_delta_volt | V |
| max_volt_sn | times |
| end_volt | V |
| avg_current | A |
| start_soc | % |
| end_soc | % |
| month | month |
shown in Figure 4. Figure 5 is a partial magnification of the upper left corner of Figure 4. It can be seen that the voltage (end_volt) at the end of the charge and the SOC (end_soc) at the end of the charge are present in the two features marked in the red box in the figure. The obvious positive correlation indicates that the features are redundant and will cause multicollinearity in the model, making the regression model lacking stability and difficult to distinguish the individual effects of each explanatory variable, so only end_soc is retained.

At this time, there are 12 features, 11 of which are numeric features, and 1 is a classification feature (month). According to the Pearson correlation coefficient, the correlation between the numerical features is calculated and the heat map is drawn. Figure 6 shows the Pearson correlation coefficient heat map and the values in the figure represent the Pearson correlation coefficient calculated based on the input features. The purpose is to visualize the degree of correlation between the input features, which is a step in feature engineering. In the figure, you can see that there is no correlation coefficient approaching 1, indicating that the features at this time have met the needs of model training. These 11 numeric features are processed according to Max-Min standardization to eliminate the
Figure 5. Partial magnification of Figure 4.

Figure 6. Feature correlation heat map.
dimensional difference between the features for subsequent model training. The metrics of
the linear analysis with Pearson and Spearman coefficients are shown in Table 3 and
Table 4.

**SOH estimation model based on linear regression**

This paper selects four linear regression models of Linear Regression, Lasso, Ridge and
Elastic Net to establish the SOH estimation model of the power battery health status,

### Table 3. Pearson coefficients.

|                      | summileage | sumcharge_time | avg_temp | avg_delta_temp |
|----------------------|------------|----------------|----------|----------------|
| Summileage           | 1.000      | 0.244          | 0.270    | 0.166          |
| sumcharge_time       | 0.244      | 1.000          | 0.624    | -0.120         |
| avg_temp             | 0.270      | 0.624          | 1.000    | -0.072         |
| avg_delta_temp       | 0.166      | -0.120         | -0.072   | 1.000          |
| max_temp_sn          | -0.170     | 0.176          | 0.124    | -0.210         |
| min_temp_sn          | 0.002      | 0.060          | 0.090    | -0.055         |
| avg_delta_volt       | -0.064     | -0.358         | -0.144   | 0.109          |
| max_volt_sn          | -0.246     | 0.113          | -0.049   | -0.115         |
| avg_current          | -0.321     | 0.373          | -0.065   | -0.221         |
| start_soc            | -0.204     | -0.052         | -0.023   | -0.212         |
| end_soc              | 0.013      | -0.091         | 0.038    | 0.097          |
| soh                  | -0.737     | -0.079         | -0.103   | -0.190         |

|                      | max_temp_sn | min_temp_sn | avg_delta_volt | max_volt_sn |
|----------------------|-------------|-------------|----------------|-------------|
| Summileage           | -0.170      | 0.002       | -0.064         | -0.246      |
| sumcharge_time       | 0.176       | 0.060       | -0.358         | 0.113       |
| avg_temp             | 0.124       | 0.090       | -0.144         | -0.049      |
| avg_delta_temp       | -0.210      | -0.055      | 0.109          | -0.115      |
| max_temp_sn          | 1.000       | 0.276       | -0.228         | 0.238       |
| min_temp_sn          | 0.276       | 1.000       | -0.208         | 0.113       |
| avg_delta_volt       | -0.228      | -0.208      | 1.000          | -0.457      |
| max_volt_sn          | 0.238       | 0.113       | -0.457         | 1.000       |
| avg_current          | 0.294       | 0.173       | -0.473         | 0.356       |
| start_soc            | -0.207      | -0.146      | 0.345          | -0.343      |
| end_soc              | -0.059      | -0.061      | 0.344          | 0.012       |
| soh                  | 0.153       | 0.059       | 0.056          | 0.165       |

|                      | avg_current | start_soc   | end_soc       | soh          |
|----------------------|-------------|-------------|---------------|--------------|
| Summileage           | -0.321      | -0.204      | 0.013         | -0.737       |
| sumcharge_time       | 0.373       | -0.052      | -0.091        | -0.079       |
| avg_temp             | -0.065      | -0.023      | 0.038         | -0.103       |
| avg_delta_temp       | -0.221      | -0.212      | 0.097         | -0.190       |
| max_temp_sn          | 0.294       | -0.207      | -0.059        | 0.153        |
| min_temp_sn          | 0.173       | -0.146      | -0.061        | 0.060        |
| avg_delta_volt       | -0.473      | 0.345       | 0.344         | 0.056        |
| max_volt_sn          | 0.356       | -0.343      | 0.012         | 0.165        |
| avg_current          | 1.000       | 0.062       | -0.188        | 0.367        |
| start_soc            | 0.062       | 1.000       | 0.343         | 0.332        |
| end_soc              | -0.188      | 0.343       | 1.000         | 0.090        |
| soh                  | 0.367       | 0.332       | 0.090         | 1.000        |
compares the differences between the four.\textsuperscript{19} The four regression algorithms are introduced below:

(1) Linear Regression

Linear regression is a kind of regression problem. Linear regression assumes that the target value and the feature are linearly correlated, that is, satisfying a multiple linear

\begin{table}[h]
\centering
\begin{tabular}{lcccc}
\hline
 & summileage & sumcharge_time & avg_temp & avg_delta_temp \\
\hline
summileage & 1.000 & 0.250 & 0.278 & 0.179 \\
sumcharge_time & 0.250 & 1.000 & 0.708 & -0.082 \\
avg_temp & 0.278 & 0.708 & 1.000 & -0.072 \\
avg_delta_temp & 0.179 & -0.082 & -0.072 & 1.000 \\
max_temp_sn & -0.177 & 0.171 & 0.128 & -0.166 \\
min_temp_sn & -0.014 & 0.110 & 0.060 & -0.005 \\
avg_delta_volt & -0.120 & -0.355 & -0.156 & 0.039 \\
max_volt_sn & -0.247 & 0.153 & -0.030 & -0.074 \\
avg_current & -0.334 & 0.344 & -0.004 & -0.167 \\
start_8^\circ c & -0.200 & -0.027 & -0.012 & -0.240 \\
end_soc & 0.117 & -0.020 & 0.093 & 0.115 \\
soc & -0.707 & -0.082 & -0.094 & -0.215 \\
\hline
\end{tabular}
\caption{Spearman coefficients.}
\end{table}
equation. By constructing a loss function, the parameters w and b when the loss function is the smallest are solved. The objective function is:

\[ J(\theta) = \frac{1}{2} \sum_{i} (y^{(i)} - \theta^T x^{(i)})^2 \]  

(3)

In the formula, \( m \) is the number of samples, and \( \theta \) is an n-dimensional vector, which means that the data has n-dimensional features. Solve the parameter vector \( \theta \) that minimizes the formula (3). Linear Regression modeling is fast and simple, but when the number of features is large and the number of samples is small, it is easy to cause the generalization performance of the model to decrease, resulting in the phenomenon of “overfitting”.

(2) Lasso

In order to solve the above-mentioned Linear Regression overfitting problem, it is proposed to add a regularization term of the L1 norm to the objective function (3), called Lasso, and its objective function is:

\[ J(\theta) = \frac{1}{2} \sum_{i} (y^{(i)} - \theta^T x^{(i)})^2 + \lambda \sum_{j} |\theta_j| \]  

(4)

In the formula, \( \lambda \) is called the regularization coefficient, which controls the degree of regularization. It can be seen from the formula that the goal is to find the \( \theta \) that minimizes \( J(\theta) \), so when \( \lambda \) is larger, in order to ensure that \( \lambda \sum_{j} |\theta_j| \) is the smallest, the smaller \( \theta \) is, the higher the degree of regularization. One of the advantages of Lasso is that it can obtain sparse solutions, so that there is a process of dimensionality reduction for features, so it is widely used in engineering applications.

(3) Ridge

Similar to Lasso, Ridge is also to solve the problem of Linear Regression overfitting. Unlike Lasso, Ridge adds a regularization term of the L2 norm to the objective function. The objective function is:

\[ J(\theta) = \frac{1}{2} \sum_{i} (y^{(i)} - \theta^T x^{(i)})^2 + \lambda \sum_{j} \theta_j^2 \]  

(5)

Ridge cannot generate a sparse solution, and the parameter can only be as small as possible but not equal to 0. The advantage of Ridge’s approach is that when there are fewer input features, it will not produce the effect of dimensionality reduction, and will not make most of the parameters before the features zero, so this method is suitable for scenarios with fewer features.

(4) Elastic Net
Elastic Net is a mixture of Lasso and Ridge. It uses both L1 and L2 regularization. The objective function is as follows:

\[ J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2 + \lambda(\rho \sum_{j=1}^{n} |\theta_j| + (1 - \rho) \sum_{j=1}^{n} \theta_j^2) \]  

(6)

In addition to \( \lambda \), \( \rho \) is also added as a regularization coefficient. The two coefficients adjust the regularization ratio of L1 and L2 together. The advantage is that it inherits some of the stability of Ridge on the basis of a certain dimensionality reduction effect.

This paper treats the preprocessed data as time series data, sorts values by the accumulated mileage from small to large, and uses forward verification method which is suitable for time series models to train the model. The sorted data is divided into training set, validation set and test set according to the ratio of 6:2:2. The cumulative mileage of all data ranges from 6930 km to 141,583 km. Therefore, the training set contains data of accumulated mileage of 6930 km to 80,792 km, the validation set contains data of accumulated mileage of 80,792 km to 107,722 km, and the test set contains data of accumulated mileage of 107,722 km to 141,583 km. The model evaluation index selected in this paper is the Mean Absolute Error (MAE), which is the average of the absolute value of the error between the predicted value and the true value:

\[ MAE(X, h) = \frac{1}{m} \sum_{i=1}^{m} |h(x_i) - y_i| \]  

(7)

Where \( h(x_i) \) is the predicted value of the model, that is, the predicted SOH; \( y_i \) is the true value, also called the label of the model, that is, the estimated SOH. The four models on the training set, validation set, and test set the performance is shown in the following Table 5.

By comparing the four linear regression models, it can be seen that Lasso and Elastic Net with L1 regularization both perform poorly, because the actual running vehicle does not upload the data of each single battery, so this article can only select 12 effects. The features of SOH have a small number of features. If L1 regularization is used for dimensionality reduction, the model will be too simple and lead to larger errors. Therefore, the research in this paper shows that Ridge with L2 regularization performs well and is a better choice.20

In this paper, the Ridge linear regression model is used to make SOH prediction, and the SOH prediction chart is drawn, as shown in Figure 7. The blue dot in the figure is the

| Model         | Training set MAE | Validation set MAE | Test set MAE |
|---------------|------------------|--------------------|--------------|
| Linear Regression | 1.57%           | 2.05%              | 3.38%        |
| Lasso         | 2.92%           | 3.80%              | 4.66%        |
| Elastic Net   | 2.92%           | 3.80%              | 4.65%        |
| Ridge         | 1.57%           | 2.06%              | 3.37%        |

Table 5. Horizontal comparison of models.
real SOH under the current accumulated mileage, and the orange dot is the SOH predicted by the model. It can be seen from the figure that the prediction made by this method is basically consistent with the real SOH and meets the actual demand.

**Conclusion**

This paper uses the operating data of 25 new energy vehicles provided by the Shanghai New Energy Vehicle Public Data Collection and Monitoring Research Center to evaluate and build an SOH estimation model. This article fully introduces the entire process, including data preprocessing, SOH evaluation, feature engineering, and training and verification of the final SOH estimation model. This paper compares the application of the four linear regression models named Linear Regression, Lasso, Ridge, and Elastic Net in the prediction of power battery health status. The results of the study show that the MAE of the four is less than 5%. However, in the low-dimensional feature application scenarios without cell data, the most commonly used Lasso model with dimensionality reduction function in the industrial field is not the most suitable. Instead, the simpler Linear Regression or Ridge with the L2 regularization term is more applicable and meets the needs of practical use.

**Author contributions**

Conceptualization and methodology were performed by Huang Bixiong and Liao Haiyu, data curation was performed by Yan xiao, supervision was performed by Liu Xintian and Wang Yiquan, reviewing and editing were performed by Wang Xu. The first draft of the manuscript was written by Huang Bixiong and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.
Data availability statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Xintian Liu https://orcid.org/0000-0002-3395-9176

References

1. Barre A, Deguilhem B, Grolleau S, et al. A review on lithium-ion battery ageing mechanisms and estimations for automotive applications[J]. J Power Sources 2013; 241: 680–689.
2. Wijewardana S, Vepa R and Shaheed MH. Dynamic battery cell model and state of charge estimation[J]. J Power Sources 2016; 380: 109–120.
3. Vichard L, Ravey A, Venet P, et al. A method to estimate battery SOH indicators based on vehicle operating data only[J]. ENERGY 2021; 225: 235–245.
4. HAN X LUL and, et al. LJ. A review on the key issues for lithium-ion battery management in electric vehicles[J]. J Power Sources 2013; 226: 272–288.
5. Li J, Mazzola M, Gafford J, et al. A new parameter estimation algorithm for an electrical analogue battery model[C]. Applied Power Electronics Conference and Exposition (APEC). In: 2012 Twenty-Seventh Annual IEEE, IEEE, 2012, pp.427–433.
6. Li J, Cheng Y, Jia M, et al. An electrochemical-thermal model based on dynamic responses for lithium iron phosphate battery[J]. J Power Sources 2014; 255: 130–143.
7. Gao D and Huang M. Prediction of remaining useful life of lithium-ion battery based on multi-kernel support vector machine with particle swarm optimization[J]. Journal of Power Electronics 2017; 17: 1288–1297.
8. L LIUZ, S YUA and Y LEEJ. Synthesis and characterization of LiNi1-x-yCoxMnyO2 as the cathode materials of secondary lithium batteries[J]. J Power Sources 1999; 9: 416–419.
9. Klass V, Behm M and Lindbergh G. A support vector machine-based state-of-health estimation method for lithium-ion batteries under electric vehicle operation[J]. J Power Sources 2014; 270: 262–272.
10. Hu JK and Vasilakos AV. Energy Big data analytics and security: challenges and opportunities[J]. IEEE Trans Smart Grid 2016; 7: 2423–2436.
11. Chen Z, Mi CC, Fu YH, et al. Online battery state of health estimation based on genetic algorithm for electric and hybrid vehicle applications[J]. J Power Sources 2013; 240: 184–192.
12. Nuhic A, Terzimehic T, Soczka-Guth T, et al. Health diagnosis and remaining useful life prognostics of lithium-ion batteries using data-driven methods[J]. J Power Sources 2013; 239: 680–688.
13. Li SQ, He HW and Li JW. Big data driven lithium-ion battery modeling method based on SDAE-ELM algorithm and data pre-processing technology[J]. Appl Energy 2019; 242: 1259–1273.
14. Peterson SB, Apt J and Whitacre JF. Lithium-ion battery cell degradation resulting from realistic vehicle and vehicle-to-grid utilization[J]. *J Power Sources* 2010; 195: 2385–2392.

15. Zhang JL and Lee J. A review on prognostics and health monitoring of Li-ion battery[J]. *J Power Sources* 2011; 196: 6007–6014.

16. Anton JCA, Nieto PJG, Viejo CB, et al. Support vector machines used to estimate the battery state of charge[J]. *IEEE Trans Power Electron* 2013; 28: 5919–5926.

17. Safari M, Morcrette M, Teyssot A, et al. Life prediction methods for lithium-Ion batteries derived from a fatigue approach II. Capacity-loss prediction of batteries subjected to complex current profiles[J]. *J Electrochem Soc* 2010; 157: A892–A898.

18. Li N, Li BL and Gao L. Transient stability assessment of power system base on XGBoost and factorization machine[J]. *IEEE ACCESS* 2020; 8: 28403–28414.

19. Andre D, Appel C, Soczka-Guth T, et al. Advanced mathematical methods of SOC and SOH estimation for lithium-ion batteries[J]. *J Power Sources* 2013; 224: 20–27.

20. Huang GB, Zhou HM, Ding XJ, et al. Extreme learning machine for regression and multiclass classification[J]. *IEEE Trans Syst Man Cybern B Cybern* 2012; 42: 513–529.

**Author biographies**

**Bixiong Huang** is currently working as a lecturer of vehicle engineering at Shanghai University of Engineering Science since 2014. His research area is focused on the state evaluation of key components of new energy vehicles.

**Haiyu Liao** is pursuing his Master’s degree at Shanghai University of Engineering Science. His research area is focused on the optimization of dynamic system characteristics.

**Yiquan Wang** has graduated from Shanghai University of Engineering Science. His research area was focused on battery algorithms and data science during his postgraduate time.

**Xintian Liu** is currently working as an associate professor of vehicle engineering at Shanghai University of Engineering Science (China) since 2007. His research area is focused on product quality control, fatigue life prediction and evaluation, uncertainty of mechanical system, reliability design theory and vehicle system dynamics.

**Xiao Yan** is currently working as a Chief Scientist at Shanghai Makesens Energy Storage Ltd. His research area is focused on online battery diagnostic, prognostic and digital management.