Article

Health Factor Extraction of Lithium-Ion Batteries Based on Discrete Wavelet Transform and SOH Prediction Based on CatBoost

Mei Zhang, Wanli Chen *, Jun Yin and Tao Feng

College of Electrical and Information Engineering, Anhui University of Science and Technology (AUST), Huainan 232001, China; xwz098@163.com (M.Z.); yj1995543079@163.com (J.Y.); 19965556751@189.com (T.F.) * Correspondence: cwl18815213451@163.com; Tel.: +86-199-655-56751

Abstract: Aiming to accurately identify the state of health (SOH) and the remaining useful life (RUL) of lithium-ion batteries, in this paper, we propose an algorithm for the health factor extraction and SOH prediction of the batteries based on discrete wavelet transform and the Cauchy–Gaussian variation tent sparrow search algorithm (DWT-CGTSSA). Firstly, concerning the inconsistent data length, discrete wavelet transform (DWT) was adopted to decompose the battery’s signals and extract features. Then, the Cauchy–Gaussian variation tent sparrow search algorithm (CGTSSA) was utilized to extract features and obtain the optimal feature subset after encoding. Finally, the optimal feature subset was used to establish a prediction model based on CatBoost for predicting the SOH of lithium-ion batteries. Experiments were conducted for verification. The experimental results showed that the model established in this research is capable of realizing the prediction between different battery packs. The B0005 battery from dataset A was taken as the training set to predict the complete SOH of B0006 and B0007 batteries. For the prediction model of CGTSSA-CatBoost, the goodness of fit ($R^2$) exceeded 0.99, and the value of mean square error (MSE) was less than 1‰. A comparison with other state-of-the-art prediction models verified the superior performance of the CGTSSA-CatBoost model. Under different working conditions, the $R^2$ of all models in dataset B exceeded 0.98.

Keywords: lithium-ion battery; SOH prediction; DWT; CGTSSA; CatBoost

1. Introduction

Due to the merits of high energy density, high power density, and long cycle life, lithium-ion batteries have been widely used to store energy [1–5]. The service life of a battery gradually declines with the prolonged service time, resulting from the battery’s internal chemical reaction and the influence of the external environment. The decline and degradation of lithium-ion batteries will increase the maintenance cost of many electronic devices. The sudden failure of batteries easily causes the crash of large equipment, including new-energy vehicles, which may lead to major accidents. Therefore, the accurate and timely prediction of battery SOH and failure time helps replace the failed batteries in a timely manner, so as to reduce the risks of accidents. As the core energy storage component of new-energy vehicles, lithium-ion batteries are of great significance in energy storage equipment. Therefore, the research on lithium-ion batteries is decisive and meaningful for the development of new-energy vehicles [6–8]. Influenced by the composition and structure of lithium-ion batteries, we cannot directly obtain the real-time SOH of batteries through sensing technology. Therefore, it is necessary to obtain the SOH and RUL of batteries by developing estimation and prediction models [9,10].

In machine learning, excellent features determine the upper bound of the model performance, and feature engineering is particularly critical in the prediction model. The extracted health factors directly determine the prediction effect of the prediction model. However, the length of the measured battery data during the operation of the lithium-ion
battery is uncertain, and the extraction of health factors is difficult, which undoubtedly leads to the inaccurate prediction of SOH and residual service life (RUL) of the battery. The determination of effective feature engineering to realize the extraction and screening of battery health factors has become the focus of research and is challenging to achieve in lithium-ion batteries.

At present, there are two main methods for the RUL prediction of lithium-ion batteries: model-based and data-driven methods. In a model-based method, it is required to establish the simulation model of batteries. Considering the complex chemical reactions in lithium-ion batteries, various partial differential equations are established, involving the calculation of relevant parameters and matrices of a large equivalent circuit model. Therefore, there are many challenges in ensuring the accuracy of the model in practical applications [11]. In a data-driven method, the typical features are usually extracted from degraded data. Then, a machine learning method is used to map the relationship between degraded data and SOH, so as to predict the SOH and RUL. In this research, the data-driven method was adopted.

The battery RUL prediction based on a data-driven method has extensively been investigated. Shi Yongsheng et al. proposed a SOH estimation method for lithium-ion batteries based on an attention-improved bi-directional gating cycle unit (BiGRU). This method realizes high-precision SOH estimation of different types of batteries. They used internal resistance, voltage change rate, test time, discharge energy, discharge depth, and terminal voltage in the discharge cycle as health factors; thus, the selected health factors are fewer, and the feature engineering is simple. When selecting health factors, subjective judgment was used as the basis for selection, and therefore the established feature engineering needs to be improved [12]. Zhang et al. proposed a battery SOH prediction method combining one-dimensional convolution (1DCNN) and long short-term memory network (LSTM). The voltage, current, and temperature intervals in the charging and discharging data were averaged and then serially connected. This method is relatively simple, but it requires high data integrity, and it is difficult to establish feature engineering in incomplete data [13]. Feng HL et al. proposed a new SOH and RUL prediction method based on GPR and established a new SOH prediction model for lithium-ion batteries by improving the basic Gaussian process regression model [14]. Cai L et al. established a more effective SOH estimator through support vector regression (SVR) and the short-term characteristics of the current pulse test and optimized the whole process of the SOH estimator through NSGA-II, taking into account the measurement cost and estimation accuracy of the characteristics. They selected four inflection points in the voltage response curve as the characteristics of the SOH estimator and used feature extraction to eliminate ineffective interference, but the extracted features were fewer, and the generalization ability of the model with fewer features in the actual prediction needs to be proved [15]. Song et al. proposed a SOH estimation method for lithium-ion batteries based on the XGBoost algorithm. The SOH of lithium-ion batteries was estimated using the XGBoost model and then corrected through the Markov chain [16]. Taking the differential voltage curve as well as charging and discharging curves as the degradation features of battery capacity, Li et al. adopted the Elman neural network to predict the RUL of batteries [17]. Park et al. used the wavelet transform method to promote feature extraction to preprocess nonlinear feature data from lib and used convolutional neural network (CNN) and long short-term memory (LSTM) technology for lithium-ion SOH estimation using the wavelet transform method [18]. Kaur K et al. proposed three models of different network architecture families. By considering the influence of parameters such as the complexity of the model, the sampling rate of battery measurable signals, and the type of battery measurable signals, they evaluated the battery capacity estimation of different models, relying on the ability to effectively process time series data by remembering long-term dependence [19]. Khumprom et al. applied the deep neural network (DNN) to predict the SOH and RUL of lithium-ion batteries, showing equivalent or more competent performance, compared with other machine learning algorithms [20]. Rossi et al. proposed a method to adjust the extended Kalman filter (EKF) covariance matrix by applying the optimization process based on the genetic
2. Algorithm Principle

Regarding the prediction of battery life, we adopted the DWT-CGTSSA to extract the health factor features and constructed a CGTSSA-CatBoost model to predict the SOH. The detailed principle is shown in Figure 1.

![Figure 1. Multiscale health factor extraction and CGTSSA-CatBoost prediction model based on DWT.](image)

The health factor extraction and the CGTSSA-CatBoost prediction model based on the DWT-CGTSSA are mainly composed of three parts: feature engineering construction, the CGTSSA-optimized CatBoost model, and SOH prediction. In feature engineering construction, DWT is used to decompose the voltage and current signals of charging and discharging into approximate signals and detailed signals. Through analyzing the signal features of signal amplitude factor and pulse factor, a 320-dimensional feature vector set is obtained. By eliminating the redundant features using the CGTSSA optimization algorithm, an 11-dimensional optimal feature subset is obtained. The CGTSSA-optimized CatBoost model adopts the CGTSSA to optimize the parameters of the CatBoost model and obtains the optimal CatBoost model through training. The obtained optimal CatBoost model predicts the SOH of lithium-ion batteries, outputs the prediction results, and conducts an evaluation of the model.

2.1. Discrete Wavelet Transform (DWT)

Wavelet transform has better performance in processing nonstationary sequence signals than Fourier transform. By replacing the sinusoidal and conical waves decomposed by Fourier transform with a set of degenerative orthogonal bases, the abrupt and nonstationary parts of the signal can be better performed [23].
The data $x[n]$ are passed through a half-band low-pass filter with impulse response $h[n]$ using

$$x[n] \times h[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot x[n-k]$$ (1)

Following the Nyquist theorem, we performed downsampling for data. One sample point is removed at an interval, one-half of the sample points of the signal are reserved, and the scale is doubled. The high-pass filtering is performed for this half by

$$y[n] = \sum_{k=-\infty}^{\infty} h[k] \cdot x[2n-k]$$ (2)

$x[n]$ is decomposed into $\alpha$ layers by wavelet transform. The low-frequency information and high-frequency information of layer $\alpha$ are expressed by

$$\begin{cases} x_{\alpha,L}[n] = \sum_{k=0}^{K-1} x_{\alpha-1,L}[2n-k] \cdot g[k] \\ x_{\alpha,H}[n] = \sum_{k=0}^{K-1} x_{\alpha-1,L}[2n-k] \cdot h[k] \end{cases}$$ (3)

where $h[k]$ is a high-pass filter, $g[k]$ is a low-pass filter, $x_{\alpha,L}[n]$ is the low-frequency information in layer $\alpha$, and $x_{\alpha,H}[n]$ is the high-frequency information in layer $\alpha$.

### 2.2. CGTSSA

When initializing the population, the sparrow search algorithm uses the random generation method. The disadvantage of this method is that it will make the sparrow population unevenly distributed and affect the subsequent iterative optimization. By taking advantage of the randomness, ergodicity, and regularity of chaotic mapping, the tent sparrow search algorithm (TSSA) optimizes the position of individual sparrows by using tent maps in chaotic mapping to avoid falling into local optimization, thus improving the global search ability and optimization accuracy [24,25].

The expression of tent mapping is given by

$$Z_{i+1} = \begin{cases} 2 \times Z_i & 0 \leq Z_i \leq \frac{1}{2} \\ 2 \times (1 - Z_i) & \frac{1}{2} < Z_i \leq 1 \end{cases}$$ (4)

where $Z_i$ is the initial value, and $Z_{i+1}$ is the value of tent after mapping.

The specific steps of TSSA are as follows:

1. Apply Equation (7) to generate a chaotic variable $Z^d$ according to initial particle $X_d$.
2. Transfer the chaotic variable carrier to the solution space of the problem to be solved as follows:

$$X_{new}^d = X_{min}^d + (X_{max}^d - X_{min}^d) \times Z^d$$ (5)

where $X_{max}^d$ and $X_{min}^d$ are the maximum and minimum values of the $d$-dimensional variable $X_{new}$, respectively.

3. Perform the chaotic disturbance to the individual according to Equation (6):

$$X'_{new} = \frac{(X + X_{new})}{2}$$ (6)

where $X$ is the individual requiring chaotic disturbance, $X_{new}$ is the generated chaotic disturbance, and $X'_{new}$ is the individual after chaotic disturbance.

During the late iteration of TSSA, the assimilation and mutation strategy among sparrows lead to a shorter search step, resulting in local optimization. To avoid local optimization and ensure the mutation sparrow has ample step size in the later stage,
we adopted the Cauchy–Gaussian mutation strategy to mutate the optimal individual position [26,27]. The specific iteration formula is expressed as

\[ U_{best}^t = X_{best}^t \left[ 1 + \lambda_1 Gauss(0, \sigma^2) + \lambda_2 Cauchy(0, \sigma^2) \right] \]  

(7)

where \( U_{best}^t \) represents the position of the optimal individual after variation, \( \sigma^2 \) represents the standard deviation of Cauchy–Gaussian mutation strategy, \( Gauss(0, \sigma^2) \) is a random variable satisfying the Gaussian distribution, \( Cauchy(0, \sigma^2) \) is a random variable satisfying the Cauchy distribution, and \( \lambda_1 \) and \( \lambda_2 \) are dynamic parameters adaptively adjusted with the number of iterations. Influenced by and the Cauchy–Gaussian mutation mode, with the increase in iteration time, \( \lambda_1 \) gradually increases, and \( \lambda_2 \) gradually decreases, which can make the algorithm locally optimal and coordinate its local development and global exploration capabilities during iteration.

The expression of \( \sigma^2 \) is

\[ \sigma = \left\{\begin{array}{ll}
1, & f(X_{best}) < f(X_i) \\
\exp\left(\frac{f(X_{best}) - f(X_i)}{|f(X_{best})|}\right), & \text{otherwise}
\end{array}\right. \]  

(8)

where \( f(X_{best}) \) is the fitness value of the best individual sparrow at present, and \( f(X_i) \) is the fitness value after variation.

2.3. CatBoost Model

Composed of categorical and boosting features, the gradient boosting decision tree (GBDT) algorithm is a base learner based on oblivious trees, with the advantages of fewer parameters, supporting the categorical variables, and high accuracy [28,29].

2.3.1. GBDT Algorithm

The GBDT algorithm mainly includes the bagging algorithm and boosting algorithm. As the mainstream algorithm, the boosting algorithm can promote a weak learner to a strong learner. It relies on the framework of an integrated algorithm, uses the previous round of weak-learner errors to strengthen the next round of weak learners through iteration, and fuses multiple learners to form a strong learner through some combination strategy.

Gradient boosting is a machine learning technique used for regression, classification, and sorting tasks. It belongs to the boosting algorithm family. The GBDT algorithm is a novel integrated algorithm fused from the CART number and gradient lifting algorithm [30].

The GBDT framework mainly has four steps:

Step 1. Initialize the weak learner:

\[ f_0(x) = \arg\min_c \sum_{i=1}^N L(y_i, c) \]  

(9)

where \( L(y_i, c) \) is the loss function, \( y_i \) is the \( i \)th label, and \( c \) is the parameter with the smallest square loss function.

Step 2. Calculate the residual error:

\[ r_{im} = -\left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right] \left| \begin{array}{c}
\text{f}(x) = f_{m-1}\left( x \right)
\end{array}\right. \]  

(10)

where \( M \) is the number of iterations \((m = 1, 2, \ldots, M)\); \( N \) is the total number of samples \((i = 1, 2, \ldots, N)\).
Step 3. Strengthen the next learner according to the residual of the last weak learner:

\[ f_m(x) = f_{m-1}(x) + \sum_{j=1}^{I} \gamma_{jm} I_{jm}(x \in R_{jm}) \]  

\[ \text{(11)} \]

2.3.2. Principle of CatBoost Model

Similar to all standard gradient lifting algorithms, CatBoost also builds a new tree to fit the gradient of the current model. However, all classical lifting algorithms have overfitting problems caused by biased pointwise gradient estimation. To solve this problem, CatBoost has made some improvements to the classical gradient lifting algorithm [31].

Set \( D \) as the training set:

\[ D = (X_i, Y_i) \]  

\[ \text{(12)} \]

where \( n \) is the number of sample groups (\( i = 1, 2, \ldots, n \)), each group of samples is \( X_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \), \( x_{im} \) is the \( m \)th feature vector of group \( i \) of samples, and \( Y_i \) is the \( i \)th target value of \( Y \).

The feature conversion value of the CatBoost model is

\[ \hat{x}_i^k = \frac{\sum_{j=1}^{n} \varphi(x_{j}^k = x_{i}^k) Y_j + \alpha p}{\sum_{j=1}^{n} \varphi(x_{j}^k = x_{i}^k)} + \alpha \]  

\[ \text{(13)} \]

where \( \varphi \) is the indicator function, \( p \) is the prior value, and \( \alpha \) is a priori weight.

Using ordered boosting to replace the gradient estimation method in the traditional algorithm, CatBoost is capable of reducing the deviation of gradient estimation and improving the generalization capability of the model.

3. Experimental

3.1. Instrument and Equipment

The instruments used in the experiment include a battery detector, a thermostat, an upper computer, and a crocodile clip.

Adopting a battery tester, we conducted a circular charge–discharge experiment on the battery and detected the relevant parameters, including voltage, current, power, and capacity during the process. The experiment was conducted in an environment with a temperature of 15 \( ^\circ \)C through a constant temperature box.

An 18,650 lithium-ion battery was taken as the experimental object, the detailed information of which is shown in Table 1.

Table 1. Battery’s detailed parameter table.

| Item                        | Specifications       | Item                        | Specifications       |
|-----------------------------|----------------------|-----------------------------|----------------------|
| Shell material              | Nickel plated steel  | charging method            | standard             |
| Nominal capacity            | 1300 mAh             | (CC/CV) fast               | 0.5C_5A \times 7.5 \text{ h} |
| Rated voltage               | 3.7 V                | charging                   | 1C_5A \times 2.5 \text{ h} |
| Charging voltage (Max)      | 4.2 V                | discharging                | 0 \text{°C}–45 \text{°C} |
| Discharge cutoff voltage    | 2.7 V                | charging                   | 32 \text{°F}–113 \text{°F} |
| Charging current (Max)      | 1 C_5A               | discharging                | \text{−15 \text{°C}–60 \text{°C}} |
| Discharge current (Max)     | 3 C_5A               | storage                    | \text{5 \text{°F}–140 \text{°F}} |
| Internal resistance (Max at 1000 Hz) | \leq 25 mΩ   | working temperature        | \text{−20 \text{°C}–60 \text{°C}} |
|                             |                      |                             | \text{−4 \text{°F}–113 \text{°F}} |

3.2. Experimental Steps

The experimental process is shown in Figure 2.
Figure 2. Experimental process.

The experimental equipment is shown in Figure 3.

![Experimental equipment diagram](image)

(a) Experimental equipment. (b) Equipment connection diagram.

As shown in Figure 3, the computer was used as the host computer; then, the host computer was connected to a Xinwei tester by using the LAN, and the battery was connected with the crocodile clip. The tester detected the voltage, current, power, and other data of the battery during charging and discharging.

3.3. Experimental Data

In this research, two datasets were used, i.e., dataset A and dataset B, to verify the generalization capability of the proposed feature extraction method and the prediction model. Taking dataset A as the research object, we established the prediction model and used B0005 as the training set to predict the SOH and RUL of B0006 and B0007. Taking dataset B as the verification object, we reestablished feature engineering for dataset B, developed the prediction model, and observed the prediction effect, so as to verify that the feature engineering proposed in the paper has a strong generalization capability.
Dataset A is from the first batch dataset of NASA’s Prediction Center of Excellence (PCoE). Taking B0005, B0006, and B0007 batteries as the data, we recorded the voltage, current, and temperature during the battery cycle charging and discharging. The experimental process is as follows: First, the battery was charged with 1.5 A constant current, followed by constant voltage charging; after the voltage reached 4.2 V, charging was stopped until the charging current was less than 20 mA; lastly, the battery was discharged with 2.0 A constant current until the voltage was less than 2.5 V. According to the above process, the battery was charged and discharged for 168 cycles. When conducting the prediction experiment for the battery SOH, the detection parameters (voltage, current, and power) of dataset A were taken as independent variables in the prediction model, and the total discharge capacity (Ah) of each cycle was taken as the prediction object of the prediction model.

Taking the 18650 battery as the research object, dataset B comprises the data measured in the experiment according to the experimental process in Section 2.2. A total of 148 cycles of battery parameters were obtained. The total discharge capacity (mAh) of each cycle was taken as the prediction object of the prediction model.

4. Extraction of Health Factor Based on DWT-CGTSSA

The charging and discharging time of lithium-ion batteries changes with the aging of the battery during the process of cycle charging, resulting in inconsistent time steps of the parameters measured in each cycle. Taking the charging current of the B0005 battery in dataset A as an example, the charging current cycle curves of different cycles are shown in Figure 4.

![Figure 4. Charging current curve of different cycles.](image)

It can be seen from Figure 4 that the charging time of each charging cycle was inconsistent, and the length of the sequence was different. With the transformation of the curve obtained by the number of cycles, the transformation trend between some characteristics on the curve and the capacity was similar. The sampling frequency of the first 31 cycles of the curve was obviously inconsistent with that of the following cycles, and the time length of the curve at 31 cycles presented a cliff-like change. There were 10 groups in dataset A, which had similar curves, and there were 6 groups of curves in dataset B. However, these data should be subjected to feature extraction before being input into the prediction model.

4.1. Extraction of Health Factor Based on DWT

Through decomposing the signal into approximate signal and detail signal using DWT, we extracted the curve features, captured the waveform information, calculated the
amplitude factor, waveform factor, pulse factor, and margin factor as the health factors for SOH prediction, and finally obtained the feature vector set.

The extracted information is shown in Table 2.

Table 2. Table of Feature information.

| Feature Value      | Symbolic Representation | Feature Value      | Symbolic Representation |
|--------------------|-------------------------|--------------------|-------------------------|
| Mean value         | Mean                    | Mean absolute value| Mae                     |
| Standard deviation | Std                     | Kurtosis           | Kur                     |
| Root mean square   | Rms                     | Energy             | En                      |
| Maximum value      | Max                     | Amplitude factor   | Cre                     |
| Minimum value      | Min                     | Waveform factor    | Sha                     |
| Peak value         | Peak                    | Impulse factor     | Imp                    |
| Skewness           | Ske                     | Clearance factor   | Cle                    |
| Variance           | Var                     | Root amplitude     | Root                   |

According to Table 2, the approximate signal and detail signal of each charging and discharging parameter had 16 kinds of feature information, with a total of 32 kinds of features. Dataset A generated 10 sets of signal curves for each charging and discharging cycle. Therefore, a 320-dimensional feature vector set was obtained for dataset A in the experiment.

4.2. Selection of Health Factors Based on CGTSSA

As mentioned in Section 3.1, we finally obtained a 320-dimensional feature vector of dataset A. Excess features increase the complexity of the model and cause overfitting. Therefore, it is necessary to remove irrelevant and redundant features.

Adopting the CGTSSA to encode the optimal feature subset, we selected the health factors. First, the 320-dimensional feature vector set was taken as a variable with a length of 320 dimensions. The variable value can be filled with 0/1, where 0 represents the unselected feature, and 1 represents the selected feature. Next, the B0005 battery of dataset A was divided into a training set and a verification set. The first 100 cycles were set as training samples, and cycles 101 to 168 were set as verification samples, aiming to verify that the fitness function of the MSE value of the verification set is the fitness value and establish a fitness function. Taking the CGTSSA as the optimization algorithm, the CatBoost model as the prediction model, and the selected feature set as variables, the features of the feature vector set were extracted by constructing a CGTSSA coding algorithm. Finally, taking the 320-dimensional vector as the independent variable, the MSE value as the fitness function value, and the CGTSSA as the optimization algorithm, we constructed a binary CGTSSA optimization algorithm to find the optimal feature subset. The fitness function curve of the health factors is shown in Figure 5.

![Fitness curve of CGTSSA screening health factors.](image_url)
As can be seen from Figure 5, the 11-dimensional optimal feature subset was obtained after CGTSSA coding. When the number of CGTSSA iterations reached 20, the fitness value no longer changed, and the optimal solution of CGTSSA was obtained. The MSE value reached 0.0298\% in the verification set.

The Pearson correlation coefficient between the optimal feature subset and the battery capacity to be predicted is shown in Figure 6.

As shown in Figure 6, there was a strong correlation between some screened features and capacity.

5. Model Establishment and Verification

5.1. Comparison of CatBoost Model and Its Parameter Optimization Algorithm

As mentioned in Section 3.1, we finally obtained a 320-dimensional feature vector of dataset A. Excess features increase the complexity of the model and cause overfitting. Therefore, it is necessary to remove irrelevant and redundant features. First, the 320-dimensional feature vector set was taken as a variable with a length of 320 dimensions. The variable value can be filled with 0/1, where 0 represents the unselected feature, and 1 represents the selected feature. Next, the B0005 battery of dataset A was selected as the prediction model, and the selected feature set as variables, the features of the feature vector set were extracted by constructing a CGTSSA coding algorithm. Finally, taking the 320-dimensional vector as the independent variable, the MSE value as the fitness function, and the CGTSSA as the optimization algorithm, we constructed a binary CGTSSA optimization algorithm to find the optimal feature subset. The fitness function curve of CGTSSA screening health factors is as follows:

![Fitness curve of CGTSSA screening health factors.](image)

As can be seen from Figure 5, the 11-dimensional optimal feature subset was obtained after CGTSSA coding.

For dataset A, we used B0005 as the training set, extracted the feature vector by adopting the proposed feature extraction method, trained the CatBoost model, and obtained the prediction model. Then, the model was used to predict the SOH and RUL of B0006 and B0007, so as to realize the prediction between different batteries under the same model and the same charging and discharging strategy. Finally, the complete cycle capacity of several other batteries was predicted based on the model constructed by available battery training.

Super-parameter setting greatly affects the performance of the CatBoost model. If the parameter learning_rate is too small, the learning step of the model will be small, resulting in the insufficient fitting of the model. If the parameter learning_rate is too large, the learning span will be large, and the model’s convergence will be reduced, resulting in the overfitting of the model. The regularization parameter L2_leaf_reg has a direct influence on the fitting capability of the model. The matched L2_leaf_reg parameter is capable of preventing the model from overfitting and enhancing the generalization capability of the model. Bayesian bagging controls the intensity of bagging_temperature, thereby affecting the performance of the model. The depth of the tree affects the prediction effect of the model. A deeper tree has an advantage in increasing the expression capability of the model but has the disadvantage of causing model overfitting [32]. The optimization algorithm was adopted to optimize the super parameters in the above four CatBoost models, thereby improving the performance of the model and enhancing the generalization capability of the model.

The commonly used parameter optimization algorithms include particle swarm optimization (PSO) and sparrow search algorithm (SSA). In this research, CGTSSA, SSA, and PSO were adopted to optimize the four super parameters of the CatBoost model, and a comparison between the algorithms was made. In the experiment, the three optimization models and the original CatBoost model were used to train the data of the B0005 battery, and then the trained model was used to predict the battery life of B0006 and B0007. The prediction results are shown in Figure 7.
### Table 3. Detailed prediction results of various algorithms.

| Data   | Method        | Really Life | Predict Life | AE     | R²     | RMSE   | SSE   | MSE (%) |
|--------|---------------|-------------|--------------|--------|--------|--------|-------|---------|
| B0006  | CGTSSA-CatBoost | 113         | 113          | 0      | 0.9938 | 0.0268 | 0.1210 | 0.7202  |
|        | SSA-CatBoost  | 112         | 112          | 1      | 0.9918 | 0.0317 | 0.1689 | 1.0056  |
|        | PSO-CatBoost  | 110         | 110          | 3      | 0.9807 | 0.0531 | 0.4728 | 2.8144  |
|        | CatBoost      | 102         | 102          | 11     | 0.9561 | 0.0708 | 0.8418 | 5.0109  |
| B0007  | CGTSSA-CatBoost | 126         | 126          | 0      | 0.9947 | 0.0118 | 0.0233 | 0.1386  |
|        | SSA-CatBoost  | 127         | 127          | 1      | 0.9926 | 0.0147 | 0.0362 | 0.2155  |
|        | PSO-CatBoost  | 128         | 128          | 2      | 0.9828 | 0.0263 | 0.1166 | 0.6941  |
|        | CatBoost      | 117         | 117          | 9      | 0.9412 | 0.0465 | 0.3631 | 2.1613  |

Combining the results in Table 3 and Figure 7, for the CGTSSA-CatBoost model, the AE was 0, R² was higher than 0.99, SSE was lower than 0.2, and MSE was lower than 1‰.

**Figure 7.** Prediction results of CatBoost model and its optimization model.

As shown in Figure 7, the battery capacity dropping to 80% of rated capacity was considered a failure, and the failure thresholds were set to 1.38 Ah and 1.5 Ah for the B0006 and B0007 batteries, respectively.

It can be seen from Figure 7 that the CatBoost model with default parameters achieved the poorest prediction effect.

To comprehensively compare the performance of the above models, in this paper, the square sum of error (SSE), root mean square error (RMSE), the goodness of fit (R²), absolute error (AE), and mean square error (RMSE) of the residual life of the batteries were taken as the indicators to compare the model performance.

The detailed comparison results of the models are shown in Table 3.

As shown in Figure 7, the battery capacity dropping to 80% of rated capacity was considered a failure, and the failure thresholds were set to 1.38 Ah and 1.5 Ah for the B0006 and B0007 batteries, respectively.

It can be seen from Figure 7 that the CatBoost model with default parameters achieved the poorest prediction effect.

To comprehensively compare the performance of the above models, in this paper, the square sum of error (SSE), root mean square error (RMSE), the goodness of fit (R²), absolute error (AE), and mean square error (RMSE) of the residual life of the batteries were taken as the indicators to compare the model performance.

The detailed comparison results of the models are shown in Table 3.
1% in B0006 and B0007 batteries. Thus, the CGTSSA-CatBoost model showed the highest prediction accuracy and the best prediction effect among the above models.

5.2. Comparison with Other Models
At present, with the development of machine learning, there are many prediction models in the market. The ELM and SVM models are the most classical and popular prediction models, which play great roles in the SOH prediction of lithium-ion batteries. Many studies in the literature are based on the ELM and SVM as prediction models. The SVM model has a good processing effect in small-sample data processing [15,33]. Since the kernel function method overcomes the problems of dimension disaster and nonlinear separability, the computational complexity is not increased when mapping to high-dimensional space. The ELM model is a simple and effective single-hidden-layer feedforward neural network SLFN learning algorithm, which has the advantages of fewer training parameters and very fast speed.

To obtain the optimal prediction model, we selected the ELM, support vector machine (SVM), and CatBoost models to predict the preprocessed data and adopted CGTSSA to optimize the ELM, SVM, and CatBoost models. The performance of the models is compared, as shown in Figure 8.

![Figure 8. Comparison results of multiple models.](image_url)
The detailed indicators of each model are shown in Table 4.

Table 4. Detailed prediction results of various models.

| Data   | Method          | Really Life | Predict Life | AE | R^2   | RMSE   | SSE   | MSE (%) |
|--------|-----------------|-------------|--------------|----|-------|--------|-------|---------|
| B0006  | CGTSSA-CatBoost | 113         | 0            | 0  | 0.9938| 0.0268 | 0.1210| 0.7202  |
|        | CGTSSA-SVM     | 110         | 3            | 0  | 0.9901| 0.0362 | 0.2206| 1.3130  |
|        | CGTSSA-ELM     | 110         | 3            | 0  | 0.9750| 0.0579 | 0.5632| 3.3523  |
| B0007  | CGTSSA-CatBoost| 126         | 0            | 0  | 0.9947| 0.0118 | 0.0233| 0.1386  |
|        | CGTSSA-SVM     | 129         | 3            | 0  | 0.9937| 0.0178 | 0.0532| 0.3169  |
|        | CGTSSA-ELM     | 104         | 22           | 0  | 0.9798| 0.0447 | 0.3356| 1.9977  |

Combined with the data in Table 4 and Figure 8, it was revealed that the CGTSSA-CatBoost model had the best performance in the B0006 and B0007 batteries. The prediction effect of the ELM model was slightly worse, and the prediction effect of the SVM model was better. When predicting the same battery, the performance of the SVM model was close to the CatBoost model, and the R^2 of the SVM model was higher than 0.99, indicating that the model has a good fitting effect. In terms of R^2, RMSE, MSE, and SSE indicators, the CatBoost model performed best, i.e., the CatBoost model had the best SOH prediction effect, followed by the SVM and ELM models. As regards the AE index, the AE value of the CatBoost model was 0, which was the lowest, indicating that the RUL prediction effect of the CatBoost model was the best. The RUL of the ELM model was the same as that of the SVM model on the B0006 battery, but the prediction effect of the SVM model was higher than that of the ELM model on the B0007 battery. Compared with the ELM and SVM models, the CatBoost model had better model performance and stronger generalization ability.

5.3. Verification of the Generalization Capability

To verify the generalization capability of the feature engineering and prediction model proposed in this paper, we re-extracted features for dataset B and performed SOH prediction based on the CGTSSA-CatBoost model. There were 148 cycle data in dataset B. The first 68 cycles were used in the training set, and the last 80 cycles were the test set. The above four models were used to predict the batteries’ capacity, with the prediction results shown in Figure 9.

Figure 9. Prediction results of dataset B.
Figure 9 shows that the CatBoost model optimized by super parameters had a better prediction performance than that with default parameters. The detailed evaluation indicators are shown in Table 5.

Table 5. Prediction results of various algorithms.

| Method           | R²   | RMSE  | SSE    | MSE   |
|------------------|------|-------|--------|-------|
| CGTSSA-CatBoost  | 0.9962 | 1.1905 | 113.3773 | 1.4172 |
| SSA-CatBoost     | 0.9961 | 1.8156 | 263.7207 | 3.2965 |
| PSO-CatBoost     | 0.9958 | 2.1051 | 354.504 | 4.4313 |
| CatBoost         | 0.9864 | 4.7566 | 1809.9964 | 22.625 |

Figure 9 and Table 5 suggest that the CGTSSA-CatBoost model had the lowest MSE value but the best model performance in dataset B. It has also achieved good prediction results in the feature engineering established in this paper, with a value of RMSE < 5, which is capable of realizing the accurate prediction of SOH of the same battery.

In summary, the DWT-CGTSSA health factor extraction method and CGTSSA-CatBoost prediction model proposed in this paper had better prediction accuracy in different datasets than other state-of-the-art models or methods. The proposed model had a satisfying generalization capability. The results in Figures 7 and 9, Tables 3 and 5 revealed that the proposed method and model were capable of realizing the SOH prediction of different batteries, under the same charging and discharging strategy and under different working conditions. Satisfying prediction accuracy was achieved, with a goodness of fit of 0.99, and an MSE value of <1‰.

6. Discussion

Most of the published papers used the first I cycle of the same battery to predict the next II cycle (I + II = N, N is the total number of cycles). However, in practical applications, it is preferable to achieve the model trained by one or more batteries to predict the complete SOH of another battery under the same model and working conditions. In this study, it was found that the SOH prediction of several other batteries by one battery was valid, and some characteristics were significantly correlated with the SOH of the battery under the same model and working conditions. Each charging and discharging of the battery produces a large volume of data. Thus, how to mine information from the data is very important. In this paper, the feature engineering established by the DWT-CGTSSA could effectively extract the characteristics of the battery. The model established by these features could obviously predict the SOH between different batteries, which was proved in dataset A, but the premise was that the models, working conditions, and working environment of these batteries (temperature, humidity, etc.) were the same. In this paper, dataset B was used to test the generalization ability of feature engineering, and it was also proven that the prediction of feature engineering in the same battery could also achieve good results.

In this paper, NASA data and NCM batteries in the laboratory were used. Their capacities are different, so it is obvious that their chemical materials are slightly different. The algorithm proposed in this paper had a good prediction effect on the two kinds of batteries with different materials. The algorithm in this paper had a markedly strong generalization ability and could adapt to the prediction of NASA batteries and NCM batteries. In our subsequent work, we plan to apply the algorithm to lithium-ion batteries of various materials to observe whether the proposed algorithm can be applied to the SOH prediction of lithium-ion batteries with different chemicals. In addition, our later work will be combined with hardware, embed the algorithm into a microcontroller, and combine the algorithm with practical applications to make a real-time detection system. In addition, it should be noted that the prediction model proposed in this paper needs to be batteries of the same model, the same working condition, and the same environment. For different batteries, we need to redefine our algorithm to establish feature engineering.
Among other relatively new research results, the authors of [34] realized prediction in the same battery, whereas in this study, we completed prediction between different batteries. Compared with the study of [34], our model is more practical. In another study [16], the SOH of the whole battery was also completely predicted. NASA data were used in the data of this paper and in the study of [16], but its training set used three groups of batteries. In contrast, the training set in this paper only used one group of batteries, but the MSE value in this paper was lower when predicting the same battery, and the prediction effect was better. Therefore, the proposed model in this study used fewer training data to achieve better results.

7. Conclusions

In this paper, we adopted the DWT-CGTSSA to extract features, establish the CGTSSA-CatBoost prediction model, and conduct the experiment on different datasets to accurately predict the SOH and RUL of lithium-ion batteries. The conclusions are summarized as follows:

1. By extracting the characteristics of signal amplitude factor and pulse factor after DWT decomposition, and then screening the characteristics by using the CGTSSA algorithm, the characteristics that were more suitable for the batter SOH prediction model could be extracted.

2. In the established feature engineering, the CatBoost and its optimization models were used to predict the SOH of different batteries in dataset A, and good prediction results were obtained. Among them, the CGTSSA-CatBoost model had the best prediction effect, with AE 0, an MSE lower than 1‰, and an SSE lower than 0.2.

3. Compared with the ELM and SVM models commonly used in the SOH prediction of battery, the CatBoost model had better performance and a better effect on multiple indicators such as MSE and RMSE. The AE index was 0, and the RUL prediction error of the model was lower.

4. The SOH prediction of the same battery in dataset B using the feature engineering and prediction model in this paper also achieved good prediction results, the RMSE was less than 5, and the proposed method had a strong generalization ability.

In summary, the proposed feature engineering and prediction model can not only accurately predict the SOH of the same battery but also accurately predict the SOH between different batteries under the same model and the same charging and discharging strategy. The latter is more meaningful in practice.

Author Contributions: Conceptualization, W.C.; data curation, M.Z.; methodology, W.C.; project administration, M.Z.; resources, M.Z. and J.Y.; validation, J.Y.; visualization, T.F.; writing—original draft preparation, W.C.; writing—review and editing, M.Z. and W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of the Higher Education Institute of Anhui Province (KJ2020A0309) and the National Natural Science Foundation of China (51874010).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The access URL for dataset A in the manuscript is: https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/, accessed on 16 November 2021.

Acknowledgments: The authors would like to thank Natural Science Foundation of the Higher Education Institute of Anhui Province for helpful discussions on topics related to this work. The authors would like to thank Natural Science Foundation of the National Natural Science Foundation of China for helpful discussions on topics related to this work.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.
30. Fu, F.C.; Jiang, J.W.; Shao, Y.X.; Cui, B. An Experimental Evaluation of Large Scale GBDT Systems. *Proc. VLDB Endow.* 2019, 12, 1357–1370. [CrossRef]

31. Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for big data: An interdisciplinary review. *J. Big Data* 2020, 7, 94. [CrossRef] [PubMed]

32. Aggarwal, A.; Chakradar, M.; Bhatia, M.S.; Kumar, M.; Stephan, T.; Gupta, S.K.; Alsamhi, S.H.; Al-Dois, H. COVID-19 Risk Prediction for Diabetic Patients Using Fuzzy Inference System and Machine Learning Approaches. *J. Healthc. Eng.* 2022, 2022, 4096950. [CrossRef] [PubMed]

33. Ge, D.D.; Zhang, Z.D.; Kong, X.D.; Wan, Z.P. Extreme Learning Machine Using Bat Optimization Algorithm for Estimating State of Health of Lithium-Ion Batteries. *Appl. Sci.* 2022, 12, 1398. [CrossRef]

34. Li, Q.L.; Li, D.Z.; Zhao, K.; Wang, L.C.; Wang, K. State of health estimation of lithium-ion battery based on improved ant lion optimization and support vector regression. *J. Energy Storage* 2022, 50, 104215. [CrossRef]