Int. J. Electrochem. Sci., 17 **(2022)** Article Number: 220823, doi: 10.20964/2022.08.30

International Journal of
ELECTROCHEMICAL
SCIENCE
www.electrochemsci.org

Hybrid Estimation Strategy for the State of Health of Lithium-ion Batteries with Different Application Needs

Zhigang He[1], Xiaodan Guo[1,], Shuai Hu[2] , Weiquan Li[3], Xianggan Ni[1]*

[1] College of Automobile and Traffic Engineering, Jiangsu University, Zhenjiang, 212013, China;
[2] Yongkang Quality And Technology Monitoring Institute (national inspection center for Hardware & Door Product Quality (Zhejiang)), Zhejiang, 321300, China;
[3] Zhejiang Fangyuan Test Group Co. Ltd, Zhejiang 310018, China.
[*]E-mail: g15751779331@163.com (Xiaodan Guo)

Received: 30 April 2022 / *Accepted:* 9 June 2022 / *Published:* 4 July 2022

An accurate estimation of the state of health (SOH) of a lithium-ion battery pack can ensure the safety of the battery. Estimating the SOH based on data-driven models is the current mainstream direction. For different application needs, focusing on the selection of health features and models is necessary. Therefore, a hybrid estimation strategy for the state of health of lithium-ion battery packs based on multiple feature dimensions and networks is proposed in this study. The real operating data of different vehicles are processed. Health features are extracted based on the frequency and number of frequency dimensions. The nonlinear degenerate relationship between health features and capacities is learned through different networks. The results show that combining frequencies and the number of frequency-based health features can significantly improve the estimation accuracy. The Long Short-Term Memory Network (LSTM) model is more advantageous when processing data with high feature dimensions, and the Gate Recurrent Unit Network (GRU) model is more advantageous when dealing with a large amount of data. It also shows the effectiveness of this hybrid estimation strategy.

Keyword: Lithium-ion battery pack; SOH; Frequency; Number of Frequencies; Network

1. INTRODUCTION

The automotive industry has responded to the changing trend of continuous developments of new energy technologies and continuous consumption of traditional energy. As of 2021, China's production of new energy vehicles has reached 3.677 million units. Moreover, lithium-ion batteries are widely used because of their safety and long life[1]. As a key component of new energy vehicles, the use of high-performance batteries is particularly important, and accurately evaluating battery performance parameters is even more important. Among them, the State of Charge (SOC), State of Health (SOH), and Remaining Useful Life (RUL) are the three popular evaluation parameters[2]. For example, the SOH

of a battery is 100% at the beginning. When the SOH is less than 80%, the battery must be replaced or recycled to ensure continued safety, which shows that accurately estimating the battery SOH is necessary[3].

The current estimation methods for the SOH are divided into the following two main approaches: model-based estimation methods and data-driven estimation methods[4]. Ref.[5] established the Thevenin battery model to analyse and verify the declining trend of $LiFePO_4$ batteries. Ref.[6] established a dual-Kalman filter (Dual-EKF, DEKF) battery model for SOH estimation. However, due to the special chemical reaction inside the battery, establishing a model is difficult. In addition, this method would not applicable when studying complex conditions. The current research that is based on data-driven methods mainly includes the point estimation method and probability density estimation method. For example, Ref.[7] used a support vector machine (SVM) to estimate the battery SOH based on capacity. Ref.[8] established a hybrid model based on empirical mode decomposition (EMD), grey relation analysis (GRA), and deep recurrent neural network (RNN) for battery SOH predictions. These data-driven methods do not need to consider the internal mechanism of the battery, so they are the current mainstream research[9]. Therefore, this method will be used for research in this paper.

When using the data-driven method to estimate the SOH, a large amount of data is needed to train the model. The existing data for estimating the SOH mainly come from three sources. The first is public data, such as the National Aeronautics and Space Administration (NASA), which is a stable dataset from a rigorous experimental environment[8, 10]. The second is the data obtained from the charge–discharge cycle experiments in the laboratory. Due to the stable environment, the data are also relatively stable[7, 11]. The third is the data that is obtained when the car is running. For example, Ref.[12] used data collected on a big data platform to estimate the SOH by deep learning with a feedforward neural network (FFNN). The complex external environment and driver behaviour can affect the battery parameters[13], and the data obtained in the laboratory are mostly from the cell[14]. Therefore, using the third data for this study is more meaningful. This study uses a one-year historical dataset of vehicles with different driving behaviours.

One of the keys to the data-driven battery SOH estimation is selected health features. With the increase in battery charge and discharge times and the accumulation of shelf time, the solid electrolyte interface (SEI) and electrolyte inside the lithium-ion battery are degraded, and lithium ions are precipitated, resulting in a gradual decrease in battery capacity and a gradual increase in internal resistance. Therefore, the state of health of the battery can be defined in terms of capacity and internal resistance. Since directly calculating the capacity or internal resistance is difficult, looking for closely related and directly measurable battery parameters and using them as the health characteristics to calculate the capacity is usually necessary. The battery contains various parameters, such as current, voltage, and temperature. The parameters are a typical nonlinear multidimensional time-series data. They can be selected by IC/DV analysis or statistical methods[15, 16]. For example, Ref.[17] extracted health features from the voltage response under a specific current pulse test; Ref.[18] studied a state-of-health estimator that is based on multiple health indicators and machine learning to estimate the SOH; and Ref.[19] proposed an ageing pattern recognition method based on open-circuit voltage matching analysis to analyse ageing mechanisms and extract health features. Because the charge–discharge cycle of electric vehicles is far less regular than that under laboratory conditions, the health characteristics

extracted from the above literature[15-19] may not accurately reflect the battery SOH. Ref. [20] only has one angle of feature extraction, and the research results are not comparable. Additionally, the selection of health features needs to be adjusted to achieve the optimal estimation under different application requirements, but there are few studies on this topic in the literature. Thus, health features are extracted from two dimensions in this paper.

Another key to data-driven battery SOH estimation is the selection of models. The effectiveness of the model greatly depends on the quality and size of the dataset. When dealing with the dataset, back propagation (BP), a common neural network, has the problem of gradient disappearance and gradient explosion[21,22]. To address this problem, RNNs have been developed. A recurrent neural network takes sequence data as input, performs recursion in the evolution direction of the sequence, and connects all nodes in a chain. At the same time, to solve the problem of long-term dependence, the following two RNN variants with good effects were proposed in 1997 and 2014: long short-term memory networks (LSTM) and gated recurrent unit networks (GRU). Some researchers have already applied these algorithms to the SOH estimation and obtained good results[23]. Ref.[11,24] established LSTM to predict the SOH. Ref.[10] designed a variable-length short-term memory neural network (AST-LSTM-NN) to estimate the battery SOH. Therefore, in this paper, a popular model, the recurrent neural network, will be used to verify and analyse the effectiveness of this strategy. The models employed in the above literature can effectively address these deficiencies, but most are validated under specific operating conditions or in a static laboratory environment rather than real vehicle data. In this study, the data generated during the operation of the actual vehicle are selected, which are affected by the weather, driving conditions, and driving behaviour of the driver. Therefore, these models are used in this paper to study vehicle data for SOH estimations.

Based on the above analysis, the main research contents of this paper are as follows. First, we analyse the method for calculating the SOH. Second, the health features are extracted from two different dimensions, and the health features with a high correlation coefficient with the capacity are selected for subsequent model training. Finally, based on different feature selections, the state of health is estimated under different recurrent neural network models. The flowchart of the battery SOH estimation is shown in Figure 1.

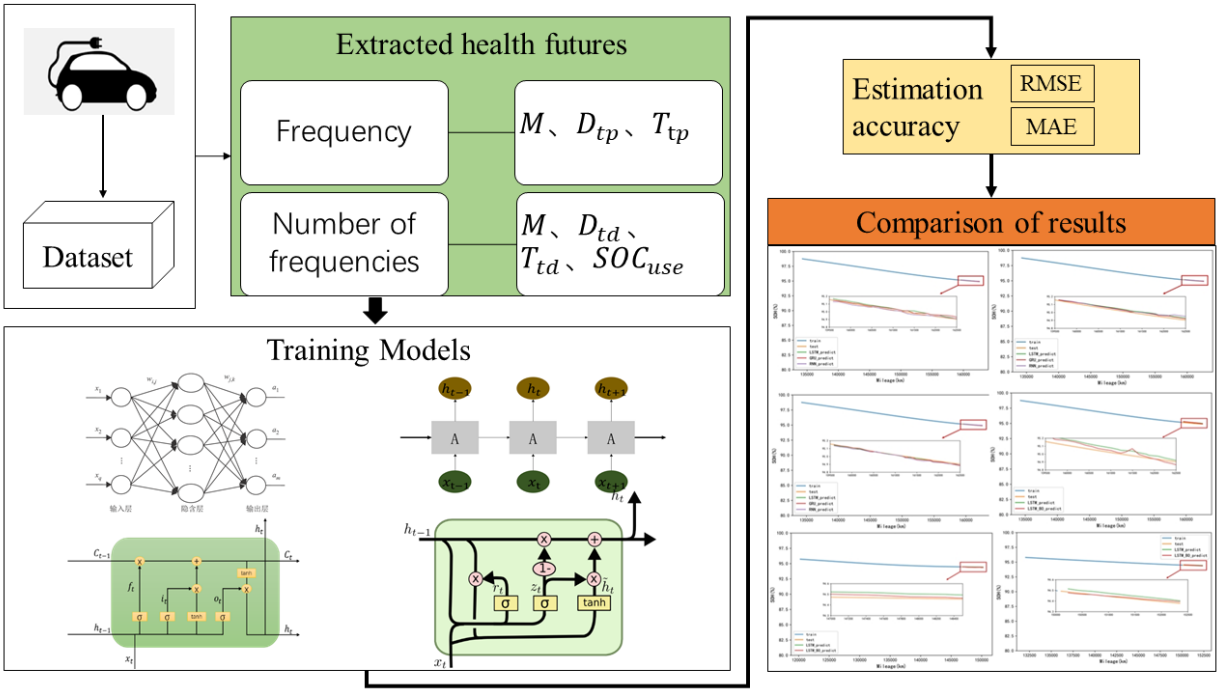


Figure 1. Flowchart of the battery SOH estimation

2. MAIN RESEARCH CONTENT

2.1. Method of calculating SOH

Due to the internal resistance of the battery changing very little when it is in use[25], accurately measuring it is difficult. Therefore, studies are mostly carried out in terms of capacity[7, 8, 26]. When capacity is used to characterize the SOH, the battery SOC can be updated in real time by the ampere-hour integration method to obtain the capacity[27, 28]. The formula for estimating the SOH is as follows:

$$\mathrm{SOH} = \frac{\mathrm{C_c}}{\mathrm{C_R}} * 100\% = \frac{\frac{\mathrm{Q_C(i)}}{\mathrm{\Delta SOC(i)}}}{\mathrm{C_R}} * 100\% = \frac{\frac{\int_{\mathrm{t_s}}^{\mathrm{t_e}} \mathrm{I(t)dt}}{\mathrm{SOC(t_e)-SOC(t_s)}}}{\mathrm{C_R}} * 100\% \quad (2.1)$$

where C_c represents the current maximum available capacity of the battery; C_R represents the rated capacity of the battery; $Q_C(i)$ represents the cumulative charging capacity of the *i-th* charging segment; $\Delta SOC(i)$ represents the amount of SOC change for the *i-th* charging segment; $I(t)$ represents the current of the battery pack at moment t; $SOC(t_s)$ and $SOC(t_e)$ represent the start SOC and end SOC of the *i-th* charging segment, respectively; and t_s and t_e represent the start time and end time of the charging segment, respectively.

2.2. Health Features Selection and Correlation Analysis

The existing operating data can be directly calculated by Formula (2.1). Since the collected operating data of the battery are affected by the air temperature, the temperature is initially corrected[29].

Since the data used in the study are nonlinear series data, the linear regression algorithm should be used to qualitatively characterize the decreasing trend of its capacity. This algorithm is designed to eliminate the capacity fluctuation and extract the label capacity value C_L, which is the training metric for subsequent machine learning models.

For data-driven models, the choice of input features has a large impact on the estimation results. Factors such as the battery operating temperature, charge–discharge current rate, and discharge depth[30] will affect battery ageing to a certain extent. To quantify these factors, health features will be extracted in this study from two dimensions. The frequency-based health feature mainly represents the changes in the battery operating temperature and charge–discharge rate. The number of frequency-based health features quantify the battery's historical working conditions.

2.2.1. Frequency-based health feature

(1) The battery charge–discharge data obtained under laboratory conditions are developed based on the number of cycles, but the battery of the actual operating vehicle is not as fully charged and discharged as under laboratory conditions. Therefore, extracting the accumulated mileage M as a health feature to characterize the SOH is necessary.

(2) To intuitively characterize the current charge–discharge rate, the multiple discharge segments between the two charging segments are divided according to different discharge rates (0-0.1C, 0.1-0.5C, 0.5-1C, >1C), and then the cumulative number of occurrences are counted and recorded as $[C(i)_1, C(i)_2, C(i)_3, C(i)_4]$, where i represents the *i-th* charging segment. The discharge rate frequency is calculated according to Formula (2.2). In general, the statistical characteristic distribution of the discharge multiplicity frequency D_{tp} is used as a health feature of the SOH, as shown in Formula (2.3).

$$P(i)_j = \frac{C(i)_j}{\sum_{j=1}^{4} C(i)_j}, j = 1,2,3,4 \quad (2.2)$$

$$D_{tp} = [P_1, P_2, P_3, P_4] \quad (2.3)$$

(3) The maximum and minimum values of the battery temperature can reflect the thermal distribution inside the battery pack, and the average temperature of the battery can classify the thermal environment of the battery. Considering that the storage function of the battery will be different at different temperatures, the average temperature between the two charging segments is divided according to different intervals (0-10℃, 10-20℃, 20-30℃, 30-40℃, >40℃). Then, the accumulated times are counted and recorded as $[T(i)_1, T(i)_2, T(i)_3, T(i)_4, T(i)_5]$. The ambient temperature frequency of the discharge segment is calculated according to Formula (2.4). In general, the statistical characteristic distribution of the ambient temperature frequency T_{tp} is used as a health feature of the SOH, as shown in Formula (2.5).

$$PP(i)_j = \frac{T(i)_j}{\sum_{j=1}^{5} T(i)_j}, j = 1,2,3,4,5 \quad (2.4)$$

$$T_{tp} = [PP_1, PP_2, PP_3, PP_4, PP_5] \quad (2.5)$$

In summary, the frequency-based health features include the accumulated mileage M, the statistical characteristic distribution of the discharge multiplicity frequency D_{tp} and the statistical characteristic distribution of the ambient temperature frequency T_{tp}.

2.2.2. Number of frequency-based health features

Since the accuracy of the frequency-based health features is affected by the driving duration of the electric vehicle, when the charging or driving duration is not enough, the accuracy of the health features extracted based on the frequency will be poor. Therefore, in this study, we further propose the number of frequency-based health features. The difference is that the historical operating conditions of the battery are quantified over a length of time.

(1) Similar to the frequency-based health feature, we extract the accumulated mileage M as a health feature to characterize the SOH.

(2) Similar to the processing of the statistical characteristic distribution of the discharge multiplicity frequency, the number of frequency of the discharge multiplier before the *i*-th charging segment is counted separately. Then, the frequency of the discharge multiplier of all the discharging segments before this charging segment is calculated according to Formula (2.6). In general, the statistical characteristic distribution of the discharge multiplicity number of frequencies D_{td} is used as a health feature of the SOH, as shown in Formula (2.7).

$$S(i)_j = \sum_{k=1}^{i} C(k)_j, k = 1,2,3 \cdots i, j = 1,2,3,4 \quad (2.6)$$

$$D_{td} = [S_1, S_2, S_3, S_4] \quad (2.7)$$

(3) Similar to the processing of the statistical characteristic distribution of the ambient temperature frequency, the number of frequency of the ambient temperature before the *i-th* charging segment are counted separately. Then, the number of frequencies of the ambient temperature of all discharging segments before this charging segment is calculated according to Formula (2.8). In general, the statistical characteristic distribution of the ambient temperature number of frequencies T_{td} is used as a health feature of the SOH, as shown in Formula (2.9).

$$SS(i)_j = \sum_{k=1}^{i} T(k)_j, k = 1,2,3 \cdots i, j = 1,2,3,4,5 \quad (2.8)$$

$$T_{td} = [SS_1, SS_2, SS_3, SS_4, SS_5] \quad (2.9)$$

(4) The entire SOC is divided into different intervals (0-20%, 20-40%, 40-60%, 60-80%, 80-100%), and the SOC number of frequency distribution of each charging segment is counted and recorded as $[D(i)_1, D(i)_2, D(i)_3, D(i)_4, D(i)_5]$. The SOC number of frequency distributions before the *i-th* charging segment is counted separately. Then, the SOC number of the frequency distribution before this charging segment is calculated according to Formula (2.10). In general, the statistical characteristic distribution of the SOC number of frequencies SOC_{use} is used as a health feature of the SOH, as shown in Formula (2.11).

$$DD(i)_j = \sum_{k=1}^{i} D(k)_j, k = 1,2,3 \cdots i, j = 1,2,3,4,5 \quad (2.10)$$

$$SOC_{use} = [DD_1, DD_2, DD_3, DD_4, DD_5] \quad (2.11)$$

In summary, the number of frequency-based health features includes the accumulated mileage M, the statistical characteristic distribution of the discharge multiplicity number of frequencies D_{td}, the statistical characteristic distribution of the ambient temperature number of frequencies T_{td} and the statistical characteristic distribution of the used SOC number of frequencies SOC_{use}.

2.2.3. Correlation Analysis

The selection of health features has a great impact on the estimation accuracy of the model. Common methods for correlation analysis are as follows: Pearson correlation [31], grey correlation [32], Spearman correlation coefficient, and maximal information coefficient (MIC). Among them, the MIC has a good effect on processing nonlinear data such as batteries, has less complexity and more stability and is more universal compared with other methods. Therefore, this study adopts this method to analyse the correlation between the health features and capacity C_L. The calculation formula of the MIC is shown in Formula (2.12), the correlation analysis results are shown in Table 1, with the results maintained to six decimal places.

$$\mathrm{mic}(x;y) = \max_{a*b<B} \frac{I(x;y)}{\log_2 min(a,b)} \quad (2.12)$$

where a and b are the number of division lattices in x and y, respectively.

2.3. Estimation Model

2.3.1. Backpropagation neural network (BPNN)

The backpropagation neural network (BPNN) is composed of the input layer, hidden layer, and output layer neurons and the transfer function between neurons. Through the transmission of signals between neurons, the mapping relationship between the input parameters and output parameters is constructed. The standard structure is shown in Figure 2, which includes the forwarding transfer and error reverse transfer. There are **q** input layer nodes, **w** hidden layer nodes, and **m** output layer nodes. The update of the connection weight $\boldsymbol{w_{ij}}$ between the *i-th* neuron in the input layer and the *j-th* neuron in the hidden layer is shown in Formula (2.13).

$$\boldsymbol{w_{ij} = \Delta w_{ij} + w_{ij}(t-1), i = 1, 2, \ldots, q, j = 1, 2, \ldots, w} \quad (2.13)$$

where $w_{ij}(t-1)$ represents the weights after previous training and Δw_{ij} represents the weight correction. The formula is shown in Formula (2.14).

$$\Delta w_{ij} = -\eta \frac{\partial E_k}{\partial w_{ij}} \quad (2.14)$$

where η represents the learning rate and E_k represents the error between the predicted output and the expected output at the *k-th* training.

The connection weight w_{jk} between the *j-th* neuron in the hidden layer and the *k-th* neuron in the output layer is the same as above.

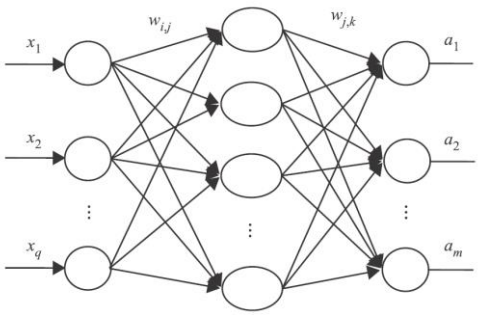


Figure 2. The structure of a standard BPNN

2.3.2. Recurrent Neural Network (RNN)

The recurrent neural network (RNN) can process long sequences of data. The structure of a standard RNN is shown in Figure 3. The current input state x_t and the previous state variable h_{t-1} are inputted into A, and the state vector h_t is updated. Then, the next input x_{t-1} is combined to input into A, and the next state vector h_{t+1} is updated. In general, the RNN is a process of combining the previous state with the current input and continuously training the parameter A to achieve the optimal value. However, this algorithm cannot take into account the previous information when the training reaches a later stage; thus, it cannot handle long-term series of data.

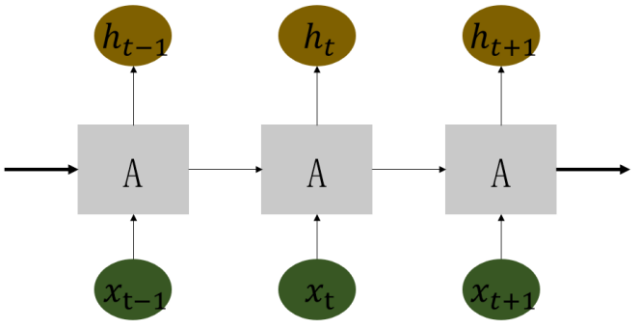


Figure 3. The structure of a standard RNN

2.3.3. Long Short-Term Memory Networks (LSTM)

Long short-term memory (LSTM) networks are a variant of RNNs. It solves the problem of vanishing gradients and the inability to solve long-time series data. The specific structure is shown in Figure 4. The LSTM consists of an input gate, output gate, forget gate, and memory unit C_t. The state of the input gate i_t is determined by the input at the current moment x_t and the output of the hidden layer at the previous moment h_{t-1}, and the memory unit to be updated is generated $\tilde{C}_t$ at the same time. The formula is as follows:

$$i_t = \sigma(W_{ix} \cdot x_t + W_{ih} \cdot h_{t-1} + b_i) \quad (2.15)$$

$$\tilde{C}_t = i_t \tanh(W_{zx} x_t + W_{zh} h_{t-1} + b_z) \quad (2.16)$$

where $W_{ix}, W_{ih}, W_{zx}, W_{zh}$ represent the weight matrices, b_i, b_z represent the bias vectors, and σ and $tanh$ represent the activation functions.

The forget gate further determines what information needs to be saved at the last moment in the memory unit C_t, and the formula is as follows:

$$f_t = \sigma(W_{fx} \cdot x_t + W_{fh} \cdot h_{t-1} + b_f) \quad (2.17)$$

where f_t represents the output vector of the forget gate at the current moment, W_{fx}, W_{fh} represent the weight matrices, and b_f represents the bias vector.

The output vector of the forget gate f_t and the memory unit to be updated $\tilde{C}_t$ are used to update the memory unit C_t, and the formula is as follows:

$$C_t = f_t C_{t-1} + i_t tanh(W_{zx} x_t + W_{zh} h_{t-1} + b_z) \quad (2.18)$$

where W_{zx} and W_{zh} represent the weight matrices and b_z represents the bias vector.

The output gate determines the update of the system state by the internal state. It combines the output of the previously hidden layer h_{t-1} with the current state x_t to obtain the output o_t and then combines it with C_t to obtain the current state output h_t. The formula is as follows:

$$o_t = \sigma(W_{ox} x_t + W_{oh} h_{t-1} + b_o) \quad (2.19)$$

$$h_t = o_t tanh(C_t) \quad (2.20)$$

where W_{ox} and W_{oh} represent the weight matrices and b_o represents the bias vector.

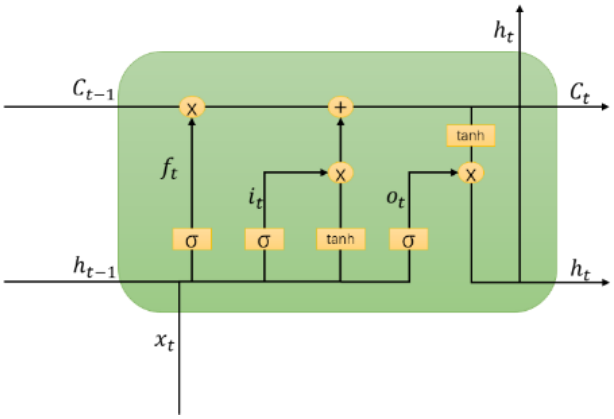


Figure 4. The structure of the LSTM

2.3.4. Gate Recurrent Unit (GRU)

The gated recurrent unit (GRU) is a new variant of RNNs proposed after the LSTM. It omits the small contributing gates and their corresponding weights in the LSTM. The GRU uses only the following two gates to control the output of the neural network: the reset gate and the update gate. Therefore, the GRU simplifies the structure and improves the training efficiency. The specific structure is shown in Figure 5. The reset gate r_t and the update gate z_t are jointly determined by the current input state x_t and the hidden state at the previous moment h_{t-1}. In addition, the candidate set of the current state $\tilde{h}_t$ is generated at the same time, with the formula as follows:

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \quad (2.21)$$

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_r) \quad (2.22)$$

$$\tilde{h}_t = \tanh(W_h[r_t * h_{t-1}, x_t] + b_n) \quad (2.23)$$

where W_r and W_z represent the weight matrices, b_r and b_n represent the bias vectors, and σ and $tanh$ represent the activation functions.

The reset gate is used to control how much information from the previous state is written into the current candidate set $\tilde{h}_t$, and the update gate is used to control the extent to which the state information of the previous moment is brought into the current state. Finally, the hidden state of the current moment h_t is calculated. The formula is as follows.

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (2.24)$$

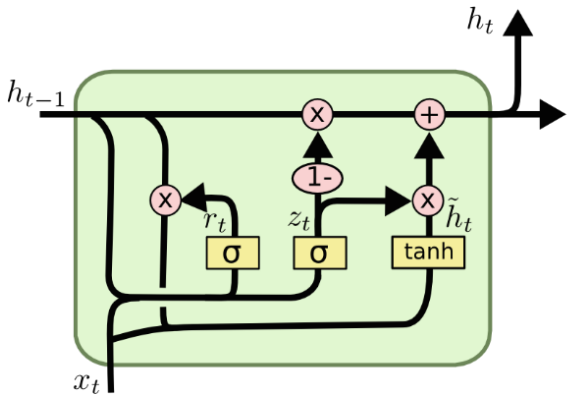


Figure 5. The structure of the GRU

2.3.5 Parameter settings of the model

In this study, the sequential model is imported to implement a linear stack of multiple network layers through Python. In the process of network training, the Adam optimization algorithm is applied to continuously update its weights. The range of the learning rate is [0.00001, 0.001]. To evaluate the estimation accuracy of the model under different hyperparameters, we set a fixed random number to ensure that the weights of the RNN model and GRU model are initialized under the same conditions. The dropout layer is used to reduce the complexity of the network. The mean absolute error (MAE), which can be used to represent the absolute error between the predicted value and the true value, is used as the loss function, and the formula is shown in Formula (2.25). Three schematic diagrams of the SOH estimation model for Li-ion battery packs are shown in Figure 6.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{C}(i) - C_L(i)| \quad (2.25)$$

where $\hat{C}(i)$ is the predictive value, and $C_L(i)$ is the true value.

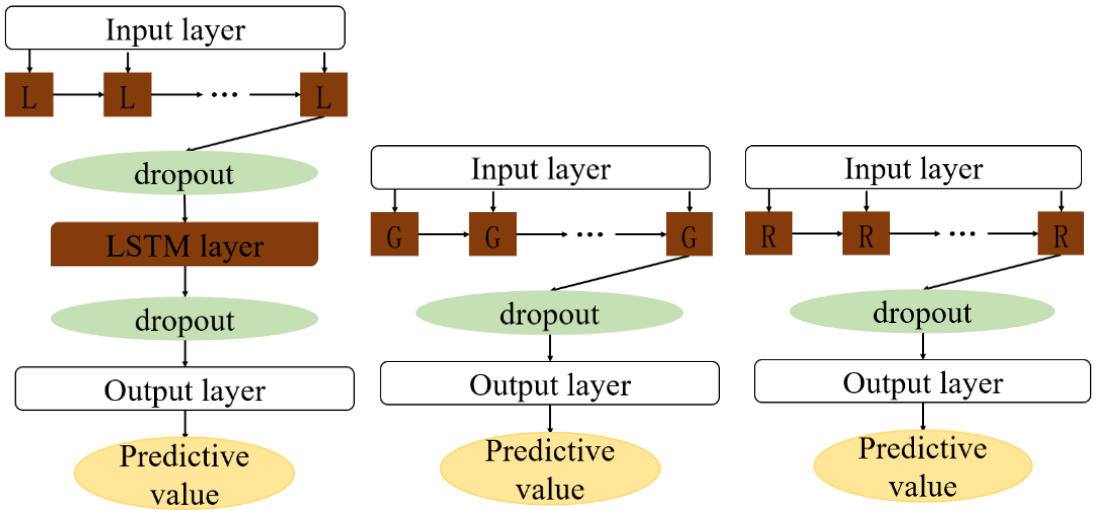


Figure 6. Three schematic diagrams of the SOH estimation model for Li-ion battery packs

3. RESULTS AND ANALYSIS

3.1. Dataset Introduction

The data used in this study come from the real environment of a car. It was sampled at a 20-second interval and lasted for one year. The rated capacity is 280 Ah and includes more than 20 parameters, such as the vehicle operating location, sampling time point, total voltage, total current, state of charge, cell maximum voltage, cell minimum voltage, maximum temperature, minimum temperature, and mileage. Moreover, these data are multi-dimensional time-series data. Due to the harsh and changeable external environment of the car during actual operation and the lithium-ion battery being easily affected by the ambient temperature[33], the data are far less stable than those collected in the laboratory. Therefore, some necessary pre-processing should be performed to improve the training effect of the model and the accuracy of the estimation[34, 35]. Different data should make different pre-processing rules. For the data in this study, a brief description of the processing procedure is given below. We need to distinguish between charged and discharged fragments. Then, we clean the records with zero battery current and the fragments with too many missing values, as well as delete clips that are too short. Then, we standardize data such as time, voltage, current, temperature, SOC, mileage, and clean abnormal data that do not meet the normal range of the battery.

3.2. Results and Analysis

The flowchart of the SOH estimation is shown in Figure 1. It contains four steps, including data acquisition, feature extraction and selection, model training, and SOH estimation. A one-year historical dataset of vehicles with different driving behaviours is used in this study. These recorded data, such as voltage and current, are inputs for the frequency-based and the number of frequency-based Health Indicators (HI) extractions. Then, three different feature selection methods are adopted to select each subset. Afterwards, different machine learning algorithms are used for model training, including the BPNN model, RNN model, LSTM model, and GRU model. Finally, the SOH is estimated based on different health features and models, and the accuracy, robustness, and computational efficiency are evaluated to demonstrate the estimation performance of each strategy.

The health features analysed above should be further screened using the maximal information coefficient. Then, we find health features that are highly correlated with the capacity to train the model. The results are as follows:

Table 1. The correlation analysis results between the frequency-based health features and C_L

feature	coefficient	feature	coefficient
M	0.999999	PP_1	0.893752
P_1	0.999997	PP_2	0.999985
P_2	0.997582	PP_3	0.999985
P_3	0.999892	PP_4	0.999624
P_4	0.999999	PP_5	0.954012

Table 2. The correlation analysis results between the number of frequency-based health features and C_L

feature	coefficient	feature	coefficient	feature	coefficient
M	0.999999	SS_1	0.409192	DD_1	0.995649
S_1	0.823188	SS_2	0.990492	DD_2	0.998995
S_2	0.845051	SS_3	0.566168	DD_3	0.999985
S_3	0.369753	SS_4	0.999985	DD_4	0.999999
S_4	0.536302	SS_5	0.793339	DD_5	0.999999

By analysing the data in Table 1 and Table 2, we can see that the calculated results are in the range of [0,1], and the correlation coefficients of most health characteristics are higher than 0.7. Therefore, the follow-up SOH estimation study selects the dataset composed of health characteristics with an MIC higher than 0.7.

We take 90% of the data as the training set to train the nonlinear relationship between battery pack capacity degradation and health features and take the last 10% of the data as the validation set and test set to verify the accuracy and generalization ability of the model. To objectively evaluate the prediction effect of each recurrent neural network model, the root mean square error (RMSE), which can be used to represent the difference between the predicted value and the true value, is used to calculate the accuracy of the estimated results. It can be used to reflect the discrete degree in the sample, with the formula is shown in Formula (3.1).

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\hat{\mathrm{C}}(\mathrm{i}) - \mathrm{C_L}(\mathrm{i}))^2} \qquad (3.1)$$

where $\hat{C}(i)$ is the predictive value, and $C_L(i)$ is the true value.

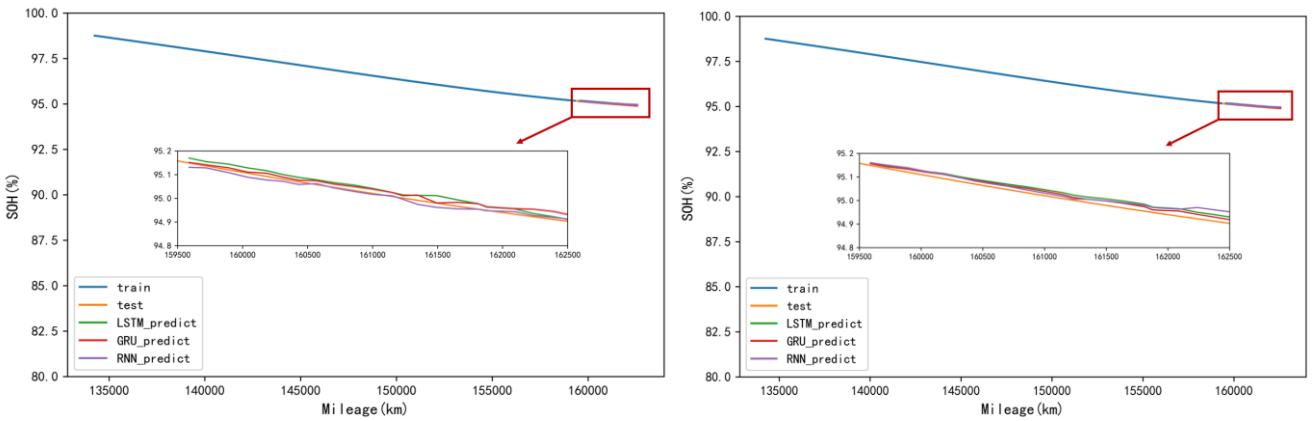
The algorithms currently include artificial neural networks (ANNs)[36], support vector machines (SVMs)[7,37], relevance vector machines (RVMs)[38], and Gaussian process regression (GPR)[39]. Ref.[40] performed correlation analysis on the health characteristics and established the SOH estimation model combined with the SVM algorithm. Ref.[39] proposed a GPR battery SOH estimation model based on the charging curve. However, the methods described above still have some disadvantages. First, these methods still depend on the sampling accuracy and calculation accuracy of experimental data; when the data quality is poor, these methods will perform poorly. Second, the validation data of these methods are usually measured in a constant current/temperature or laboratory environment. When using real-world data to estimate battery SOHs, their stability and practicality need to be verified. The

degradation of a lithium-ion battery can cover hundreds of cycles or more, and the methods proposed above do not reflect the long-term dependence of SOH degradation.

Next, we will compare and analyse the estimation results from the perspective of combining different feature selections and different models. In this study, the real running data of a certain vehicle were selected to verify the proposed method, and the verification results are shown in Figures 7~9. When processing a large amount of data, such as batteries, the BP neural network will have obvious problems of slow convergence and insufficient accuracy. Table 3 shows that the SOH estimation accuracy is significantly reduced compared with other neural networks; thus, this model will not be used in subsequent research. Combining Table 3 and Figure 10, we can see that the combination-based health feature (frequency-based health feature and number of frequencies-based health feature) has the highest estimation accuracy. Compared with the frequency-based and number of frequency-based health features, the MAE accuracy under the LSTM model is increased by 78% and 58%, and the RMSE accuracy is increased by 50% and 37%, respectively. The accuracies of the GRU model and RNN model are similar to that of the LSTM model, because the combined features not only reflect the battery's ageing information and quantify the battery's historical working conditions but also add features that reflect the driver's behaviour.

Table 3. Comparison of the accuracy of the SOH estimation results based on different feature selection and models

		LSTM	GRU	RNN	BP
Frequency	MAE（%）	0.023	0.018	0.042	0.054
	RMSE（%）	0.069	0.057	0.089	0.091
Number of Frequencies	MAE（%）	0.012	0.010	0.035	0.047
	RMSE（%）	0.054	0.042	0.072	0.084
Combination	MAE（%）	0.005	0.008	0.032	0.039
	RMSE（%）	0.034	0.037	0.047	0.059



(A) (B)

Figure 7. (A) Comparison of the results of different models based on the frequency of health features (B) Comparison of the results of different models based on the number of frequencies of health features

At the same time, no matter which feature selection is used, the SOH estimation accuracy based on the LSTM model and the GRU model is not much different. However, both are higher than the estimation accuracy based on the RNN model, which shows that, as a variant of the RNN, the LSTM model and the GRU model have better results for estimating the SOH. As seen from Figure 11, the training time of the GRU model under the three healthy feature selections is reduced by 22%, 18%, and 10% compared with the LSTM model, because the structure of the GRU model is simpler than that of the LSTM model. Therefore, when the amount of data processed is large, the GRU model will have better results.

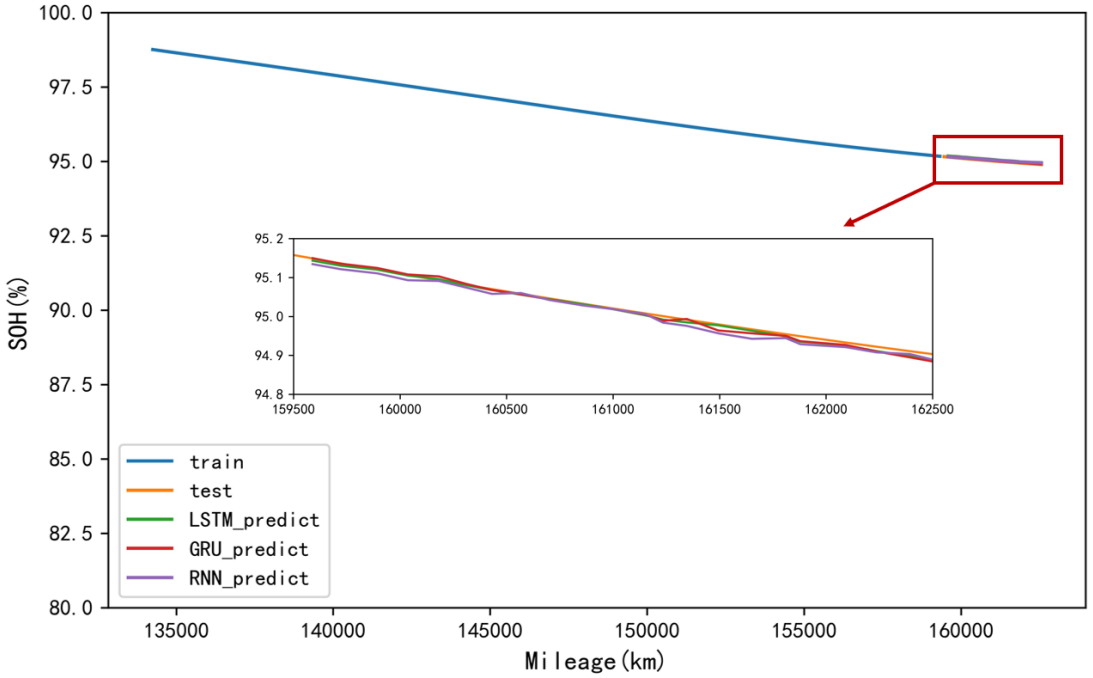


Figure 8. Comparison results of different models based on the combined health features

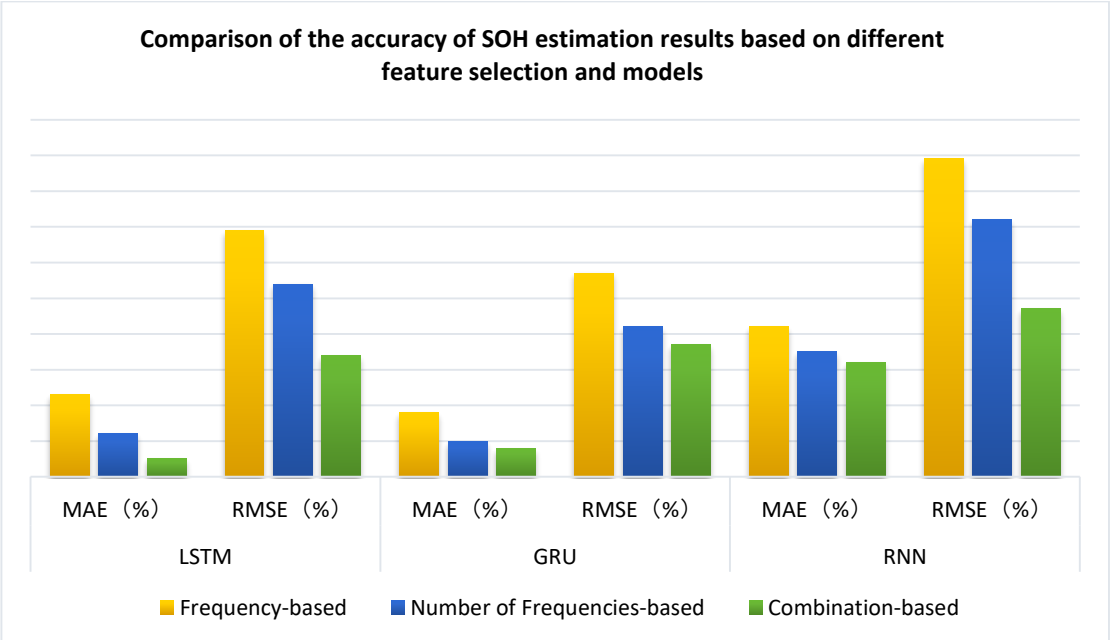


Figure 9. Comparison of the accuracy of the SOH estimation results based on different feature selections and models

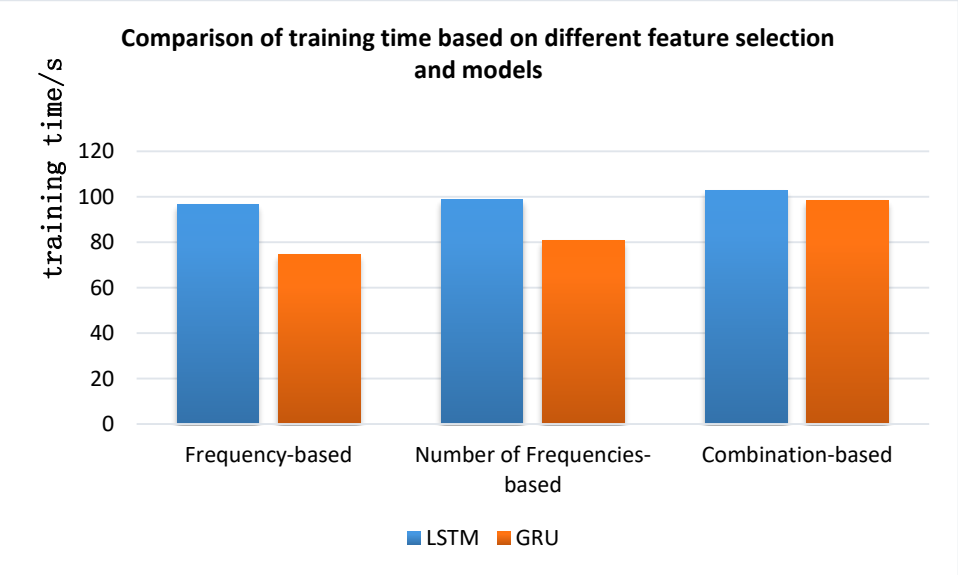


Figure 10. Comparison of the training time based on different feature selections and models

From the above analysis, we can see that the estimation accuracy of the LSTM model based on combined feature selections is slightly higher than that of the GRU model. Compared with the GRU model, the LSTM model has a more complex structure and a larger number of hyperparameters. Therefore, when the dimension of the processed data features is high, the LSTM model has an advantage over the GRU model. However, at the same time, there will be problems that it is difficult to ensure the accuracy and generalization by adjusting parameters based on manual experience. Next, we will conduct a separate study on the LSTM model based on combined features. The Bayesian optimization algorithm (BO) is added to the LSTM model, and the actual running data of three vehicles were selected for comparison and verification. The data of three vehicles are randomly selected to verify this method. The Bayesian optimization algorithm is a very popular hyperparameter optimization algorithm, which can reduce the attempts to select hyperparameters to obtain the maximum optimal solution. In short, it can improve the accuracy of the model and solve the limitations of manual parameter adjustment. By analyzing Figure 11 and Table 4, we can see that the MAE and RMSE accuracy of the LSTM model optimized by the algorithm for car A is improved respectively by 55% and 32% compared without optimization. Car B and Car C also have the same effect.

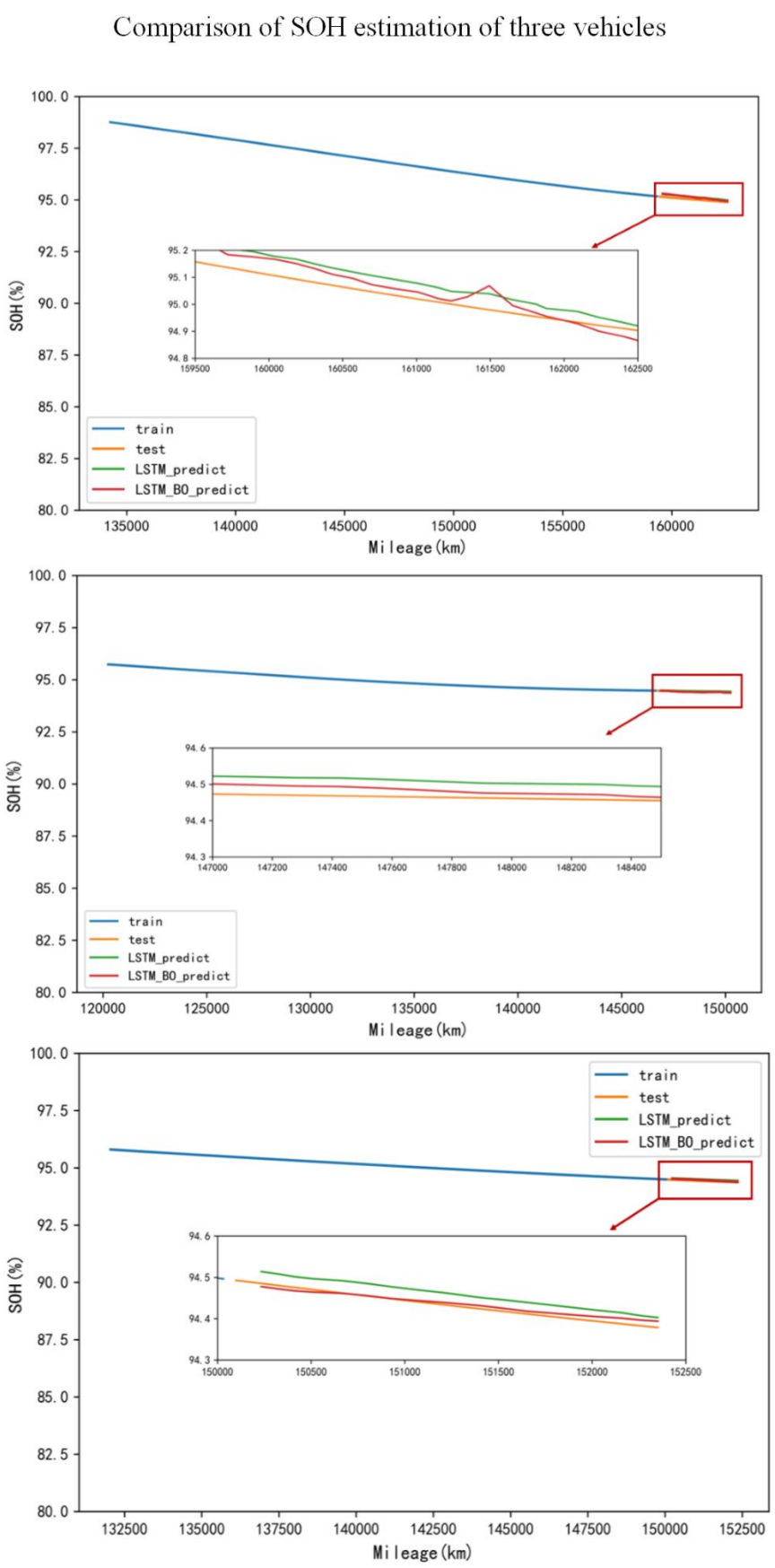


Figure 11. Comparison of the SOH estimation of three vehicles based on the combined features and LSTM-BO model

Table 4. Comparison of the SOH estimation of three vehicles based on the combined features and LSTM-BO model

	error	**A**	**B**	**C**
LSTM model	MAE（%）	0.011	0.015	0.012
	RMSE（%）	0.025	0.042	0.038
LSTM-BO model	MAE（%）	0.005	0.014	0.009
	RMSE（%）	0.017	0.032	0.024

4. CONCLUSIONS

In the era of big data, large amounts of data generated by the actual operation of the car provides many possibilities for the estimation of the SOH. Three different feature dimensions and three different estimation models are combined in this paper to estimate the battery state of health and analyse the accuracy of the estimation results. The results show that the strategy is effective and that the capacity decline of the car is in line with the normal mechanism of battery ageing. At the same time, the results show that the LSTM model and GRU model with combined features have higher accuracies. Since the training time of the GRU model is shorter, the GRU model is more advantageous when dealing with a large amount of data. Since the LSTM model has many hyperparameters, it is more advantageous when processing data with high feature dimensions. Finally, since the LSTM model has many hyperparameters, a simple hyperparameter optimization is performed. The results show that the estimation accuracy of the optimized model is significantly improved.

In the future, we aim to overcome the limitations of this study, mainly by using more kinds of models, such as the temporal convolutional network and Gaussian process regression, to further improve the applicability of the proposed strategy.

ACKNOWLEDGMENTS
This work was financially supported by Zhejiang market supervision bureau “chick eagle project” (CY2022362) fund.

References

1. M.S. Hossain, M.A. Hannan, T.F. Karim, A. Hussain, M.H.M. Saad, A. Ayob, M.S. Miah and T.M.I. Mahlia, *J. Cleaner Prod*., 292 (2021) 126044.
2. H.X. Tian, P.L. Qin, K. Li and Z. Zhao, *J. Cleaner Prod*., 261 (2020) 120813.
3. Y. Li, K.L. Liu, A.M. Foley, A. Zulke, M. Berecibar, E. Nanini-Maury, J. Van Mierlo and H.E. Hoster, *Renewable Sustainable Energy Rev*., 113 (2019) 109254.
4. B.Y. Liu, X.P. Tang and F.R. Gao, *Electrochim. Acta*., 344 (2020) 136098.
5. S. Panchal, J. McGrory, J. Kong, R. Fraser, M. Fowler, I. Dincer and M. Agelin-Chaab, *Int. J. Energy Res.*, 41 (2017) 2565-2575.
6. N. Wassiliadis, J. Adermann, A. Frericks, M. Pak, C. Reiter, B. Lohmann and M. Lienkamp, *J. Energy Storage.,* 19 (2018) 73-87.
7. X.N. Feng, C.H. Weng, X.M. He, X.B. Han, L.G. Lu, D.S. Ren and M.G. Ouyang, *IEEE Trans. Veh.*

Technol., 68 (2019) 8583-8592.
8. J.C. Chen, T.L. Chen, W.J. Liu, C.C. Cheng and M.G. Li, *Adv. Eng. Inf*., 50 (2021) 101405.
9. L.F. Wu, X.H. Fu and Y. Guan, *Appl. Sci.-Basel*., 6 (2016) 166.
10. P.H. Li, Z.J. Zhang, Q.Y. Xiong, B.C. Ding, J. Hou, D.C. Luo, Y.J. Rong and S.Y. Li, *J. Power Sources*, 459 (2020) 228069.
11. Y.T. Wu, Q. Xue, J.W. Shen, Z.Z. Lei, Z. Chen and Y.G. Liu, *IEEE Access*., 8 (2020) 28533-28547.
12. L.J. Song, K.Y. Zhang, T.Y. Liang, X.B. Han and Y.J. Zhang, *J. Energy Storage*, 32 (2020) 101836.
13. J.C. Hong, Z.P. Wang, W. Chen, L.Y. Wang, P. Lin and C.H. Qu, *J. Cleaner Prod*., 294 (2021) 125814.
14. H.B. Ren, Y.Z. Zhao, S.Z. Chen and T.P. Wang, *Energy*, 166 (2019) 908-917.
15. Y. Li, C.F. Zou, M. Berecibar, E. Nanini-Maury, J.C.W. Chan, P. van den Bossche, J. Van Mierlo and N. Omar, *Appl. Energy*., 232 (2018) 197-210.
16. Y. Gao, J.C. Jiang, C.P. Zhang, W.G. Zhang and Y. Jiang, *J. Power Sources*., 400 (2018) 641-651.
17. J.H. Meng, L. Cai, G.Z. Luo, D.I. Stroe and R. Teodorescu, *Microelectron. Reliab*., 88 (2018) 1216-1220.
18. H.H. Pan, Z.Q. Lu, H.M. Wang, H.Y. Wei and L. Chen, *Energy*., 160 (2018) 466-477.
19. J.Q. Tian, R.L. Xu, Y.J. Wang and Z.H. Chen, *Energy*., 221 (2021) 119682.
20. Z.G. He, X.Y. Shen, Y.Y. Sun, S.C. Zhao, B. Fan and C.F. Pan, *J. Energy Storage*, 41 (2021) 102867.
21. B. Saha, K. Goebel, and J. Christophersen, *Trans. Inst. Meas. Control*, 31(2009) 293-308.
22. D.E. Choe, H.C. Kim and M.H. Kim, *Renewable Energy*., 174 (2021) 218-235.
23. J.T. Qu, F. Liu, Y.X. Ma and J.M. Fan, *IEEE Access*, 7 (2019) 87178-87191.
24. K.F. Qian and X.T. Liu, *J. Energy Storage*, 44 (2021) 103319.
25. K. Li, P. Zhou, Y.F. Lu, X.B. Han, X.J. Li and Y.J. Zheng, *J. Power Sources*, 468 (2020) 228192.
26. C.P. Zhang, Y. Jiang, J.C. Jiang, G. Cheng, W.P. Diao and W.G. Zhang, *Appl. Energy*, 207 (2017) 510-519.
27. J. Jaguemont, L. Boulon and Y Dubé, *Appl. Energy*, 164 (2016) 99-114.
28. K. Qian, Y.Y. Li, Y.B. He, D.Q. Liu, Y. Zheng, D. Luo, B.H. Li and F.Y. Kang, *RSC Adv*, 6 (2016) 76897-76904.
29. Y. Gao, J.C. Jiang, C.P. Zhang, W.G. Zhang and Y. Jiang, *J. Power Sources*, 400 (2018) 641-651.
30. D. Yang, X. Zhang, R. Pan, Y.J. Wang and Z.H. Chen, *J. Power Sources*, 384 (2018) 387-395.
31. X.Y. Li, C.G. Yuan, and Z.P. Wang, *J. Power Sources*, 467 (2020) 228358.
32. X.Y. Li, C.G. Yuan and Z.P. Wang, *ENERGY*, 190 (2020) 116467.
33. C. Chen, R. Xiong, R.X. Yang, W.X. Shen and F.C. Sun, *J. Cleaner Prod*., 234 (2019) 1153-1164.
34. A. Wadi, M.E. Abdel-Hafez and A.A. Hussein, *IEEE Trans. Veh. Technol*., 68 (2019) 8593-8600.
35. C. Huang, Z.H. Wang, Z.H. Zhao, L. Wang, C.S. Lai and D. Wang, *IEEE Access*, 6 (2018) 27617-27628.
36. J. Wu, C.B. Zhang and Z.H. Chen, *Appl. Energy*, 173 (2016) 134-140.
37. D.T. Liu, Y.C. Song, L. Li, H.T. Liao, and Y. Peng, *J. Cleaner Prod*., 199 (2018) 1050-1065.
38. Y. Zhang and B. Guo, *ENERGIES*, 8 (2015) 12439-12457.
39. D. Yang, X. Zhang, R. Pan, Y.J. Wang and Z.H. Chen, *J. Power Sources*, 384 (2018) 387-395.
40. J.H. Meng, L. Cai, G.Z. Luo, D.I. Stroe and R. Teodorescu, *Microelectron. Reliab*. 88-90 (2018) 1216-1220.