Journal of Power Sources 623 (2024) 235417

Contents lists available at ScienceDirect

Journal of Power Sources

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

A multi-dimensional machine learning framework for accurate and efficient battery state of charge estimation

Sijing Wang [a], Meiyuan Jiao [a], Ruoyu Zhou [a], Yijia Ren [a], Honglai Liu [a,b], Cheng Lian [a,b,*]

[a] State Key Laboratory of Chemical Engineering, Shanghai Engineering Research Center of Hierarchical Nanomaterials, School of Chemical Engineering, East China University of Science and Technology, Shanghai, 200237, China

[b] School of Chemistry and Molecular Engineering, East China University of Science and Technology, Shanghai, 200237, China

HIGHLIGHTS

- Proposed a multi-dimensional machine learning framework for SOC estimation.
- Applied median filtering to remove noise from raw data, enhancing data quality.
- Employed CWT to extract time-frequency features from battery voltage signals.
- Enhanced model generalization capabilities by feature cross and random forest.
- The model was validated under various temperatures and operating conditions.

ARTICLE INFO

Keywords:
State of charge
Median filtering
Continuous wavelet transform
Feature cross
Random forest

ABSTRACT

Accurate state of charge (SOC) estimation is essential for battery safe and efficient utilization. As artificial intelligence technologies evolve, data-driven methods have become mainstream for estimating SOC. However, the technique can significantly deteriorate model performance when encountering poor or insufficient data quality. In this paper, we apply median filtering to eliminate extreme noise and utilize continuous wavelet transform to extract time-frequency features from voltage signals. Additionally, we generate novel features via feature crossing. We then apply dimensionality reduction via the random forest method to decrease computational expense. Finally, we select a convolutional neural network (CNN) as the base model to learn optimized features for more precise SOC estimation. To confirm the efficacy of our proposed method, this study compares it with CNN, long short-term memory (LSTM), bidirectional LSTM (BILSTM), and a CNN-BILSTM model combined with an attention mechanism. These comparisons are conducted under different temperatures and operating conditions. The results indicate that this method achieves a mean absolute error and a root mean square error of less than 2.89 % and 3.71 %, respectively, in SOC estimation, demonstrating superior accuracy compared to other models. This study underscores the significance of feature engineering techniques in SOC estimation.

1. Introduction

Lithium-ion batteries are extensively employed as the principal energy storage units in modern portable electronic devices and electric vehicles due to their long cycle life, low self-discharge rate, high energy density, and no environmental pollution [1,2]. However, due to the complexity of internal chemical reactions and external operating conditions' uncertainty, an efficient and reliable battery management system (BMS) is essential for monitoring battery status [3]. Among the numerous parameters of the BMS, the state of charge (SOC) directly affects the efficiency and safety of the battery [4]. Accurate assessment of SOC is essential to ensure stable system operation and extend battery life [5,6]. However, since the SOC is an internal variable that cannot be directly measured using sensors, it must be indirectly estimated through observable parameters like voltage and current [7]. Therefore, the accurate prediction of SOC faces many challenges [8].

To meet the challenges in SOC estimation, four main methods have been proposed: ampere-hour integration (AH) method, open-circuit

* Corresponding author. State Key Laboratory of Chemical Engineering, Shanghai Engineering Research Center of Hierarchical Nanomaterials, School of Chemical Engineering, East China University of Science and Technology, Shanghai, 200237, China.

E-mail address: liancheng@ecust.edu.cn (C. Lian).

https://doi.org/10.1016/j.jpowsour.2024.235417
Received 8 June 2024; Received in revised form 30 August 2024; Accepted 7 September 2024
Available online 13 September 2024

voltage (OCV) method, model-based filtering method, and data-driven method [9,10]. The AH calculates the SOC by integrating the current. Yet, this approach is susceptible to the current measurement noise and cumulative error, and thus the accuracy gradually decreases in the long-term application [11]. The OCV method estimates the SOC by measuring the relationship between the battery terminal voltage and the SOC, and although it is easy to operate, its accuracy is susceptible to changes in the battery operating state [12,13]. In contrast, model-based filtering methods attempt to predict SOC by constructing a battery model and combining it with a Kalman filtering (KF) algorithm [14]. However, since the internal reactions of the battery show obvious nonlinear characteristics during the charging and discharging cycles and the internal parameters of the battery change in real-time, this puts higher requirements on the selection of the battery model [15]. An accurate battery model can effectively simplify the estimation steps of battery SOC [16,17]. Although the pseudo-two-dimensional model offers a more precise depiction of the electrochemical reactions occurring within the battery, its application in BMS is limited by the complexity of its partial differential equation expressions, the large computational effort, and the lack of the required accuracy of the simplified model [18]. While the combination of the equivalent circuit model and the KF algorithm is a more mature approach, the identification of the parameters needs to be done in a laboratory environment, which limits the updating capability of the SOC estimation algorithm in BMS [19,20].

In recent years, data-driven methods, particularly those utilizing neural networks, have gained popularity for simulating nonlinear system behaviors and predicting SOC [21]. Additionally, generative artificial intelligence technologies can further enhance these capabilities by modeling the complex relationships between a battery's internal state and external characteristics, thereby providing a deeper and more nuanced understanding of battery behaviors [22]. With the advantage of data-driven methods, algorithms such as support vector machines [23], Gaussian process regression [24], and artificial neural networks [25] have been widely studied. Javid et al. [26] used recurrent neural networks (RNN) for SOC estimation, but with the increase of the time step, the RNN exposes the drawbacks of the gradient vanishing and gradient exploding, which makes the network training more difficult. Therefore, Kollmeyer et al. [27] introduced a long short-term memory (LSTM) network to solve the long time-series problem, but the LSTM model only considers capturing time-dependent aspects in the forward direction and ignores capturing information in the backward direction, which has certain limitations [28]. For this reason, Bian et al. [29] used a bidirectional LSTM network model to capture the forward-backward time dependence of SOC estimation, which improves the performance of the model. In addition, the convolutional neural network (CNN) [30] has also been used in SOC estimation due to its excellent feature extraction capabilities, and Hu et al. [31] proposed a hybrid model that utilizes a temporal convolutional network (TCN) to extract features and an LSTM layer is utilized to discern both short-term and long-term dependencies within the battery data. However, the performance of the data-driven model may be impacted as it does not concentrate on the raw features extracted. This oversight could result in failing to learn sufficient critical features or learning an excess of irrelevant features. Li et al. [32] therefore used four different data enhancement methods to preprocess the data and explore the impact of data augmentation on the accuracy of SOC estimation. Since the attention mechanism can focus on feature selection, which improves the learning ability and prediction accuracy of the model, Zhou et al. [33] added the attention mechanism into the TCN network, enhancing the model's focus on critical time steps within the battery dataset. This modification significantly boosts the model's robustness and generalization capabilities. Combined with the above discussion, the data-driven approach provides a solution that is both efficient and feasible for battery SOC estimation, laying the foundation for further development of BMS [34].

Although the data-driven approach for estimating SOC has the benefits of algorithms' simple and easy implementation, it is significantly reliant on the quality and volume of the data available [35]. If there is noise in the data collection process or sparse battery data can reduce the model's accuracy and reliability, thereby weakening the robustness of the estimation results. Second, the training data for the model must cover all possible operating states and environmental conditions of the battery, otherwise, the model's performance may dramatically decrease in unknown states. Additionally, these models usually need to learn from a large amount of data, which not only increases model complexity and training time but also may lead to overfitting, i.e., the model exhibits strong performance on the training dataset but its performance degrades in real-world applications. Thus, it is essential to identify strategies to surmount these limitations, thereby improving the model's performance.

This paper proposes a combination method of advanced feature engineering techniques to enhance SOC estimation accuracy. Considering that battery data often contain noise, we use median filtering (MF) to denoise the acquired raw data. MF is an effective nonlinear digital filtering technique commonly used to remove noise, which helps to minimize the influence of noise on the feature extraction process, thereby ensuring the integrity and quality of the data [36]. We then employ the continuous wavelet transform (CWT) for multi-scale analysis of battery voltage. CWT is a time-frequency analysis tool that can capture the frequency-domain characteristics of batteries under varying conditions [37]. In addition, to enhance the model's generalization capabilities, feature cross (FC) and random forest method (RF) are introduced to increase the amount of data. FC enhances the model's nonlinear capabilities by multiplying two or more features, thereby achieving a nonlinear transformation of the sample space. This technique has proven to enhance model accuracy across various domains [38]. RF, on the other hand, is a feature selection algorithm based on the integrated learning Bagging framework, which automatically calculates the importance of each feature, and then selects the corresponding subset of features based on the ranking of feature importance, thus ensuring that the data dimensionality is reduced while retaining key information [39,40]. Finally, we input the more representative data obtained through these feature engineering techniques into CNN for SOC estimation and compare it with some existing models under different environmental temperatures and operating conditions.

2. Methodology

2.1. Median filtering

In real life, the accuracy of data-driven SOC estimation methods is significantly influenced by the quality of battery data, including voltage, current, and temperature. However, this data often contains disturbances or outliers. To improve the data's quality and reliability, it is essential to filter out these disturbances.

MF is an effective method that removes transient extreme disturbances from data. It works by sorting a set of sampled values within a defined window and selecting the median of these values as the output [36]. Specifically, we concentrate on the target data points and their neighbors within the window, sort the data, and then select the sorted median as the output, replacing the original data points. In this paper, we set the window size to 201 to balance smoothness with detail preservation, ensuring optimal filtering effect.

Compared to other filtering methods, MF not only effectively removes impulse noise and mutations but also preserves the original characteristics of the signal relatively well. It does not cause significant changes to the overall shape of the signal, making it more suitable for application scenarios that require the preservation of signal details or features. Therefore, its use in battery data that requires high detail preservation can improve the accuracy of SOC estimation.

2.2. Continuous wavelet transform

CWT is a technique used to analyze signals simultaneously in the time and frequency domains, and the method is particularly suitable for extracting time-frequency features in unbalanced signals [37]. In practice, the core idea of CWT involves decomposing the signal into wavelets at various scales and positions, and calculating the wavelet coefficients at different scales, to obtain the local characteristics of the signal. The formula of CWT is expressed as:

$$CWT_f(a,b) = [f(t), \psi_{a,b}(t)] = \frac{1}{\sqrt{a}} \int f(t) \psi^* \left(\frac{t-b}{a} \right) dt \tag{1}$$

Among them, $\psi_{a,b}(t)$ is the wavelet basis function adjusted by the scale factor a and the translation factor b:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a} \right) \quad a, b \in R, a > 0 \tag{2}$$

In this work, we choose the real part of the Morlet wavelet as the mother wavelet because it effectively balances time and frequency localization when dealing with signals related to battery voltage. Where, a is determined by the scale range, which is 1:128 used in this paper, and b is implicitly determined by the length of the signal.

To deeply analyze the voltage signal's dynamic properties at various charging and discharging phases, we calculated the energy of the wavelet coefficients at each scale. This calculation is achieved by summing the absolute values of the squares of the coefficients at each scale, as outlined by the following equation:

$$Energy(s) = \sum\nolimits_t |CWT(a,b)|^2 \tag{3}$$

The Energy features calculated by Eq. (3) provide the energy distribution of the voltage at different frequency levels and identify the key features that affect battery performance. This provides deeper insights into the model, thus improving its accuracy. Consequently, CWT proves to be an effective tool for enhancing the accuracy of battery state estimation and helps us better understand the battery behavior under various operating conditions.

2.3. Feature cross

FC also known as feature combinations, creates new features by combining two or more existing features from the original data through multiplicative or other functional relationships [38]. Therefore, it can be argued that the core of this approach is to reveal the non-linear relationships between features, thus enhancing the model's ability to handle complex data. Traditional models often fail to accurately represent the nonlinear relationship between a single feature and its corresponding output, but the cross-combination of features allows the creation of new features representing these relationships. Moreover, in practical applications, certain outputs depend not only on a single feature but are determined by the joint actions of multiple features. FC reveals these interactions and provides more comprehensive information to the model.

During the charging and discharging processes of a battery, the voltage accurately represents the difference in electrochemical potential within the battery, while the current indicates the rate of ion migration. The change in battery temperature affects the ionic conductivity of the electrolyte and the reaction kinetics of the electrode materials, as well as the internal battery's resistance. This, in turn, indirectly influences the charging and discharging efficiency and the battery's lifespan. Based on this understanding, we constructed new features to capture the complex interactions among these parameters, thus enriching the data for data-driven modeling through the following:

(1) Voltage-current crossover feature (power): As a function of voltage and current, power not only reflects the battery's energy conversion rate but also serves as an important thermodynamic parameter influencing the exothermic or absorptive nature of the reactions inside the battery. This feature aids in capturing the energy conversion efficiency and thermal management needs of the battery under different charging and discharging states.

(2) Voltage-temperature crossover feature: The interaction between voltage and cell temperature shows how the rate of electrochemical reactions in a battery is affected by temperature. Temperature variations impact the ionic conductivity of the electrolyte and the rate constants of the electrode reactions, thereby affecting the battery's voltage response. This capability allows the model to more accurately predict the SOC across different temperature conditions.

(3) Current-temperature crossover feature: The interaction between current and temperature illustrates how the current intensity affects the thermal balance inside the battery. Under conditions of high current charging and discharging, an increase in internal battery temperature may degrade the electrolyte performance or accelerate the aging process, making this feature crucial for predicting battery behavior under high load conditions.

By integrating these interacting features into the original data, we not only preserve the information from the original features but also enhance the understanding of the nonlinear interactions between them. This provides more informative and characteristic data for the training of data-driven models. This FC strategy based on physical and chemical mechanisms, significantly enhances the accuracy of battery SOC estimation models and illustrates the crucial role of feature engineering in predicting battery states.

2.4. Random forest

The input data of current data-driven models are often large and complex, with characteristics like high-dimensional and non-linearity. Confronted with vast high-dimensional data, the elimination of redundant features for effective feature selection poses a significant challenge to the field. RF is a composite learning algorithm that utilizes numerous decision trees to train samples and consolidate their predictions [40]. The working principle is illustrated in Fig. 1. As the number of decision trees increases, the higher the accuracy of the algorithm and the stronger the robustness. Furthermore, RF can analyze datasets with complex and interacting features and enhance the model's accuracy by using the variable importance metric as a reference criterion to screen high-dimensional features and reduce the impact of low-relevant features on model training.

The key to feature dimensionality reduction using RF lies in choosing the optimal number of features, and this paper solves this problem by using out-of-bag (OOB) errors to determine the relative importance of feature variables. Subsequently, it sorts and filters these variables to enhance the model's effectiveness. Since the RF generation process uses random sampling with put-back, not all samples will be used, and the unused samples are called OOB samples. The OOB error rate provides an unbiased estimation of the random forest's generalization error, yielding results comparable to those obtained through k-fold cross-validation. If the OOB error rate of a feature is significantly increased by adding noise, it means that the feature has a high impact on the classification result and has a high degree of importance.

Indeed, the essence of feature dimensionality reduction is selecting the optimal number of features. By calculating the OOB error rate, the most contributive features to the model's predictions can be effectively identified, facilitating efficient data dimensionality reduction. The procedure is outlined in the following steps:

1) Initialize the RF model, set the number of trees to 60 in this paper, and specify the use of regression methods;

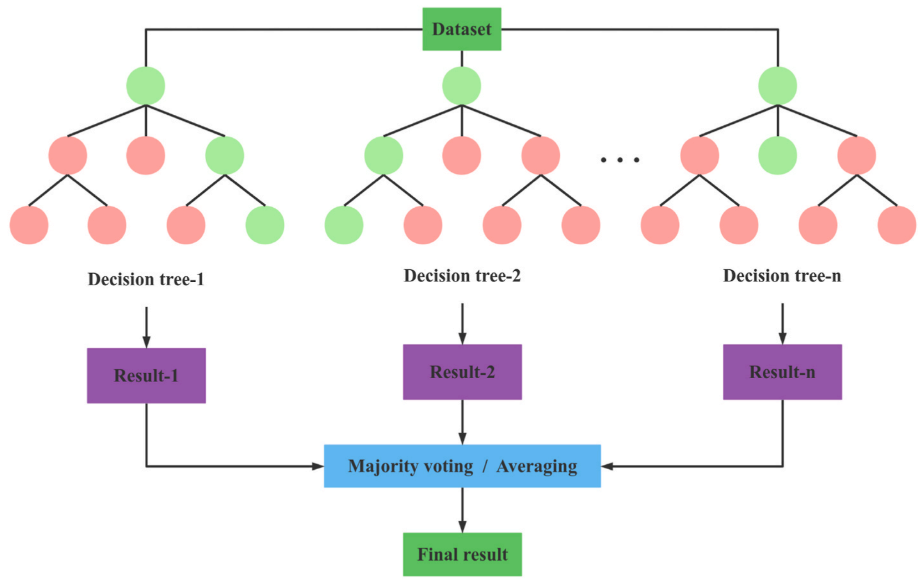

Fig. 1. Working principle of random forest.

2) After training the model, assess the importance of each feature by the OOB error increase;
3) Sort feature importance in descending order and select the top three features as the final subset;
4) Use these selected features to construct a reduced data matrix, including features and labels.

2.5. Convolutional neural network

CNN is a neural network inspired by the structure of the biological visual system, and its core idea is to gradually extract the features from the input data through multiple layers of convolutional and pooling operations and perform classification or regression through the fully connected layer. The traditional fully connected neural network has the disadvantages of a high number of parameters, high computational cost, and easy overfitting. While CNN adopts parameter sharing and sparse connectivity, it can effectively lower the number of parameters and computation and improve the generalization ability of the model. Therefore, in this study, CNN is adopted to estimate the SOC. The CNN receives features after RF dimensionality reduction through the input layer. The core components of a convolutional module typically include a convolutional layer, an activation layer, a pooling layer, and a batch normalization (BN) layer. The convolutional layer extracts features using 1D convolution, the activation layer provides nonlinear representation capability through rectified linear units, the pooling layer reduces data dimensionality and captures high-level features, and the BN layer reduces the effect of distribution shifts during training and improves the network's robustness by normalizing a small batch of inputs. Finally, the extracted features are input into the fully connected layer, which produces the final SOC estimation output. Table 1 lists some parameter settings inside the model.

Table 1
Parameter settings of CNN model.

Parameters	Value
Filter Size (Layer 1/Layer 2)	[3,1]
Number of Filters (Layer 1/Layer 2)	16/32
Stride (Layer 1/Layer 2)	[1,1]
Padding (Layer 1/Layer 2)	[1,1]
Dropout Rate	0.2
Optimizer	Adam
Max Epochs	100
Initial learn rate	0.005
Learn Rate Drop Factor	0.1
Learn Rate Drop Period	100
L2 Regularization	0.0001

2.6. Proposed framework

Fig. 2 shows the general framework of the MF-CWT-FC-RF-CNN model. The acquired raw data are denoised by MF to minimize the impact of noise on the feature extraction process and guarantee the data quality. We then use CWT for multi-scale analysis of battery voltage, a typical time-series data, to capture the frequency domain features of the battery during various operating stages. Considering the inter-correlation between voltage, current, and temperature, we create new features through FC to provide more perspectives for the model, which helps to capture the subtle differences in battery behavior. In addition, we employ RF to further explore the intrinsic structure of the data, which not only reduces the redundancy of the data but also simplifies the model complexity and speeds up the training process while helping to prevent overfitting. Finally, we input these features into the CNN model for accurate SOC prediction.

3. Experimental settings

3.1. Battery dataset description

To confirm the effectiveness of the proposed algorithm, we used a publicly available dataset of Panasonic 18650 PF batteries tested at the University of Wisconsin-Madison [41]. In each test, the battery was initially charged to its full capacity and then discharged until it reached a cutoff voltage of 2.5 V. The dataset details the changes in battery voltage, current, and temperature across five different ambient temperatures ranging from −20 °C to 25 °C. Besides, HPPC conditions and nine different drive cycle conditions (including Cycle 1, Cycle 2, Cycle 3, Cycle 4, NN, UDDS, LA92, US06, and HWFET) were also tested at each ambient temperature. Fig. 3 shows the voltage, current, battery temperature, and SOC curves recorded during the UDDS drive cycle tests conducted at −20 °C and 25 °C ambient temperatures. Table 2 provides the specifications of the batteries utilized in these tests. It is noteworthy that the chemical properties of batteries are notably influenced by temperature, resulting in a nonlinear alteration of their effective capacity, especially under extreme thermal conditions. This nonlinearity

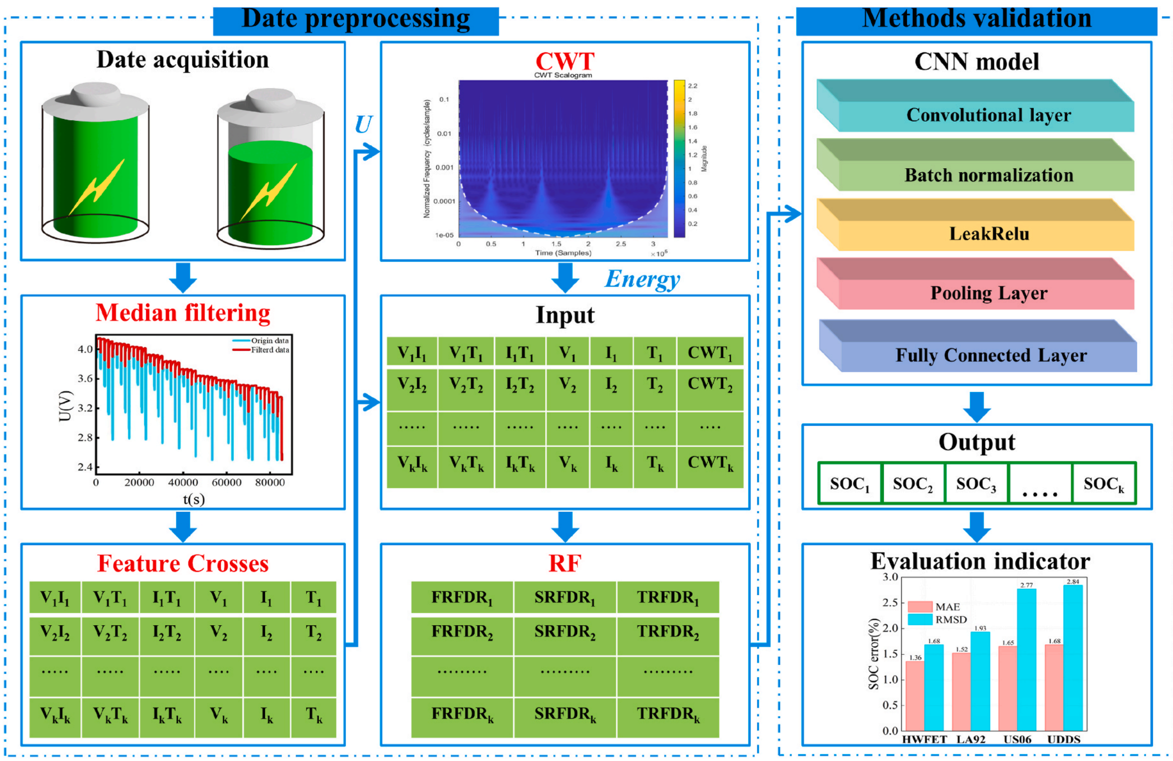

Fig. 2. Flowchart of the MF-CWT-FC-RF-CNN model.

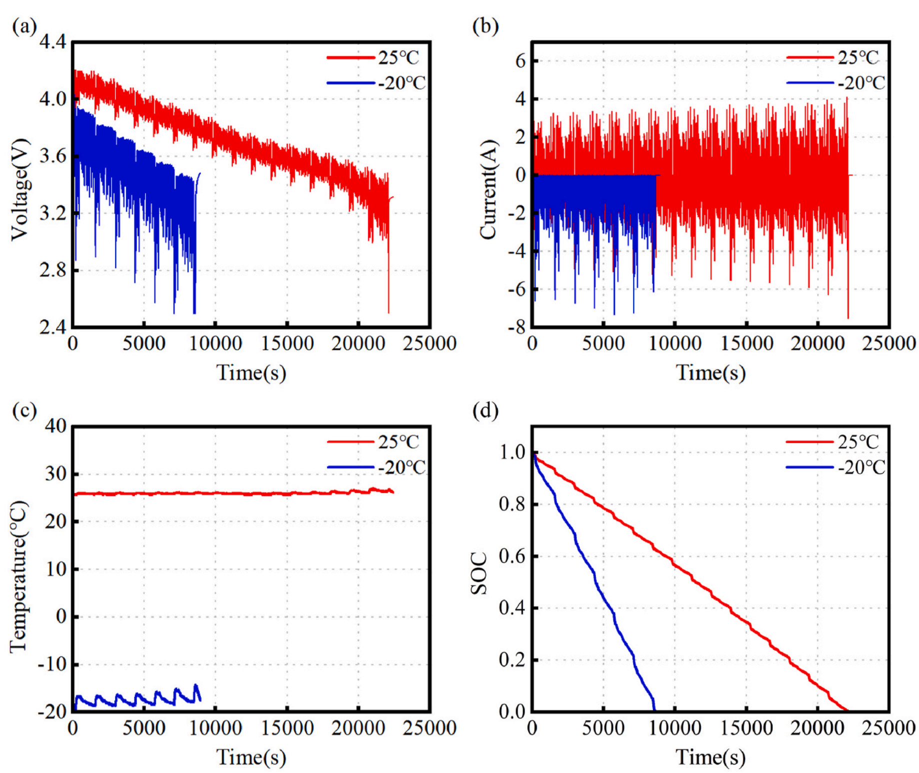

Fig. 3. Voltage, Current, temperature and SOC profiles of the UDDS at 25 °C and −20 °C: (a) voltage; (b) current; (c) temperature; (d) SOC.

complicates the accurate representation of a battery's actual performance using a static maximum capacity value. To tackle this, we determined the maximum average discharge capacity at five distinct constant temperatures through experimental measurements. Table 3 illustrates the relationship between temperature and capacity. Using this data, we apply a quadratic polynomial fitting method to estimate the maximum capacity at unmeasured temperatures, thus ensuring that the capacity values employed in the SOC calculations accurately mirror real conditions across varying temperatures. The resulting temperature-capacity curve is depicted in Eq. (4).

Table 2
Detailed parameters of Panasonic 18650 PF cell.

Battery parameters	Value
Nominal capacity	2.9 Ah
Nominal open circuit voltage	3.6 V
Maximum charging voltage	4.2 V
Minimum discharging voltage	2.5 V
Mass	48 g
Energy storage	9.9 Wh
Minimum charging temperature	10 °C
Cycles to 80 % capacity	500(100 % DOD,25 °C)

Table 3
The temperature-capacity correspondence relationship.

Temperature (°C)	25	10	0	−10	−20
Mean capacity (Ah)	2.65	2.44	2.32	2.03	1.74

$$Q_{MAX}(T) = 2.2845 + 0.0215T - 0.003T^2 \quad (4)$$

3.2. Evaluation criteria

We used the following three performance metrics to assess the SOC estimation performance of the model: R-squared (R^2), mean absolute error (MAE), and root mean square error (RMSE). R^2 quantifies the proportion of the variance in the actual data that is captured by the model's predictions, with a range from 0 to 1. A value closer to 1 indicates a superior fit of the model to the data. MAE and RMSE indicate the mean error and the standard deviation of the model's prediction errors, respectively, and the smaller their values, the higher the estimation accuracy and the better the performance of the model. The detailed calculation formula is presented below:

$$\text{Error} = y_i - \widehat{y_i} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \quad (6)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y_i}| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2} \quad (8)$$

In the formula, n represents the total number of samples, y_i and $\widehat{y_i}$ are the actual and estimated SOC values at the i-th moment, respectively, and $\overline{y}$ is the average of the actual SOC values.

3.3. Comparison models

To confirm the efficacy of our proposed method, this study compares it with the following models:

LSTM [28] networks, equipped with memory cells that retain information from previous time steps recurrently, are adept at learning the context and long-term dependencies within time series data. In this work, the number of hidden layers is fixed at 4.

BILSTM [30] networks, equipped with the capability to learn information from both forward and backward directions in sequence data by integrating two independent LSTM layers. This structure allows BILSTMs to effectively capture the context and long-term dependencies within time series data. In this work, the number of hidden layers is fixed at 4.

CNN-BILSTM-AT model, merges the strengths of CNN and BILSTM architectures, allowing the network to effectively capture both local contextual features and temporal patterns from the input time series. Additionally, the attention mechanism introduces a dynamic weighting strategy, enhancing the model's ability to concentrate on the most pertinent segments of the data for more precise predictions. In this work, the CNN configuration is the same as previously described, with number of attention layers for point multiplication is fixed at 2 and the number of hidden layers is fixed at 4.

To guarantee a fair and thorough comparison across models, we have aligned the initial settings of the hyperparameters for all comparison models with those outlined in Table 1 of this document. Furthermore, we have optimized these parameters through a grid search method to enhance the robustness and reliability of the model results.

4. Results and discussion

To demonstrate the superiority of the MF-CWT-FC-RF-CNN prediction model proposed in this paper, this study evaluates the effectiveness of the feature engineering techniques and illustrates the significant improvement these methods bring to SOC estimation accuracy. Our experiment is divided into three parts. In the first part, we used the dataset of Cycle1-4 at room temperature 25 °C as training data for the data-driven model. The datasets of HWFET, LA92, and US06 at 25 °C serve as test data to evaluate the applied feature engineering techniques as explained in Section II. We assess the impact of individual feature engineering techniques and their combinations on SOC prediction performance by comparing the results of six models: CNN, MF-CNN, CWT-CNN, FC-CNN, FC-RF-CNN, and MF-CWT-FC-RF-CNN (see Section 4.1). This comparison not only demonstrates the benefits of using feature engineering but also highlights the significant enhancement in prediction accuracy when multiple techniques are combined within our proposed model. The second part focuses on comparing the estimation results of the proposed method with those of various models like CNN, LSTM, BILSTM, and CNN-BILSTM-AT at low temperatures, where cyclic data at 10 °C and 0 °C are used for training and testing (see Section 4.2). This comparison validates the effectiveness of the feature engineering technique and tests the model's adaptability, demonstrating that our method maintains high prediction accuracy and reliability even in harsh environments. The final section explores the effect of the model under unknown ambient temperature conditions by selecting the datasets of UDDS, HPPC, and LA92 at four known temperatures as the training data, and the datasets at unknown temperatures as the test data. This approach evaluates the value of including temperature as an explicit input variable, helping the model to learn more directly about the impact of temperature changes on SOC, and validating the model's ability to generalize at unknown temperatures (see Section 4.3). When used for SOC estimation, the model considers battery voltage, current, and temperature per second, $X_t = \{V(t), I(t), T\}$, as inputs, with the output being the current SOC value.

4.1. The impact of feature engineering technology on SOC estimation

In this section, the CNN combines several feature engineering methods described in Section 2 and is trained on the Cycle1-4 discharge dataset. To validate the performance of the proposed model under different operating conditions of the battery, we exclusively use three datasets, HWFET, LA92, and US06, as test sets. These datasets are only used for final performance evaluation after model training and were never used during the training process.

The SOC estimation outcomes and errors for the three conditions are shown in Fig. 4. Analysis of the estimation results shows that MF improves the quality of the data by removing the tip noise in the data, which improves the estimation accuracy of the model, but the improvement is only significant under the US06 operating condition with high data noise. CWT extracts the energy features that respond to voltage changes by performing a time-frequency analysis of the voltage signal, and combining this new feature with the original feature into the

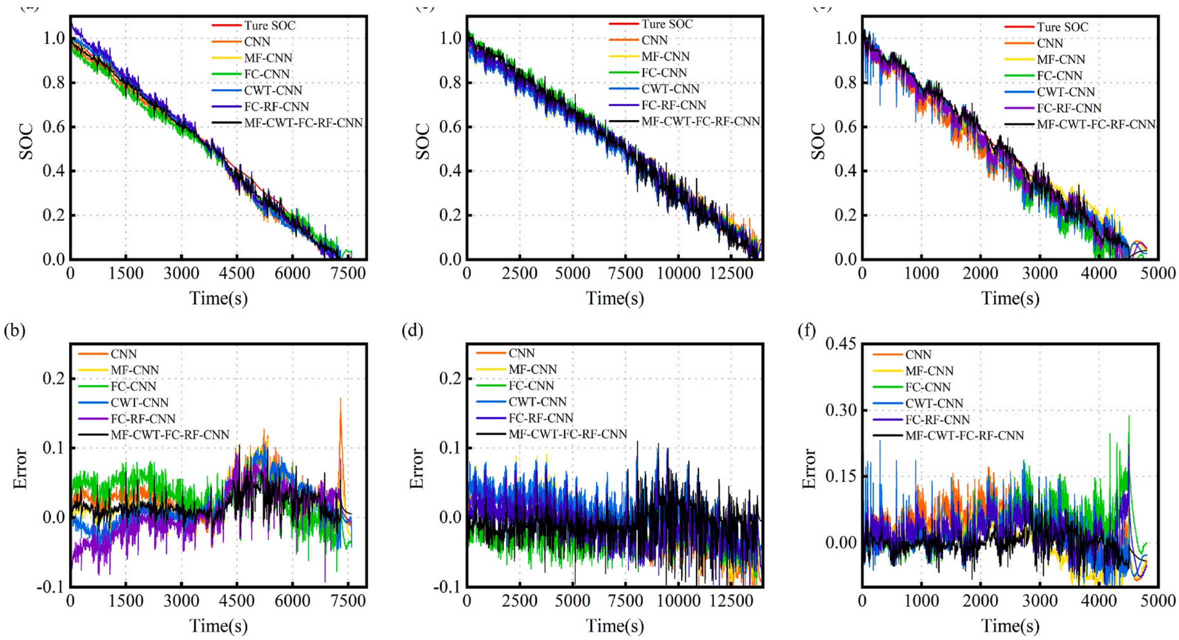

Fig. 4. SOC estimation results and error by different feature engineering technologies at 25 °C: (a) (b) HWFET; (c) (d) LA92; (e) (f) US06.

model reduces the MAE and RMSE to a certain extent and improves the R^2. The results indicate that the prediction accuracy of the above two models has improved to a certain extent, which is in line with the expectations before the experiment.

We also used the FC approach to increase the amount of data to capture the nonlinear relationship between the three features of voltage, current, and battery temperature. However, the accuracy of this method only increased in the LA92 operating condition, where the data fluctuated a lot, but decreased in the other two cases instead. We speculate that this may be due to the fact that the method introduces a large number of redundant features, which leads to model overfitting. To solve this problem, we used RF to identify and retain features with high contribution rates, while removing redundant features for the purpose of data dimensionality reduction. The results show that the accuracy of the model is significantly improved after performing feature dimensionality reduction, a finding that suggests that reasonable dimensionality reduction is necessary in high-dimensional feature spaces.

Although the accuracy of FC-RF-CNN estimation is already high, the model is still affected by anomalous noise at some data points, which makes the prediction results not optimal. Therefore, we combined the methods mentioned above to form the MF-CWT-FC-RF-CNN algorithm to further optimize the model performance. Table 4 lists the comparisons of the results of the evaluation metrics of all the models under three different working conditions. The MAE and RMSE of the final proposed MF-CWT-FC-RF-CNN prediction model are improved by 63.37 % and 58.37 %, while the R^2 is improved by 3.3 % compared with the single CNN model. These results not only demonstrate the usefulness of combining multiple feature engineering techniques in improving model estimation accuracy, but also show that an appropriate combination of feature engineering techniques can overcome the limitations of a single method. By ensuring consistency in the hyperparameters of the CNN model, we can more accurately measure the impact of different feature engineering techniques, rather than changes brought about by hyperparameter adjustments.

Table 4
Performance indicators of different feature engineering technologies at 25 °C.

Working condition	Algorithm	R^2	MAE	RMSE
HWFET	CNN	0.98345	0.0302	0.0384
	MF-CNN	0.98595	0.0254	0.0354
	FC-CNN	0.98315	0.0341	0.0388
	CWT-CNN	0.98570	0.0264	0.0357
	FC-RF-CNN	0.98719	0.0279	0.0338
	MF-CWT-FC-RF-CNN	0.99368	0.0181	0.0238
LA92	CNN	0.98399	0.0299	0.0369
	MF-CNN	0.98667	0.0285	0.0337
	FC-CNN	0.98781	0.0281	0.0322
	CWT-CNN	0.99003	0.0237	0.0291
	FC-RF-CNN	0.99097	0.0224	0.0277
	MF-CWT-FC-RF-CNN	0.99300	0.0190	0.0244
US06	CNN	0.96028	0.0516	0.0603
	MF-CNN	0.98059	0.0342	0.0421
	FC-CNN	0.95517	0.0498	0.0640
	CWT-CNN	0.97648	0.0368	0.0464
	FC-RF-CNN	0.98156	0.0330	0.0411
	MF-CWT-FC-RF-CNN	0.99311	0.0189	0.0251

4.2. *SOC estimation results at two lower ambient temperatures: 10 °C and 0 °C*

In Section 4.1, we explored the performance of the MF-CWT-FC-RF-CNN model solely at a room temperature of 25 °C, without comparing it to other popular models. However, temperature is a critical parameter in battery modeling and significantly impacts SOC estimation. In real-world conditions, it is unrealistic to expect batteries to operate at a constant ambient temperature, and the heat generated during charging and discharging can also cause battery temperatures to rise. Therefore, to comprehensively evaluate the performance of the proposed model under varied temperature conditions, this section focuses on verifying the model's effectiveness at 10 °C and 0 °C. Under these ambient temperatures, driving cycle data such as HWFET, LA92, and US06 continue to serve as the test set, while Cycle1-4 driving cycle data at the corresponding temperatures are used for training. The one-dimensional voltage data was converted to a two-dimensional image using CWT as shown in Fig. 5, from which it can be seen that the time-frequency plots generated for different temperatures and operating conditions are different, and therefore the Energy features we extracted are also different. This transformation enables the model to discern the impacts of temperature and operational changes on voltage from the original one-dimensional data, providing richer and more detailed feature information for the model. With these enhanced features, our model not

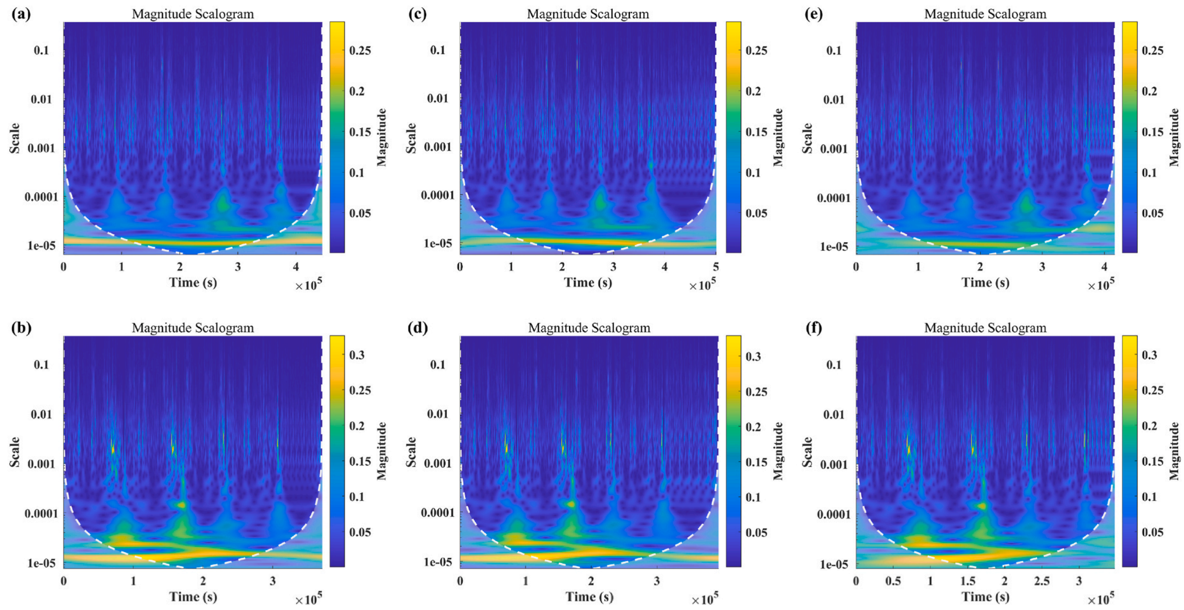

Fig. 5. The time-frequency plots of voltage: (a) HWFET at 10 °C; (b) HWFET at 0 °C; (c) LA92 at 10 °C; (d) LA92 at 0 °C; (e) US06 at 10 °C; (f) US06 at 0 °C.

only learns and understands battery behavior under various states more accurately but also significantly improves the accuracy and reliability of state estimation.

We also compare the proposed MF-CWT-FC-RF-CNN model with existing models such as CNN, LSTM, BILSTM, and CNN-BILSTM-AT. The estimation results and corresponding errors for these five methods are presented in Figs. 6 and 7. The results indicate that the CNN model exhibits the lowest estimation accuracy, as it does not specialize in handling temporal sequences. In contrast, the BILSTM model, which accounts for temporal dynamics in both forward and backward directions, demonstrates higher accuracy than the LSTM model. The CNN-BILSTM model that combines the attention mechanism not only has the advantages of the separate models, but also captures the nonlinear dynamics of the cell, thus achieving higher estimation accuracy. In comparison, our proposed MF-CWT-FC-RF-CNN model outperforms the aforementioned models under two different temperature settings and three distinct operating conditions. The results show that even the poorly performing CNN models can achieve good estimation results when combined with advanced feature engineering and learning high-quality data. Specific evaluation metrics are presented in Tables 5 and 6.

Compared with the results at room temperature, we find that the estimation error increases as the temperature decreases, indicating that the low-temperature environment has a detrimental effect on the battery performance and poses a challenge to the accurate estimation of SOC. However, the comparison of the results above shows that our model achieves superior SOC estimation regardless of the temperature or operating conditions, and the performance is superior to existing models.

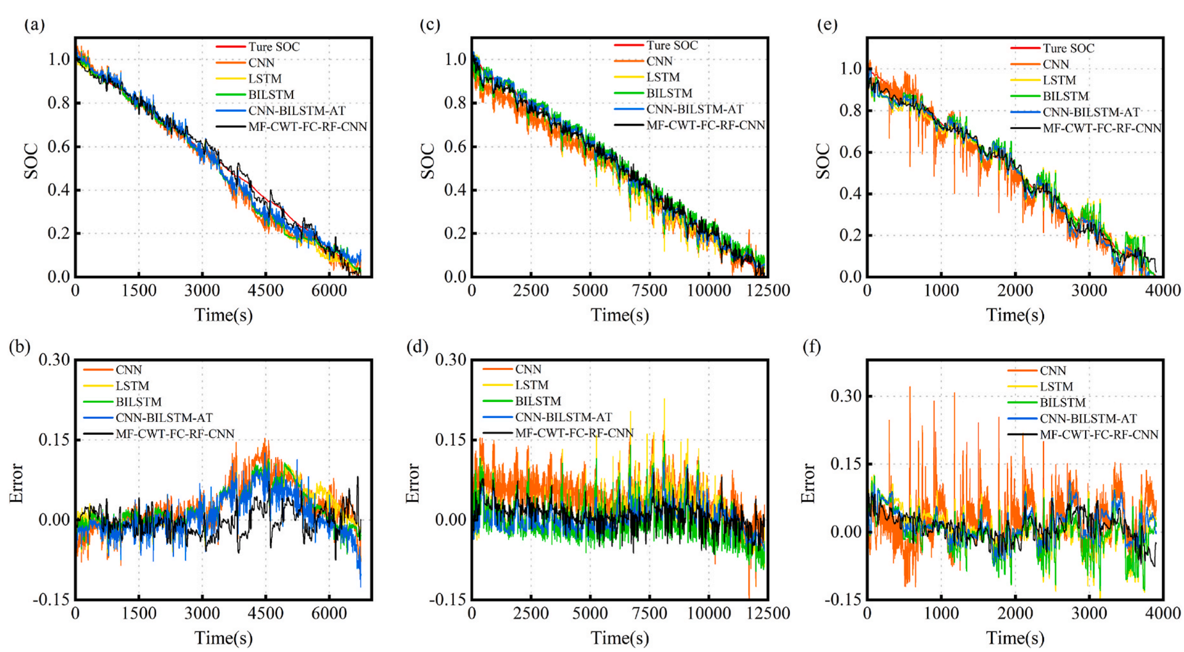

Fig. 6. SOC estimation results and error by different models at 10 °C: (a) (b) HWFET; (c) (d) LA92; (e) (f) US06.

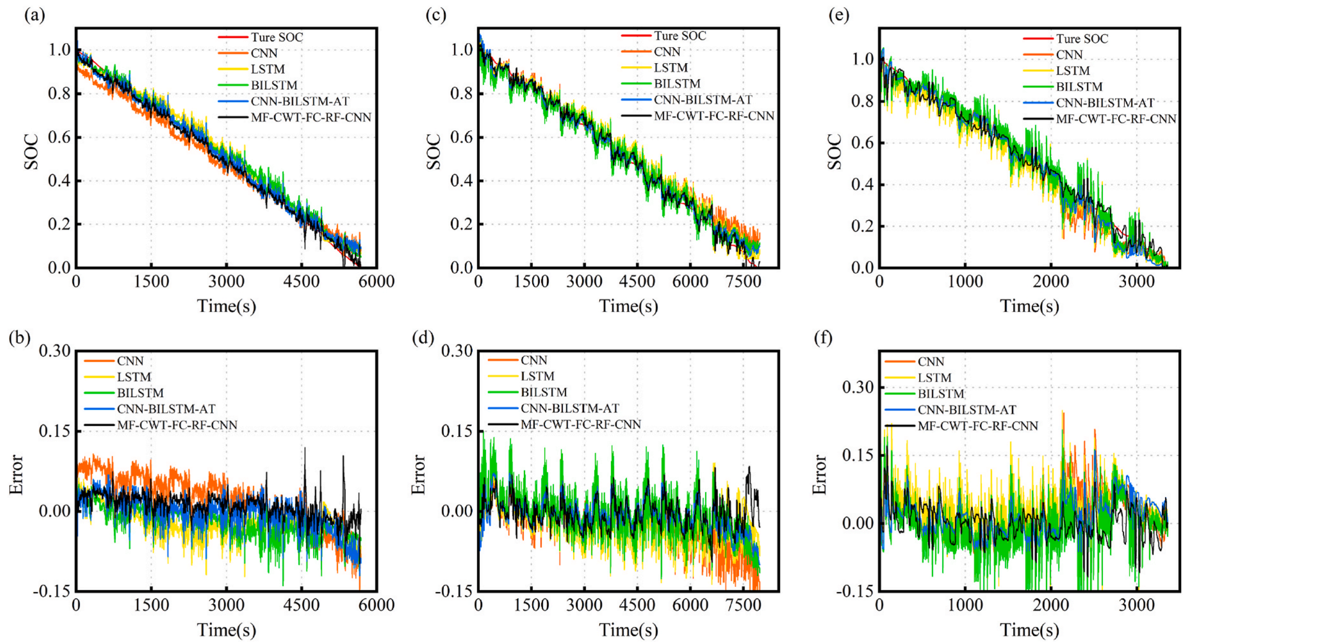

Fig. 7. SOC estimation results and error by different models at 0 °C: (a) (b) HWFET; (c) (d) LA92; (e) (f) US06.

Table 5
Performance indicators of different models at 10 °C.

Working condition	Algorithm	R^2	MAE	RMSE
HWFET	CNN	0.97039	0.0365	0.0491
	LSTM	0.97891	0.0315	0.0414
	BILSTM	0.97957	0.0297	0.0407
	CNN-BILSTM-AT	0.98358	0.0283	0.0365
	MF-CWT-FC-RF-CNN	0.99219	0.0201	0.0252
LA92	CNN	0.96104	0.0486	0.0563
	LSTM	0.98455	0.0262	0.0355
	BILSTM	0.98742	0.0254	0.0320
	CNN-BILSTM-AT	0.99125	0.0199	0.0267
	MF-CWT-FC-RF-CNN	0.99273	0.0187	0.0243
US06	CNN	0.96369	0.0427	0.0541
	LSTM	0.97296	0.0359	0.0467
	BILSTM	0.97706	0.0322	0.0429
	CNN-BILSTM-AT	0.98050	0.0315	0.0396
	MF-CWT-FC-RF-CNN	0.98826	0.0245	0.0308

Table 6
Performance indicators of different models at 0 °C.

Working condition	Algorithm	R^2	MAE	RMSE
HWFET	CNN	0.96468	0.0479	0.0543
	LSTM	0.98250	0.0326	0.0382
	BILSTM	0.98410	0.0296	0.0364
	CNN-BILSTM-AT	0.98833	0.0235	0.0312
	MF-CWT-FC-RF-CNN	0.99302	0.0191	0.0241
LA92	CNN	0.97387	0.0362	0.0469
	LSTM	0.98031	0.0333	0.0407
	BILSTM	0.98253	0.0294	0.0383
	CNN-BILSTM-AT	0.98851	0.0244	0.0311
	MF-CWT-FC-RF-CNN	0.99068	0.0224	0.0279
US06	CNN	0.95660	0.0456	0.0592
	LSTM	0.96294	0.0421	0.0547
	BILSTM	0.96788	0.0384	0.0509
	CNN-BILSTM-AT	0.97675	0.0346	0.0433
	MF-CWT-FC-RF-CNN	0.98291	0.0289	0.0371

4.3. SOC estimation results at unknown temperatures

The MF-CWT-FC-RF-CNN model exhibits good SOC estimation performance at constant ambient temperature. However, in practical applications, batteries are subject to varying ambient temperatures throughout the day and it is impossible for the neural network to learn the charging and discharging behavior at each temperature. Consequently, results obtained at a single temperature may not accurately reflect the model's overall performance. To assess whether our model can reliably estimate SOC at unknown temperatures, we conducted the following experiments: First, we used discharge data from UDDS conditions at ambient temperatures of −20 °C, −10 °C, 10 °C, and 25 °C as a training set to predict the SOC at 0 °C. Second, we used discharge data from HPPC conditions at −20 °C, −10 °C, 0 °C, and 25 °C as a training set to predict the SOC at 10 °C. Finally, the discharge data at −20 °C, −10 °C, 0 °C and 10 °C for the LA92 condition was used to predict the SOC at 25 °C.

As illustrated in Fig. 8, despite the networks not being trained at 0 °C and 10 °C, the five models were able to predict the SOC at these temperatures with relatively high accuracy. This indicates that although the method of integrating data from different ambient temperatures does not fully capture the dynamic characteristics of the battery over the entire temperature range, this method still provides valuable insight into the performance of the models over a wide range of temperature conditions. However, since 25 °C is an extrapolated temperature not directly included in our training dataset, and thus the accuracy of the SOC estimation at this temperature significantly decreases.

As shown in the performance metrics comparison in Table 7, the estimation performance of standalone CNN, LSTM, and BILSTM models remains sub-optimal. In contrast, the accuracy of the CNN-BILSTM-AT model has significantly improved but still falls short of that exhibited by our proposed MF-CWT-FC-RF-CNN model. Overall, our model demonstrates robust generalization capabilities, producing optimal estimation results under both lower and unknown temperature conditions. Notably, even without having been trained on the charging and discharging data at 25 °C, the model achieves an R^2 value of 0.98766 and an RMSE value of 3.17 %.

This section of the study encompasses a temperature range from −20 °C to 25 °C, totaling 45 °C. This approach not only simulates real-world conditions but also offers an effective method for SOC estimation

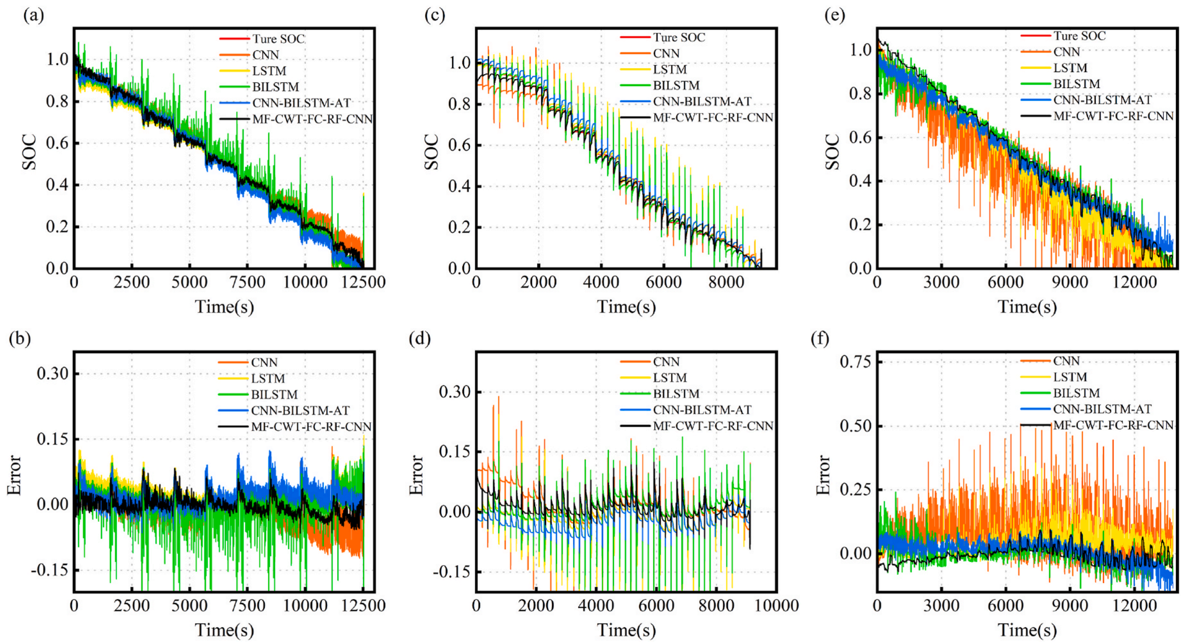

Fig. 8. SOC estimation results and error by different models: (a) (b) UDDS at 0 °C; (c) (d) HPPC at 10 °C; (e) (f) LA92 at 25 °C.

Table 7
Performance indicators of different models at unknown temperatures.

Working condition	Algorithm	R^2	MAE	RMSE
UDDS	CNN	0.97571	0.0344	0.0446
	LSTM	0.98021	0.0302	0.0402
	BILSTM	0.98345	0.0251	0.0368
	CNN-BILSTM-AT	0.98751	0.0240	0.0320
	MF-CWT-FC-RF-CNN	0.99390	0.0171	0.0223
HPPC	CNN	0.97599	0.0345	0.0479
	LSTM	0.98173	0.0271	0.0434
	BILSTM	0.98355	0.0265	0.0412
	CNN-BILSTM-AT	0.98806	0.0295	0.0351
	MF-CWT-FC-RF-CNN	0.99233	0.0179	0.0281
LA92	CNN	0.89595	0.0693	0.0923
	LSTM	0.97137	0.0360	0.0484
	BILSTM	0.97514	0.0332	0.0451
	CNN-BILSTM-AT	0.98206	0.0324	0.0383
	MF-CWT-FC-RF-CNN	0.98766	0.0252	0.0317

using neural networks. By employing this method, we gain a deeper understanding of battery performance and behavior at various ambient temperatures. This knowledge, in turn, informs the optimization and design of BMS.

5. Conclusions

In this paper, a machine learning model, MF-CWT-FC-RF-CNN, based on the combination of advanced feature engineering techniques and CNN, is proposed to significantly improve the accuracy of SOC estimation for lithium-ion batteries under a wide range of operating conditions and temperatures. The model relies mainly on comprehensive feature engineering of the input data to optimize data quality and extract key information. Learning from this feature-engineered optimized data enables the CNN model to predict the SOC of the battery more accurately. This fusion of feature engineering and deep learning greatly enhances the model's ability to adapt to different environmental variations, which in turn leads to more reliable SOC estimation under unknown conditions.

Our study comprehensively compares the performance of various types of neural network models in three different environmental scenarios, including SOC estimation under room temperature conditions applying different feature engineering techniques, SOC estimation at low ambient temperatures, and SOC estimation at unknown ambient temperatures.

The main contributions of our work include the following:

1. Data quality improvement: We employed MF to denoise the input raw data, removing outliers and noise, thereby enhancing the data quality. This improvement directly increases the model's estimation accuracy, resulting in a significant enhancement in performance across various test conditions.
2. Enhanced data representation: We also feature extracted the denoised voltage data through CWT to identify the key features affecting the battery performance, which provided deep information for the model. Meanwhile, the two-by-two cross-tabulation analysis between denoised voltage, denoised current and denoised temperature increases the dimensionality of the data to better capture and represent the complex relationships among the data. Finally, to prevent model overfitting, we use RF to downscale these dimensionally increased features and filter out the three most influential features, which enhances the model's generalization capability and ensures excellent performance in adapting to temperature variations.
3. Comparison of model performance: Experimental results show that our proposed models outperform several popular models, with MAE and RMSE below 2.89 % and 3.71 %, respectively, under various operating conditions and temperatures.

In this paper, we have only explored the method of generating new features by multiplying two and two features, which improves the nonlinear learning ability and estimation accuracy of the model to a certain extent, but we can also further improve the SOC by studying the polynomial relationship between the voltage and current, respectively, and the temperature, and then generating more useful information for the model through the polynomial feature crossover or higher-order feature crossover, thus further improving the SOC estimation accuracy. In addition, in our future research, we intend to extend the feature space by including some features related to the chemical properties of the battery in addition to the basic features of external variables like voltage, current, and battery temperature, to further improve the accuracy of the SOC estimation of lithium-ion batteries.

CRediT authorship contribution statement

Sijing Wang: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Meiyuan Jiao:** Writing – review & editing, Validation, Software, Conceptualization. **Ruoyu Zhou:** Writing – original draft, Validation, Formal analysis, Conceptualization. **Yijia Ren:** Writing – original draft, Methodology, Formal analysis. **Honglai Liu:** Validation, Supervision, Funding acquisition. **Cheng Lian:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 22278127), the Fundamental Research Funds for the Central Universities (No. 2022ZFJH004), and the Shanghai Pilot Program for Basic Research (22T01400100-18).

References

[1] S. Nyamathulla, C. Dhanamjayulu, A review of battery energy storage systems and advanced battery management system for different applications: challenges and recommendations, J. Energy Storage 86 (2024) 111179.

[2] K. Luo, X. Chen, H. Zheng, et al., A review of deep learning approach to predicting the state of health and state of charge of lithium-ion batteries, J. Energy Chem. 74 (2022) 159–173.

[3] M.S.H. Lipu, M.A. Hannan, T.F. Karim, et al., Intelligent algorithms and control strategies for battery management system in electric vehicles: progress, challenges and future outlook, J. Clean. Prod. 292 (2021) 126044.

[4] M. Kurucan, M. Özbaltan, Z. Yetgin, et al., Applications of artificial neural network based battery management systems: a literature review, Renew. Sustain. Energy Rev. 192 (2024) 114262.

[5] H. Ruan, J.V. Barreras, T. Engstrom, et al., Lithium-ion battery lifetime extension: a review of derating methods, J. Power Sources 563 (2023) 232805.

[6] D. Gräf, J. Marschewski, L. Ibing, et al., What drives capacity degradation in utility-scale battery energy storage systems? The impact of operating strategy and temperature in different grid applications, J. Energy Storage 47 (2022) 103533.

[7] O. Kadem, J. Kim, Real-time state of charge-open circuit voltage curve construction for battery state of charge estimation, IEEE Trans. Veh. Technol. 72 (7) (2023) 8613–8622.

[8] F. Naseri, S. Karimi, E. Farjah, et al., Supercapacitor management system: a comprehensive review of modeling, estimation, balancing, and protection techniques, Renew. Sustain. Energy Rev. 155 (2022) 111913.

[9] G. Zubi, R. Dufo-López, M. Carvalho, et al., The lithium-ion battery: state of the art and future perspectives, Renew. Sustain. Energy Rev. 89 (2018) 292–308.

[10] M. Adaikkappan, N. Sathiyamoorthy, Modeling, state of charge estimation, and charging of lithium-ion battery in electric vehicle: a review, Int. J. Energy Res. 46 (3) (2022) 2141–2165.

[11] N. Yang, X. Zhang, G. Li, State of charge estimation for pulse discharge of a LiFePO4 battery by a revised Ah counting, Electrochim. Acta 151 (2015) 63–71.

[12] R. Xiao, Y. Hu, X. Jia, et al., A novel estimation of state of charge for the lithium-ion battery in electric vehicle without open circuit voltage experiment, Energy 243 (2022) 123072.

[13] S. Zhang, X. Zhang, A novel non-experiment-based reconstruction method for the relationship between open-circuit-voltage and state-of-charge/state-of-energy of lithium-ion battery, Electrochim. Acta 403 (2022) 139637.

[14] S. Wang, P. Huang, C. Lian, et al., Multi-interest adaptive unscented Kalman filter based on improved matrix decomposition methods for lithium-ion battery state of charge estimation, J. Power Sources 606 (2024) 234547.

[15] J. Liu, L. Zhou, Y. Zhang, et al., Aging behavior and mechanisms of lithium-ion battery under multi-aging path, J. Clean. Prod. 423 (2023) 138678.

[16] M. Hossain, M.E. Haque, M.T. Arif, Kalman filtering techniques for the online model parameters and state of charge estimation of the Li-ion batteries: a comparative analysis, J. Energy Storage 51 (2022) 104174.

[17] L. He, Y. Wang, Y. Wei, et al., An adaptive central difference Kalman filter approach for state of charge estimation by fractional order model of lithium-ion battery, Energy 244 (2022) 122627.

[18] A. Bartlett, J. Marcicki, S. Onori, et al., Electrochemical model-based state of charge and capacity estimation for a composite electrode lithium-ion battery, IEEE Trans. Control Syst. Technol. 24 (2) (2015) 384–399.

[19] Y. Jiang, J. Xu, M. Liu, et al., An electromechanical coupling model-based state of charge estimation method for lithium-ion pouch battery modules, Energy 259 (2022) 125019.

[20] O.O. Solomon, W. Zheng, J. Chen, et al., State of charge estimation of Lithium-ion battery using an improved fractional-order extended Kalman filter, J. Energy Storage 49 (2022) 104007.

[21] R.G. Widjaja, M. Asrol, I. Agustono, et al., State of charge estimation of lead acid battery using neural network for advanced renewable energy systems, Emerging Science Journal 7 (3) (2023) 691–703.

[22] Y. Liu, Z. Yang, Z. Yu, et al., Generative artificial intelligence and its applications in materials science: current situation and future perspectives, Journal of Materiomics 9 (4) (2023) 798–816.

[23] A. Manoharan, K.M. Begam, V.R. Aparow, et al., Artificial neural networks, gradient boosting and support vector machines for electric vehicle battery state estimation: a review, J. Energy Storage 55 (2022) 105384.

[24] K. Fan, Y. Wan, B. Jiang, State-of-charge dependent equivalent circuit model identification for batteries using sparse Gaussian process regression, J. Process Control 112 (2022) 1–11.

[25] Z. Cui, L. Kang, L. Li, et al., A hybrid neural network model with improved input for state of charge estimation of lithium-ion battery at low temperatures, Renew. Energy 198 (2022) 1328–1340.

[26] G. Javid, D. Ould Abdeslam, M. Basset, Adaptive online state of charge estimation of EVs lithium-ion batteries with deep recurrent neural networks, Energies 14 (3) (2021) 758.

[27] E. Chemali, P.J. Kollmeyer, M. Preindl, et al., Long short-term memory networks for accurate state-of-charge estimation of Li-ion batteries, IEEE Trans. Ind. Electron. 65 (8) (2017) 6730–6739.

[28] Y. Tian, R. Lai, X. Li, et al., A combined method for state-of-charge estimation for lithium-ion batteries using a long short-term memory network and an adaptive cubature Kalman filter, Appl. Energy 265 (2020) 114789.

[29] C. Bian, H. He, S. Yang, Stacked bidirectional long short-term memory networks for state-of-charge estimation of lithium-ion batteries, Energy 191 (2020) 116538.

[30] E. Buchicchio, A. De Angelis, F. Santoni, et al., Uncertainty characterization of a CNN method for Lithium-Ion Batteries state of charge estimation using EIS data, Measurement 220 (2023) 113341.

[31] C. Hu, F. Cheng, L. Ma, et al., State of charge estimation for lithium-ion batteries based on TCN-LSTM neural networks, J. Electrochem. Soc. 169 (3) (2022) 030544.

[32] M. Li, C. Li, C. Chen, et al., Effect of data enhancement on state-of-charge estimation of lithium-ion battery based on deep learning methods, J. Energy Storage 82 (2024) 110573.

[33] F. Li, W. Zuo, K. Zhou, et al., State of charge estimation of lithium-ion batteries based on PSO-TCN-Attention neural network, J. Energy Storage 84 (2024) 110806.

[34] M. Waseem, M. Ahmad, A. Parveen, et al., Battery technologies and functionality of battery management system for EVs: current status, key challenges, and future prospectives, J. Power Sources 580 (2023) 233349.

[35] Y. Liu, Z. Yang, X. Zou, et al., Data quantity governance for machine learning in materials science, Natl. Sci. Rev. 10 (7) (2023) nwad125.

[36] K. Balasamy, D. Shamia, Feature extraction-based medical image watermarking using fuzzy-based median filter, IETE J. Res. 69 (1) (2023) 83–91.

[37] D. Łuczak, Machine Fault diagnosis through vibration analysis: continuous wavelet transform with complex morlet wavelet and time-frequency RGB image recognition via convolutional neural network, Electronics 13 (2) (2024) 452.

[38] Z. Gu, Y. Li, H. Luo, et al., Cross attention guided multi-scale feature fusion for false-positive reduction in pulmonary nodule detection, Comput. Biol. Med. 151 (2022) 106302.

[39] W. Jia, M. Sun, J. Lian, et al., Feature dimensionality reduction: a review, Complex & Intelligent Systems 8 (3) (2022) 2663–2693.

[40] J.D. Loyal, R. Zhu, Y. Cui, et al., Dimension reduction forests: local variable importance using structured random forests, J. Comput. Graph Stat. 31 (4) (2022) 1104–1113.

[41] Phillip Kollmeyer, Panasonic 18650PF Li-Ion Battery Data, Mendeley Data, V1, 2018, https://doi.org/10.17632/wykht8y7tg.1.