IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 29, 2022, accepted February 8, 2022, date of publication February 15, 2022, date of current version February 24, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3151975*

# Transformer Network for Remaining Useful Life Prediction of Lithium-Ion Batteries

**DAOQUAN CHEN[1], WEICONG HONG[2], AND XIUZE ZHOU[2]**
[1]School of Intelligent Transportation, Zhejiang Institute of Mechanical and Electrical Engineering, Hangzhou 310053, China
[2]Shuye Technology Company Ltd., Hangzhou, China

Corresponding author: Xiuze Zhou (zhouxiuze@foxmail.com)

**ABSTRACT** Accurately predicting the Remaining Useful Life (RUL) of a Li-ion battery plays an important role in managing the health and estimating the state of a battery. With the rapid development of electric vehicles, there is an increasing need to develop and improve the techniques for predicting RUL. To predict RUL, we designed a Transformer-based neural network. First, battery capacity data is always full of noise, especially during battery charge/discharge regeneration. To alleviate this problem, we applied a Denoising Auto-Encoder (DAE) to process raw data. Then, to capture temporal information and learn useful features, a reconstructed sequence was fed into a Transformer network. Finally, to bridge denoising and prediction tasks, we combined these two tasks into a unified framework. Results of extensive experiments conducted on two data sets and a comparison with some existing methods show that our proposed method performs better in predicting RUL. Our projects are all open source and are available at https://github.com/XiuzeZhou/RUL.

**INDEX TERMS** Li-ion battery, remaining useful life, transformer, denoising auto-encoder, neural network.

## I. INTRODUCTION

Having light weight, high-energy density, good performance and a long lifetime, the rechargeable Lithium-ion (Li-ion) battery is widely applied in various devices [1]–[4]. However, as the charge-discharge cycle increases, capacity generally degrades. Prognostics and Health Management (PHM) methods, of which the prediction of Remaining Useful Life (RUL) is a very important component, are necessary to ensure the reliability and safety of an electronic device [5]–[7]. To ensure safety, prediction of RUL in advance provides some key information about the maintenance and replacement of batteries [8]–[10]. Fig. 1 illustrates a toy-example of battery use.

An accurate prediction of lifetime and estimation of health for batteries are crucial for durable electronic devices. Recent advancements and achievements in machine learning in various fields have piqued interest in the estimation of data-driven battery health [4], [11], [12]. For example, to account for the effects of discharge current and ambient temperature, Ng *et al.* [13] proposed a naive Bayes model to predict the RUL. Regarding the prognostics of battery health, Nuhic *et al.* [14] explored applying Support Vector Machine (SVM) to learn the decay process. Also, for prognostics and to model battery degradation, Liu *et al.* [15], by online learning, developed the Relevance Vector Machine (RVM).

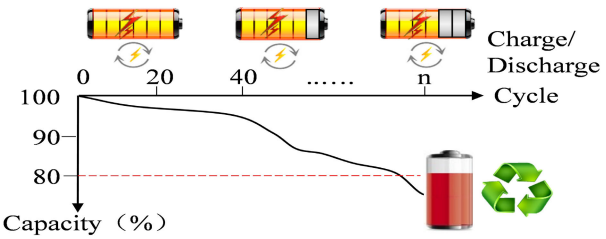



**FIGURE 1.** Example of battery capacity degradation.

Recently, remarkable success has been achieved by deep learning in various fields, such as recommender systems [16]–[18], Computer Vision (CV) [19]–[21], and Natural Language Processing (NLP) [22]–[24]. To learn about the nonlinear nature of battery capacity, deep learning models are also widely applied to RUL prediction. For example, to capture the relationship between RUL and a charge curve, Multi-Layer Perceptron (MLP) was used to describe the charge process and the terminal voltage curve of a battery [2], [25], [26]. To assess the State Of Health (SOH), Recurrent Neural Network (RNN) was developed to simulate the nonlinear trend [27]–[29]. To learn about inclination of

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Zhang.

battery degradation, Long Short-Term Memory (LSTM) was applied to the capacity sequence [8], [30], [31].

When predicting RUL, RNN-based frameworks, including Gate Recurrent Unit (GRU) and LSTM, are effective solutions for modeling sequential data. Although most existing RNN-based frameworks have shown promising performance, they have the following three major problems:

(1) Using RNN-based networks to model sequential data in a recurrent manner not only results in high time costs for training, but also degrades performance due to long-term dependency [32]–[34];

(2) To learn representation, raw data is fed directly to the neural networks; however, the training data are always full of noise, especially when capacity regeneration occurs. The highly dynamic and nonlinear capacity curve affects RNN-based methods [28], [31], [35].

(3) In most methods, data denoising and model prediction are two separate tasks; thus, the correlation between the two tasks is ignored [28], [36], [37].

To address these problems, we designed a novel neural network to model sequential capacity patterns. In the network, a Transformer, which effectively and efficiently captures useful information of the sequences, serves as the body of the model. To learn trends from the sequences, the multi-head attention network of the Transformer accelerates the training performance of the neural networks. To the best of our knowledge, this is the first Transformer-based architecture to predict RUL in the field of Li-ion batteries.

Also, to build a robust network, it is necessary to deal with noise, outliers, and irrelevant data. The representation ability of a neural network heavily relies on the quality of the source. Thus, to accurately predict RUL, the Denoising Auto-Encoder (DAE), with its powerful ability to learn representation from noisy raw data, is used to reconstruct input data.

Finally, for better generalization, we propose an objective function to bridge denoising and prediction, instead of solving these two tasks separately. The learning procedure optimizes both tasks simultaneously in a unified framework.

## II. RELATED WORK

### A. PROBLEM DEFINITION

An accurate, timely RUL predictor is important for a battery to maintain advance warning of potential risk [11]. For batteries, SOH, a health indicator for battery aging, represents the states of battery in each charge-discharge cycle [31], [38]. RUL is defined by the following capacity ratio:

$$SOH(t) = \frac{C_t}{C_0} \times 100\%, \tag{1}$$

where $C_0$ denotes rated capacity, and $C_t$ denotes the measured capacity of cycle, $t$. As the number of times a battery is charged/discharged increases, capacity degrades. For a battery, End of Life (EOL) which is closely related to its capacity [39], is defined as the point when remaining capacity reaches 70-80% of initial capacity [31], [40]. Fig. 2 illustrates an example of RUL prediction.

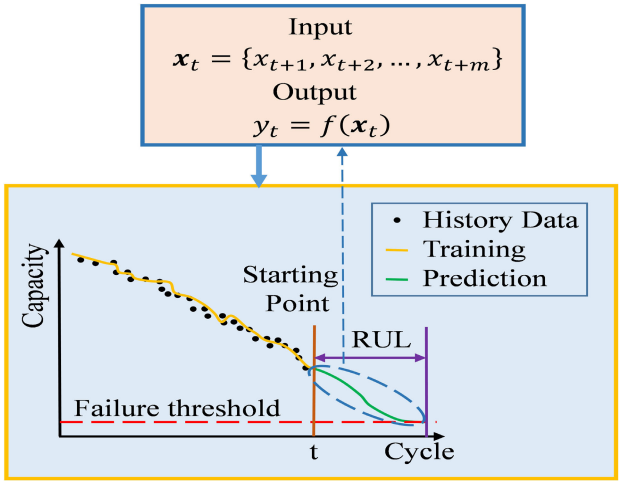



**FIGURE 2.** Example of RUL prediction.

### B. DEEP LEARNING FOR RUL

Because Li-ion batteries are a source of power for many devices, it is critical to ensure their reliability and safety. RUL prediction and SOH evaluation have become increasingly important topics and have received considerable attention in recent years. Methods to predict RUL for batteries are classified into two kinds: model-based and data-driven [4], [41], [42].

#### 1) MODEL-BASED

To fit the degradation curve of a battery, mathematical models are built to describe the physical properties. However, in practice, for a battery working under some noisy and uncertain environments, it is difficult for mathematical models to accurately assess the SOH [27], [43]–[46].

#### 2) DATA-DRIVEN

Data-driven methods are modeled on historical data without considering any properties of the battery. Because of their flexibility and ease of operation, data-driven methods receive more popular attention [47]–[50].

Neural network based data-driven methods possess good generalization and powerful feature extraction ability [2], [51]. To predict RUL for Li-ion batteries, many deep learning models have been proposed. For RUL prediction, MLP is applied to learn nonlinear degradation [2], [25]. However, it poorly captures the temporal information from the input sequence. To deal with the sequence data, many RNN-based frameworks, including RNN, LSTM and GRU, have been designed [8], [31], [52]. However, RNN-based frameworks with a recurrent manner have a high time cost for training and degrade performance due to long-term dependency. To speed up training, CNN is used [36], [53]. But when it comes to time series, CNN, as MLP, runs into the same problem: it achieves limited performance in degradation trend.

**TABLE 1. Major notations used in this paper.**

| Notations | Description |
|---|---|
| $\mathbf{x}$ | input sequence of capacity |
| $x_t$ | the $t$-th capacity of $\mathbf{x}$ |
| $\mathbf{x}'_t$ | slice of $\mathbf{x}$ |
| $n$ | length of $\mathbf{x}$ |
| $m$ | length of $\mathbf{x}'_t$ |
| $C_0$ | rated capacity |
| $\|\cdot\|_F^2$ | Fibonacci-norm |
| $\widehat{x}_t$ | predicted value of $x_t$ |
| $\widehat{\mathbf{x}}_t$ | predicted value of $\mathbf{x}_t$ |
| $\widetilde{\mathbf{x}}_t$ | corrupted vector of $\mathbf{x}_t$ |
| $\boldsymbol{z}$ | output of the DAE encoder |
| $a(\cdot)$ | activation function |
| $\ell(\cdot)$ | loss function |
| $f(\cdot)$ | function |
| $\Omega(\cdot)$ | regularization |
| $\Theta$ | learning parameters of models |
| $\lambda$ | a regularization parameter |
| $\mathcal{L}_d$, $\mathcal{L}$ | objective functions |
| $\boldsymbol{Q}$ | *query* of *Attention* |
| $\boldsymbol{K}$ | *key* of *Attention* |
| $\boldsymbol{V}$ | *value* of *Attention* |
| $\boldsymbol{W}$, $\boldsymbol{W}'$, $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{W}_p$ | weights of networks |
| $\boldsymbol{b}$, $\boldsymbol{b}'$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{b}_p$ | biases of networks |
| $\boldsymbol{W}_Q^l$, $\boldsymbol{W}_K^l$, $\boldsymbol{W}_V^l$, $\boldsymbol{W}^O$ | weights of *Multi-Head Attention* |

Transformers, which perform well in encoding text, have been explored for various applications, such as recommendation systems [54], [55] and CV [56], [57]. They parallelly and effectively capture long-range dependencies by an attention mechanism. Owing to the effectiveness and efficiency of Transformers in modeling long sequences, we explore using them to capture the weight of capacities at different time steps in the prediction of RUL.

Also, to further improve accuracy, various modules are combined to gain their advantages [11], [36], [58], [59]. Although those deep learning methods have achieved great success in exploring battery decay trends, they train directly on noisy data, which limits the model to learn accurate representations. To denoise and get clean input data, Lu *et al.* [37] proposed AE-GRU, in which an autoencoder was used in the data pre-processing step to extract the features of the original data, and GRU was used to learn the long-term inclination. However, in AE-GRU, data denoising and RUL prediction are two separate tasks; thus, the correlation between the two tasks is ignored. In this paper, we propose an objective function to bridge denoising and prediction and optimize both tasks simultaneously in a unified framework.

## III. THE PROPOSED METHOD

The main goal of our model is to predict RUL from historical records. Therefore, first we describe our architecture in detail. Then, we describe the objective function, which jointly combines DAE and prediction loss.

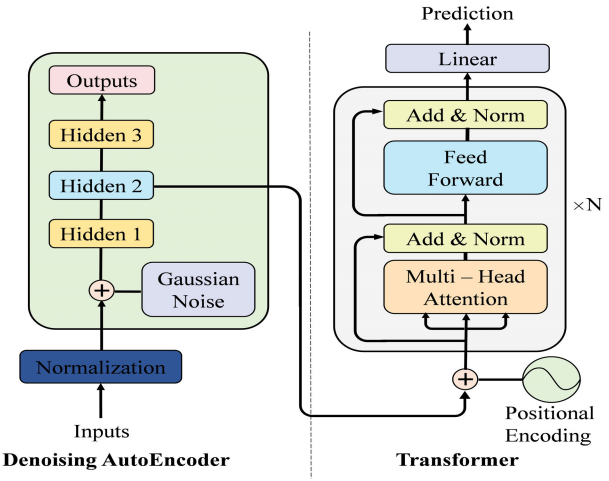



**FIGURE 3. Denoising transformer network for RUL prediction.**

### A. DENOISING TRANSFORMER NETWORK FOR RUL

To provide for uninterrupted battery operation and determine appropriate maintenance, accurate and timely prediction of RUL is important. To solve the problem of most existing RNN-based methods, we designed a deep learning architecture, Denoising Transformer (DeTransformer) network, consisting of four parts: input and normalization, denoising, Transformer, and prediction. The architecture is shown in Fig. 3. Table 1 lists the major notations used throughout this paper.

#### 1) INPUT AND NORMALIZATION

To reduce the influence of input data distribution changes on neural networks, the data must be normalized. Let $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ denote the input sequence of capacity with length $n$, which is mapped to $(0, 1]$:

$$\mathbf{x}' = \frac{\mathbf{x}}{C_0}, \tag{2}$$

where $C_0$ denotes rated capacity.

#### 2) DENOISING

Raw input is always full of noise, especially when charge/discharge regeneration occurs. In most methods, raw data is fed directly to the neural networks without any denoising. These noise data seriously affect the prediction accuracy of the methods. To maintain stability and robustness, input data must be denoised before being fed into deep neural networks. DAE, an unsupervised method in learning useful features, which is adopted by our method, reconstructs input data from lower-dimensional representation preserving as much information as possible in the process.

Let $\mathbf{x}'_t = \left\{x'_{t+1}, x'_{t+2}, \ldots, x'_{t+m}\right\} \in \mathbf{x}'$ denote the slice of input with $m$ samples of a sequence. Gaussian noise is added to the normalized input to obtain the corrupted vector, $\widetilde{\mathbf{x}}_t$. DAE serves two purposes: denoising the raw input and learning nonlinear representation:

$$\mathbf{z} = a\left(\boldsymbol{W}^T\widetilde{\mathbf{x}}_t + \boldsymbol{b}\right), \tag{3}$$

where $\boldsymbol{W}$, $\boldsymbol{b}$, $a(\cdot)$, and $\boldsymbol{z}$ denote weight, bias, activation function, and output of the DAE encoder, respectively.

Then, to reconstruct the input vector, the latent representation is mapped back to the input space, defined as follows:

$$\widehat{\mathbf{x}}_t = f'\left(\boldsymbol{W}'\boldsymbol{z} + \boldsymbol{b}'\right), \tag{4}$$

where $\boldsymbol{W}', \boldsymbol{b}', \boldsymbol{z}$ and $f'(\cdot)$ denote weight, bias, output, and map function of the output layer of the DAE encoder, respectively.

In our network, *identity* and *ReLU* functions are used as the decoding and encoding activation, respectively. Finally, the objective function is defined as follows:

$$\mathcal{L}_d = \frac{1}{n}\sum_{t=1}^{n} \ell(\widetilde{\boldsymbol{x}}_t - \widehat{\boldsymbol{x}}_t) + \lambda\left(\|\boldsymbol{W}\|_F^2 + \left\|\boldsymbol{W}'\right\|_F^2\right), \tag{5}$$

where $\ell(\cdot)$ denotes a loss function.

Because the structure of DAE is symmetrical, some weights can be tied, i.e. $\boldsymbol{W} = \boldsymbol{W}'$, thereby accelerating training by reducing the number of weights of the model.

#### 3) TRANSFORMER

The standard Transformer is a sequence-to-sequence architecture, consisting of an encoder and a decoder. The encoder takes the input sequence and maps it into a higher dimensional vector, which is then fed into the decoder to generate an output sequence. In this paper, the encoder of the Transformer is used to learn long-term dependencies of the capacity degradation from battery working records.

The Transformer layers are a stack of Transformer encoders that extract the degradation features from the reconstructed data, with two sub-layers: *Multi-Head Self-Attention* and *Feed-Forward*. To fully use the position information of the sequence, we inject some relative position tokens into the sequence. In this paper, we use sine and cosine functions of different frequencies [60]:

$$PE(t, 2k) = sin(t/10000^{2k/m}) \tag{6}$$
$$PE(t, 2k+1) = cos(t/10000^{2k/m}), \tag{7}$$

where $t$ denotes the time step.

The *Multi-Head Self-Attention* sub-layer aims to capture the dependencies between features and ignores their distances in the sequence [54]–[57]. Given the representation of the $(l-1)$-th layer, $\boldsymbol{H}^{l-1}$ and $h$ parallel attention functions, the $i$-th ($i \in [1, h]$) attention is defined:

$$head_i = Attention\left(\boldsymbol{H}^{l-1}\boldsymbol{W}_Q^l, \boldsymbol{H}^{l-1}\boldsymbol{W}_K^l, \boldsymbol{H}^{l-1}\boldsymbol{W}_V^l\right), \tag{8}$$

where $\boldsymbol{W}_Q^l$, $\boldsymbol{W}_K^l$, and $\boldsymbol{W}_V^l \in \mathbb{R}^{d\times d_h}$ are projection weights. Let $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ denote *query, key*, and *value*, *Scaled Dot-Product Attention* defined as follows::

$$Attention\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_h}}\right)\boldsymbol{V}, \tag{9}$$

where $d_h = d/h$, which avoids avoiding extremely small gradients and produces a softer attention distribution [60].

Then, the *Multi-Head Attention* is defined as follows:

$$MultiHead\left(\boldsymbol{H}^{l-1}\right) = [head_1; head_2; \cdots; head_h]\boldsymbol{W}^O, \tag{10}$$

where $\boldsymbol{W}^O$ is a trainable weight.

*Feed-Forward*, which has two different mappings (linear and *ReLU* nonlinear), is applied to each time step identically and separately. Then, we obtain $\boldsymbol{H}^l$ from the previous $MultiHead\left(\boldsymbol{H}^{l-1}\right)$ as follows:

$$\boldsymbol{H}^l = FFN\left(MultiHead\left(\boldsymbol{H}^{l-1}\right)\right), \tag{11}$$
$$FFN\left(\boldsymbol{x}\right) = ReLU\left(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1\right)\boldsymbol{W}_2 + \boldsymbol{b}_2. \tag{12}$$

#### 4) PREDICTION

Finally, to predict unknown capacity, a full connection layer is used to map the representation learned by the last Transformer cell to arrive at the final prediction $\widehat{x}_t$, namely, $\widehat{x}_{i+m+1}$:

$$\widehat{x}_t = f\left(\boldsymbol{W}_p\boldsymbol{H}^h + \boldsymbol{b}_p\right), \tag{13}$$

where $\boldsymbol{W}_p$, $\boldsymbol{b}_p$, $\boldsymbol{H}^h$, and $f(\cdot)$ denote weight, bias, input, and map function of the prediction layer, respectively.

### B. LEARNING

There are two tasks in our model: denoising and prediction. Instead of solving these two tasks separately, we propose an objective function to bridge these tasks. The learning procedure optimizes both tasks simultaneously in a unified framework. Mean Square Error (MSE) is used to evaluate loss, and the objective function is defined as follows:

$$\mathcal{L} = \sum_{t=T+1}^{n}(x_t - \widehat{x}_t)^2 + \alpha\sum_{i=1}^{n}\ell(\widetilde{\mathbf{x}}_i - \widehat{\mathbf{x}}_i) + \lambda\Omega\left(\Theta\right), \tag{14}$$

where $\alpha$ denotes a parameter to control the relative contribution of each task; $\lambda$ denotes a regularization parameter; $\Omega(\cdot)$ denotes the regularization; $\Theta$ denotes the learning parameters of our model.

## IV. EXPERIMENTS

### A. EXPERIMENTAL SETTINGS

#### 1) DATA SETS

We conducted experiments using two public data sets: NASA and CALCE. The NASA data set, available from the NASA Ames Research Center web site,[1] contains the record of four different Li-ion batteries, with each Li-ion battery repeating three operations: charge, discharge, and impedance measurements [61], [62]. Similarly, the CALCE data set is available from the Center for Advanced Life Cycle Engineering (CALCE) of the University of Maryland[2] [63]–[65].

#### 2) BASELINE APPROACHES

We compared our models to the following methods:

- **MLP** [2], with multiple fully connected layers, is used to learn the dynamic and nonlinear degradation trend of a battery.
- **RNN** [27], with multiple RNN units, is used to predict RUL.

[1] https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#battery

[2] https://calce.umd.edu/data#CS2

**TABLE 2.** Overall performance on NASA and CALCE data sets.

| NASA Data Set | | | | | | |
|---|---|---|---|---|---|---|
| | MLP | RNN | LSTM | GRU | Dual-LSTM | DeTransformer |
| RE | 0.3851 | 0.2851 | 0.2648 | 0.3044 | 0.2557 | **0.2252** |
| MAE | 0.1379 | 0.0749 | 0.0829 | 0.0806 | 0.0815 | **0.0713** |
| RMSE | 0.1541 | 0.0848 | 0.0905 | 0.0921 | 0.0879 | **0.0802** |
| CALCE Data Set | | | | | | |
| | MLP | RNN | LSTM | GRU | Dual-LSTM | DeTransformer |
| RE | 0.4018 | 0.1614 | 0.0902 | 0.1319 | 0.0885 | **0.0764** |
| MAE | 0.1557 | 0.0938 | **0.0582** | 0.0671 | 0.0636 | 0.0613 |
| RMSE | 0.2038 | 0.1099 | 0.0736 | 0.0946 | 0.0874 | **0.0705** |

- **LSTM** [8], with multiple LSTM units, is used to learn the degradation trend from the input sequence.
- **GRU** [66], with multiple GRU units, is used to learn features from sequences.
- **Dual-LSTM** [52], with two different LSTM cells in point detection, is used to capture the non-linearity between capacities.

#### 3) EVALUATION METRICS

First, the three evaluation metrics used to evaluate the prediction performance of RUL are the following: Relative Error (RE), Mean Absolute Error (MAE) and RMSE. The three evaluation metrics are defined as follows:

$$RE = \frac{|RUL^{pred} - RUL^{true}|}{RUL^{true}} \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n-T}\sum_{t=T+1}^{n}(x_t - \widehat{x}_t)^2} \tag{16}$$

$$MAE = \frac{1}{n-T}\sum_{t=T+1}^{n}\|x_t - \widehat{x}_t\| \tag{17}$$

where $n$ denotes the length of a sequence, and $T$ denotes the length of samples generated from a sequence for training.

Then, a leave-one-out evaluation is used to evaluate our models: one battery is sampled randomly; the remainder are used for training. Finally, after five iterations, the average score over all batteries is determined.

#### 4) PARAMETER SETTINGS

Our model has six key parameters: sampler size ($m$), learning rate ($\tau$), depth ($l$) and hidden size ($s$) of Transformer, regularization for learning ($\lambda$), and ratio of each task ($\alpha$). $m$ can be set about 5∼10% of the length of a sequence. In our experiments, $m$ is fixed at 16 and 64 for NASA and CALCE, respectively.

The rest parameters were determined by grid-search on the validation error: $\tau$ is chose from $\{10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}\}$; $s$ is chose from $\{8, 16, 32, 64\}$; $l$ is chose from $\{1, 2, 3, 4\}$; $\lambda$ is chose from $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$; $\alpha$ is set from $(0, 1]$.

Because RE is highly related to the RUL of a battery, we chose RE as our major evaluation metric. In terms of the RE, optimal parameters of all methods for the two data sets are listed in Table 3. All codes are run on Pytorch 1.8.0, Python 3.7, and Cenos 7 Systems with i9 CPU.

**TABLE 3.** Optimal parameters of RE scores for two data sets.

| Data sets | Models | $m$ | $\tau$ | $l$ | $s$ | $\lambda$ |
|---|---|---|---|---|---|---|
| NASA | MLP | 16 | 0.01 | 2 | 8 | $10^{-6}$ |
| | RNN | 16 | 0.001 | 2 | 64 | $10^{-6}$ |
| | LSTM | 16 | 0.001 | 2 | 64 | $10^{-6}$ |
| | GRU | 16 | 0.001 | 2 | 64 | $10^{-6}$ |
| | Dual-LSTM | 16 | 0.001 | 2 | 64 | $10^{-6}$ |
| | Our Model | 16 | 0.005 | 1 | 32 | $10^{-6}$ |
| CALCE | MLP | 64 | 0.01 | 4 | 64 | $10^{-6}$ |
| | RNN | 64 | 0.001 | 2 | 32 | $10^{-6}$ |
| | LSTM | 64 | 0.001 | 2 | 32 | $10^{-6}$ |
| | GRU | 64 | 0.001 | 2 | 32 | $10^{-6}$ |
| | Dual-LSTM | 64 | 0.001 | 2 | 32 | $10^{-6}$ |
| | Our Model | 64 | 0.001 | 1 | 32 | $10^{-6}$ |

### B. RESULTS AND ANALYSIS

#### 1) OVERALL PERFORMANCE

First, experiments were conducted to verify the performance of our methods on different data sets. Table 2 shows the RE, MAE and RMSE scores obtained for all methods. The best results are shown in bold.

From the results shown in Table 2, we conclude the following: (1) Among all methods, our models achieve the best experimental results. The results demonstrate that our model extracts useful temporal information in the modeling capacity sequences. (2) On both data sets, DeTransformer is stable and robust and always makes good predictions, regardless of whether a capacity sequence is long or short. Also, DeTransformer shows an even greater improvement on NASA. Possibly, for networks, a long sequence offers sufficient information to train; however, representation ability is limited when only one feature is fed into the network for a short sequence. (3) Among the baseline methods, MLP performs the worst, because it fails to take into account the effects of temporal information. Our model and all RNN-based models predict trends better than MLP, which means that it is necessary to add sequential information to predict RUL well. The attention networks of Transformer capture the overall

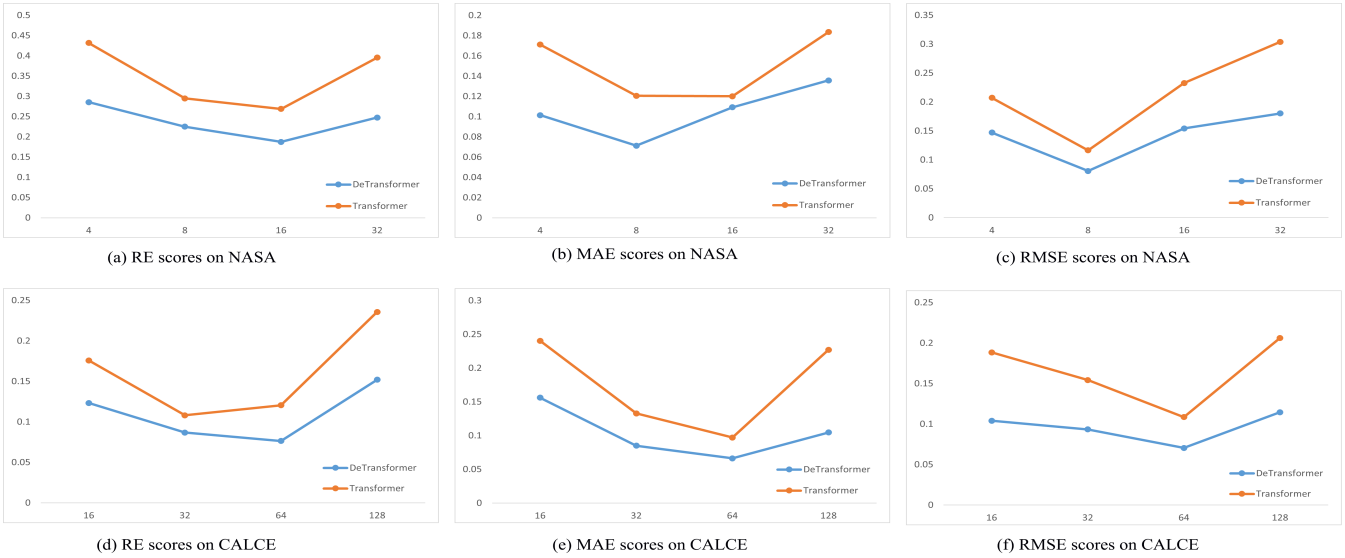



**FIGURE 4.** Effect of autoencoder.

inclination by modeling correlations among historical capacity features. Thus, our model simulates well the effects of historical capacities in sequence states. (4)RNN achieves better scores for MAE and RMSE on NASA, but, worse on CALCE when compared with LSTM, GRU, and Dual-LSTM. The best possible reason is that the sequence length is different in the two data sets. LSTM and GRU are better at learning features from long sequences than RNN, which is also a deficiency in the nature of RNN. In all metrics, DeTransformer does exceptionally well on RE, which is directly related to the RUL of a battery. Potentially, the reason is that battery charge/discharge regeneration degrades the learning of the model on the trends. To refine the representation, our models reduce noise in a raw sequence with an autoencoder. In summary, our method outperforms other competitive approaches, which suggests that our method is effective for extracting meaningful temporal features to more accurately predict the RUL of a battery.

### 2) EFFECT OF AUTOENCODER

Then, we demonstrate the improvement in performance by using an autoencoder. We compared our models with their simplified versions without an autoencoder by setting different values for the hidden size. The average scores of RE, MAE, and RMSE with changing hidden sizes are shown in Fig. 4.

From Fig. 4, it is seen that all scores first increase and then decrease with an increase in hidden size. The most probable reason is that Transformer has limited weights to obtain sufficient temporal information, leading to under-fitting when hidden size is too small. When hidden size is too large, Transformer has too many weights to learn temporal information. Also, for all metrics, with an increase in the hidden size of Transformer in most cases, our models performed better than their simplified versions, indicating that an added autoencoder improves performance in the prediction of RUL.

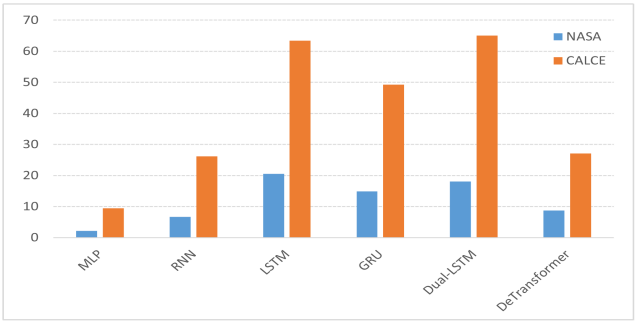



**FIGURE 5.** Time cost (seconds) of all neural networks on two data sets.

The nonlinear capacity curve contains much noise, especially when capacity regeneration occurs. Most neural networks are trained directly on the raw data, which influences the model in learning representation. However, to make better predictions, our models are trained on the refined data generated by DAE, which has the powerful ability to learn useful features from input with much noise. Therefore, an autoencoder shows strong improvement over our methods.

### 3) TIME COST

Finally, we studied the time cost of all neural networks on two data sets (See Fig. 5). From Fig. 5, it is seen that the training time of LSTM, GRU, and Dual-LSTM, is much longer than for the other models. Potentially, the reason is that all RNN-based networks modeled on sequences in a recurrent manner lead to higher time costs for training and inference. Although MLP is rapid, it does not work very well experimentally on sequential data. RNN and DeTransformer are very close in training speed. However, our approach always yields the best results. To learn trends from sequences, a multi-head attention network of a Transformer accelerates the training performance of neural networks. Finally, we conclude that, with a multi-head attention network applied in our model,

Transformer learns features in parallel, which are more suitable for predicting RUL.

## V. CONCLUSION

With a RUL predictor, an accurate estimation of RUL, a safer battery system, and longer battery service life can be achieved. We proposed a novel neural network model for RUL prediction. First, DAE was used to learn representation from corrupted input and then used to reconstruct input. Second, from the reconstructed input, Transformer networks were used to learn the feature for capacity fading. Finally, we designed an objective function, which combines jointly DAE loss and prediction loss. Compared with existing RUL methods, our models achieve better performance as indicated by lower RE, MAE, and RMSE scores.

In the future, we plan to extend our methods to more practical applications. First, training a model using part of a record may not be robust enough and may be lopsided. Thus, to fully train, more charge-discharge data will be added to our model. Also, in practice, a battery will be examined under operational conditions, such as different working temperatures and currents, which have a large impact on degradation trends. Thus, an estimation of RUL for a battery under different operating conditions will be studied further.

## ACKNOWLEDGMENT

The authors would like to thank Michael McAllister for proofreading this manuscript.

## REFERENCES

[1] Y. Song, D. Liu, C. Yang, and Y. Peng, "Data-driven hybrid remaining useful life estimation approach for spacecraft lithium-ion battery," *Microelectron. Rel.*, vol. 75, pp. 142–153, Aug. 2017.

[2] Y. Wu, W. Li, Y. Wang, and K. Zhang, "Remaining useful life prediction of lithium-ion batteries using neural network and bat-based particle filter," *IEEE Access*, vol. 7, pp. 54843–54854, 2019.

[3] L. Wu, X. Fu, and Y. Guan, "Review of the remaining useful life prognostics of vehicle lithium-ion batteries using data-driven methodologies," *Appl. Sci.*, vol. 6, no. 6, p. 166, 2016.

[4] A. Samanta, S. Chowdhuri, and S. S. Williamson, "Machine learning-based data-driven fault detection/diagnosis of lithium-ion battery: A critical review," *Electronics*, vol. 10, no. 11, p. 1309, May 2021.

[5] H. Meng and Y.-F. Li, "A review on prognostics and health management (PHM) methods of lithium-ion batteries," *Renew. Sustain. Energy Rev.*, vol. 116, Dec. 2019, Art. no. 109405.

[6] D. Wang, K.-L. Tsui, and Q. Miao, "Prognostics and health management: A review of vibration based bearing and gear health indicators," *IEEE Access*, vol. 6, pp. 665–676, 2018.

[7] J.-Z. Kong, F. Yang, X. Zhang, E. Pan, Z. Peng, and D. Wang, "Voltage-temperature health feature extraction to improve prognostics and health management of lithium-ion batteries," *Energy*, vol. 223, May 2021, Art. no. 120114.

[8] Y. Zhang, R. Xiong, H. He, and M. G. Pecht, "Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 5695–5705, Jul. 2018.

[9] H. Dong, X. Jin, Y. Lou, and C. Wang, "Lithium-ion battery state of health monitoring and remaining useful life prediction based on support vector regression-particle filter," *J. Power Sources*, vol. 271, pp. 114–123, Dec. 2014.

[10] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, "Long short-term memory network for remaining useful life estimation," in *Proc. IEEE Int. Conf. Prognostics Health Manage. (ICPHM)*, Jun. 2017, pp. 88–95.

[11] B. Zhou, C. Cheng, G. Ma, and Y. Zhang, "Remaining useful life prediction of lithium-ion battery based on attention mechanism with positional encoding," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 895, no. 1, 2020, Art. no. 012006.

[12] Y. Zhang, R. Xiong, H. He, X. Qu, and M. Pecht, "Aging characteristics-based health diagnosis and remaining useful life prognostics for lithium-ion batteries," *eTransportation*, vol. 1, Aug. 2019, Art. no. 100004.

[13] S. S. Y. Ng, Y. Xing, and K. L. Tsui, "A naive Bayes model for robust remaining useful life prediction of lithium-ion battery," *Appl. Energy*, vol. 118, pp. 114–123, Apr. 2014.

[14] A. Nuhic, T. Terzimehic, T. Soczka-Guth, M. Buchholz, and K. Dietmayer, "Health diagnosis and remaining useful life prognostics of lithium-ion batteries using data-driven methods," *J. Power Sources*, vol. 239, pp. 680–688, Oct. 2013.

[15] D. Liu, J. Zhou, H. Liao, Y. Peng, and X. Peng, "A health indicator extraction and optimization framework for lithium-ion battery degradation modeling and prognostics," *IEEE Trans. Syst., Man, Cybern. Syst*, vol. 45, no. 6, pp. 915–928, Jun. 2015.

[16] M. Chen, T. Ma, and X. Zhou, "CoCNN: Co-occurrence CNN for recommendation," *Expert Syst. Appl.*, vol. 195, Jun. 2022, Art. no. 116595.

[17] M. Chen, Y. Li, and X. Zhou, "CoNet: Co-occurrence neural networks for recommendation," *Future Gener. Comput. Syst.*, vol. 124, pp. 308–314, Nov. 2021.

[18] M. Chen and X. Zhou, "DeepRank: Learning to rank with neural networks for recommendation," *Knowl.-Based Syst.*, vol. 209, Dec. 2020, Art. no. 106478.

[19] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, Feb. 2018.

[20] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[21] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[22] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[23] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604–624, Feb. 2021.

[24] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48–57, May 2014.

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

[26] A. Khalid and A. I. Sarwat, "Unified univariate-neural network models for lithium-ion battery state-of-charge forecasting using minimized Akaike information criterion algorithm," *IEEE Access*, vol. 9, pp. 39154–39170, 2021.

[27] J. Liu, A. Saxena, K. Goebel, B. Saha, and W. Wang, "An adaptive recurrent neural network for remaining useful life prediction of lithium-ion batteries," in *Proc. Annu. Conf. Prognostics Health Manage. Soc.*, 2010, pp. 1–9.

[28] N. Gugulothu, V. Tv, P. Malhotra, L. Vig, P. Agarwal, and G. Shroff, "Predicting remaining useful life using time series embeddings based on recurrent neural networks," *Int. J. Prognostics Health Manage.*, vol. 9, no. 1, pp. 1–10, Nov. 2020.

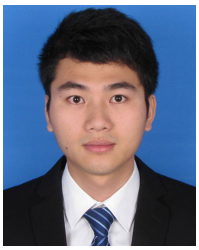
[29] M. Catelani, L. Ciani, R. Fantacci, G. Patrizi, and B. Picano, "Remaining useful life estimation for prognostics of lithium-ion batteries based on recurrent neural network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[30] A. Khalid, A. Sundararajan, I. Acharya, and A. I. Sarwat, "Prediction of Li-ion battery state of charge using multilayer perceptron and long short-term memory models," in *Proc. IEEE Transp. Electrific. Conf. Expo. (ITEC)*, Jun. 2019, pp. 1–6.

[31] K. Park, Y. Choi, W. J. Choi, H.-Y. Ryu, and H. Kim, "LSTM-based battery remaining useful life prediction with multi-channel charging profiles," *IEEE Access*, vol. 8, pp. 20786–20798, 2020.

[32] Y. Mo, Q. Wu, X. Li, and B. Huang, "Remaining useful life estimation via transformer encoder enhanced by a gated convolutional unit," *J. Intell. Manuf.*, vol. 32, no. 7, pp. 1997–2006, 2021.

[33] J. Hao, X. Wang, B. Yang, L. Wang, J. Zhang, and Z. Tu, "Modeling recurrence for transformer," in *Proc. Conf. North*, 2019, pp. 1198–1207.
[34] E. Egonmwan and Y. Chali, "Transformer and seq2seq model for paraphrase generation," in *Proc. 3rd Workshop Neural Gener. Transl.*, 2019, pp. 249–255.
[35] T. Qin, S. Zeng, J. Guo, and Z. Skaf, "A rest time-based prognostic framework for state of health estimation of lithium-ion batteries with regeneration phenomena," *Energies*, vol. 9, no. 11, p. 896, Nov. 2016.
[36] L. Ren, J. Dong, X. Wang, Z. Meng, L. Zhao, and M. J. Deen, "A data-driven auto-CNN-LSTM prediction model for lithium-ion battery remaining useful life," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3478–3487, May 2021.
[37] Y.-W. Lu, C.-Y. Hsu, and K.-C. Huang, "An autoencoder gated recurrent unit for remaining useful life prediction," *Processes*, vol. 8, no. 9, p. 1155, Sep. 2020.
[38] C. Wang, N. Lu, S. Wang, Y. Cheng, and B. Jiang, "Dynamic long short-term memory neural-network-based indirect remaining-useful-life prognosis for satellite lithium-ion battery," *Appl. Sci.*, vol. 8, no. 11, p. 2078, Oct. 2018.
[39] L. Lu, X. Han, J. Li, J. Hua, and M. Ouyang, "A review on the key issues for lithium-ion battery management in electric vehicles," *J. Power Sources*, vol. 226, pp. 272–288, Mar. 2013.
[40] K. Goebel, B. Saha, A. Saxena, J. R. Celaya, and J. P. Christophersen, "Prognostics in battery health management," *IEEE Instrum. Meas. Mag.*, vol. 11, no. 4, pp. 33–40, Aug. 2008.
[41] Y. Li, P. Chattopadhyay, S. Xiong, A. Ray, and C. D. Rahn, "Dynamic data-driven and model-based recursive analysis for estimation of battery state-of-charge," *Appl. Energy*, vol. 184, pp. 266–275, Dec. 2016.
[42] X. Lai, W. Yi, Y. Cui, C. Qin, X. Han, T. Sun, L. Zhou, and Y. Zheng, "Capacity estimation of lithium-ion cells by combining model-based and data-driven methods based on a sequential extended Kalman filter," *Energy*, vol. 216, Feb. 2021, Art. no. 119233.
[43] L. Ungurean, G. Cârstoiu, M. V. Micea, and V. Groza, "Battery state of health estimation: A structured review of models, methods and commercial devices," *Int. J. Energy Res.*, vol. 41, no. 2, pp. 151–181, 2017.
[44] K. Sun and Q. Shu, "Overview of the types of battery models," in *Proc. 30th Chin. Control Conf.*, Jul. 2011, pp. 3644–3648.
[45] X. Hu, F. Sun, and Y. Zou, "Comparison between two model-based algorithms for Li-ion battery SOC estimation in electric vehicles," *Simul. Model. Pract. Theory*, vol. 34, pp. 1–11, 2013.
[46] H. He, Y. Zhang, R. Xiong, and C. Wang, "A novel Gaussian model based battery state estimation approach: State-of-energy," *Appl. Energy*, vol. 151, pp. 41–48, Aug. 2015.
[47] G. Zhao, G. Zhang, Y. Liu, B. Zhang, and C. Hu, "Lithium-ion battery remaining useful life prediction with deep belief network and relevance vector machine," in *Proc. IEEE Int. Conf. Prognostics Health Manage. (ICPHM)*, Jun. 2017, pp. 7–13.
[48] G.-W. You, S. Park, and D. Oh, "Real-time state-of-health estimation for electric vehicle batteries: A data-driven approach," *Appl. Energy*, vol. 176, pp. 92–103, Aug. 2016.
[49] B. Gou, Y. Xu, and X. Feng, "State-of-health estimation and remaining-useful-life prediction for lithium-ion battery using a hybrid data-driven method," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, p. 10854–10867, 2020.
[50] K. Liu, Y. Shang, Q. Ouyang, and W. D. Widanage, "A data-driven approach with uncertainty quantification for predicting future capacities and remaining useful life of lithium-ion battery," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3170–3180, Apr. 2021.
[51] X. Wu, W. Zeng, F. Lin, and X. Zhou, "NeuRank: Learning to rank with neural networks for drug–target interaction prediction," *BMC Bioinf.*, vol. 22, no. 1, pp. 1–17, Dec. 2021.
[52] Z. Shi and A. Chehade, "A dual-LSTM framework combining change point detection and remaining useful life prediction," *Rel. Eng. Syst. Saf.*, vol. 205, Jan. 2021, Art. no. 107257.
[53] J. Hong, D. Lee, E.-R. Jeong, and Y. Yi, "Towards the swift prediction of the remaining useful life of lithium-ion batteries with end-to-end deep learning," *Appl. Energy*, vol. 278, Nov. 2020, Art. no. 115646.
[54] L. Wu, S. Li, C.-J. Hsieh, and J. Sharpnack, "SSE-PT: Sequential recommendation via personalized transformer," in *Proc. 14th ACM Conf. Recommender Syst.*, Sep. 2020, pp. 328–337.
[55] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 1441–1450.
[56] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4055–4064.
[57] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10578–10587.
[58] J. W. Song, Y. I. Park, J.-J. Hong, S.-G. Kim, and S.-J. Kang, "Attention-based bidirectional LSTM-CNN model for remaining useful life estimation," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2021, pp. 1–5.
[59] Y. Song, L. Li, Y. Peng, and D. Liu, "Lithium-ion battery remaining useful life prediction based on GRU-RNN," in *Proc. 12th Int. Conf. Rel., Maintainability, Saf. (ICRMS)*, Oct. 2018, pp. 317–322.
[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[61] B. Saha and K. Goebel, *Battery Data Set*. Moffett Field, CA, USA: NASA Ames Prognostics Data Repository, NASA Ames Research Center, 2008.
[62] B. Saha and K. Goebel, "Uncertainty management for diagnostics and prognostics of batteries using Bayesian techniques," in *Proc. IEEE Aerosp. Conf.*, Mar. 2008, pp. 1–8.
[63] W. He, N. Williard, M. Osterman, and M. Pecht, "Prognostics of lithium-ion batteries based on Dempster–Shafer theory and the Bayesian Monte Carlo method," *J. Power Sour.*, vol. 196, pp. 10314–10321, Dec. 2011.
[64] Y. Xing, E. W. M. Ma, K.-L. Tsui, and M. Pecht, "An ensemble model for predicting the remaining useful performance of lithium-ion batteries," *Microelectron. Rel.*, vol. 53, pp. 811–820, Jun. 2013.
[65] N. Williard, W. He, M. Osterman, and M. Pecht, "Comparative analysis of features for determining state of health in lithium-ion batteries," *Int. J. Prognostics Health Manage.*, vol. 4, no. 1, pp. 1–7, Oct. 2020.
[66] B. Xiao, Y. Liu, and B. Xiao, "Accurate state-of-charge estimation approach for lithium-ion batteries by gated recurrent unit with ensemble optimizer," *IEEE Access*, vol. 7, pp. 54192–54202, 2019.



**DAOQUAN CHEN** received the B.S. degree from the Zhejiang University of Science and Technology, in 2012, and the M.S. degree from Zhejiang University, in 2015. He is currently a Researcher with the Zhejiang Institute of Mechanical and Electrical Engineering. His current research interests include machine learning, deep learning, and automatic control.

**WEICONG HONG** received the M.S. degree from the Zhejiang University of Technology, in 2018. He is currently a Cloud Computing Engineer with Shuye Technology Company Ltd., China. His current research interests include cloud computing and the IoT applications.

**XIUZE ZHOU** received the M.S. degree from Xiamen University, in 2016. He is currently working as the Chief Technology Officer at Shuye Technology Company Ltd., China. His current research interests include machine learning, computer vision, and recommender systems.