Front. Comput. Sci., 2024, 18(5): 185329
https://doi.org/10.1007/s11704-023-3277-4

RESEARCH ARTICLE

A MLP-Mixer and mixture of expert model for remaining useful life prediction of lithium-ion batteries

Lingling ZHAO[1], Shitao SONG[2], Pengyan WANG[3], Chunyu WANG[1], Junjie WANG (✉)[4], Maozu GUO (✉)[5]

1 Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China
2 School of Electrical Engineering, Liaoning University of Technology, Jinzhou 121004, China
3 School of Computer Science, Northeast Electric Power University, Jilin 132000, China
4 Department of Medical Informatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China
5 School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

Abstract Accurately predicting the Remaining Useful Life (RUL) of lithium-ion batteries is crucial for battery management systems. Deep learning-based methods have been shown to be effective in predicting RUL by leveraging battery capacity time series data. However, the representation learning of features such as long-distance sequence dependencies and mutations in capacity time series still needs to be improved. To address this challenge, this paper proposes a novel deep learning model, the MLP-Mixer and Mixture of Expert (MMMe) model, for RUL prediction. The MMMe model leverages the Gated Recurrent Unit and Multi-Head Attention mechanism to encode the sequential data of battery capacity to capture the temporal features and a re-zero MLP-Mixer model to capture the high-level features. Additionally, we devise an ensemble predictor based on a Mixture-of-Experts (MoE) architecture to generate reliable RUL predictions. The experimental results on public datasets demonstrate that our proposed model significantly outperforms other existing methods, providing more reliable and precise RUL predictions while also accurately tracking the capacity degradation process. Our code and dataset are available at the website of github.

Keywords lithium-ion battery, remaining useful life, deep learning, MLP-Mixer, mixture-of-experts

1 Introduction

Lithium-Ion Batteries (LIBs) have been widely used in various applications, such as Electric Vehicles (EVs), Automated Guided Vehicles (AGVs), and aerospace, owing to their outstanding characteristics, including high energy density, long service life, and low self-discharge rate [1]. However, the performance of LIBs gradually deteriorates as they age, when the maximum discharging capacity of a LIB degrades to 70%–80% of its rated capacity, the LIB reaches its End-of-Life (EOL). Using a battery beyond its EOL may cause equipment failure or even catastrophic occurrences [2]. Therefore, to ensure the safety and reliability of battery-powered systems, developing a method that accurately estimates the Remaining Useful Life (RUL) of LIBs is imperative to provide early warning of battery failure and ensure reliable battery operation.

Over the past decade, there have been extensive approaches to predicting RUL of Lithium-Ion Batteries (LIBs). Generally, these methodologies can be divided into two main groups: model-based methods and data-driven methods [3,4]. Model-based methods construct mathematical or physical models from measurement data to capture battery fading dynamics [5,6]. These methods often require prior knowledge to describe the internal physical-chemical reaction mechanisms of the LIB, such as electrochemical models [7] and equivalent circuit models [8], along with estimation/filtering observers like Kalman filters [9], and particle filters [10,11] to determine model parameters. However, due to the complex internal reaction mechanism of batteries, it is difficult to establish a robust mathematical or physical model under various working conditions.

Recently, a plethora of data-driven solutions have emerged in the literature, relying on statistical analysis or artificial intelligence. Data-driven methods do not require prior knowledge of LIBs but instead employ related data and select suitable learning algorithms to directly model and predict RUL [12]. Various data-driven techniques have been proposed for RUL prediction, such as Support Vector Regression (SVR) [13,14], Extreme Learning Machine (ELM) [15], and Wiener process models [16]. The ever-growing availability of data sources and computational resources has enabled the advancement of deep learning techniques for RUL prediction,

Received April 7, 2023; accepted July 18, 2023

E-mail: junjie2021@njmu.edu.cn; guomaozu@bucea.edu.cn

leading to exponential growth in the domain. Deep learning methods accurately capture intricate and nonlinear relationships between input and output. As a result, popular deep learning architectures such as Convolutional Neural Networks (CNNs) [17,18], Recurrent Neural Networks (RNNs) [19,20], and Transformer [21] are extensively employed in RUL prediction for LIBs. In particular, deep learning methods based on CNNs utilize one-dimensional convolutional neural networks to capture the relationship between observed data and RUL [22]. However, their ability to predict RUL is still limited, as they are not adept at modeling long-range sequential dependencies in sensory data [23].

A Recurrent Neural Network (RNN) is a classical deep learning network that considers both the current input and past hidden state as input. By accounting for the number of time steps, RNN is capable of effectively storing information from the past, making it well-suited for time-series analysis. Recently, RUL prediction models based on RNN and its variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have received increasing attention in the literature [24−27]. To improve the performance of RNN-based models, the attention mechanism has been incorporated into LSTM, leading to further improvements in prediction accuracy [28].

Recently, several deep learning models based on Multi-Layer Perceptron (MLP), including MLP-Mixer [29], ResMLP [30], CycleMLP [31], and S^2MLP [32], have regained attention in the literature. The pioneering works MLP-Mixer and ResMLP stack two types of MLP layers, namely, token-mixing MLP and channel-mixing MLP, alternately. The token-mixing MLP encodes spatial information by interacting across all tokens, while the channel-mixing MLP mixes information across all channels of each token. The results of these studies have shown that MLP can be comparable to the Transformer in computer vision tasks, which has had a positive impact on the field and encouraged researchers to apply MLP to RUL prediction.

Although deep learning-based methods have demonstrated promising results in estimating the RUL, most methods consider that each time step's features hold equal importance. It should be noted that certain early steps in the sequence may carry more significant contributions to the final RUL prediction [28]. When data with varying degrees of contribution are treated with equal weights, it can potentially restrict the model's feature extraction capability. In light of this, there is value in exploring methods that can effectively handle the varying contributions of different time steps in the RUL estimation.

In this paper, we propose a novel approach called MMMe for the accurate prediction of the RUL of lithium-ion batteries. Specifically, MMMe first captures sequence information of the inputs through a Bi-directional Gated Recurrent Unit with Multi-Head Attention (BiGRU-MHA). Subsequently, the MLP-Mixer module is introduced, which allows for the communication of learned features in both temporal and channel directions. Additionally, an ensemble predictor based on the Mixture-of-Expert architecture [33] is devised to determine the RUL value. The experiments conducted show that MMMe has achieved state-of-the-art results on two public datasets, demonstrating the competitiveness of the proposed approach for RUL prediction.

2 Problem formulation

Battery RUL refers to the estimated number of charging and discharging cycles before a battery reaches its EOL. For batteries, EOL is closely related to their capacity. On this basis, this paper defines the RUL of the battery as the time interval between the current moment and the time when the State of Health (SOH) drops to 80% for the first time.

Data-driven methods for predicting the RUL aim to establish a mapping model f between available data x related to battery life and RUL y. Depending on the nature of x, there are two problem formulations that can be further distinguished. 1) x represents monitoring data that indirectly correlates with RUL, such as current, voltage, temperature, and other observables in each charging/discharging cycle. In this case, f attempts to express the relationship between these observables and RUL. 2) x represents the battery capacity time series, i.e., $x = \{c_1, c_2, \ldots, c_k\}$, where k is the length of the known time series fragment. Then the mapping f between x and the future capacity time series fragment $x' = \{c_{k+1}, c_{k+2}, \ldots, c_{k+m}\}$ are attempted to be established, where m is the length of the unknown time series fragment to be predicted. By the function $x' = f(x)$, x' can be iteratively updated by switching the predicted time series fragment into the "known" time series fragment until a c_{k+n} in the currently predicted time series fragment reaches the EOL condition. At this point, the RUL y can be estimated. This formulation essentially transforms the RUL prediction into a time series prediction.

In this paper, we address the RUL prediction based on deep learning in the second problem formulation framework. Historical data is first used to construct N sample pairs (x_i, x_i'), and then a deep neural network model is trained using these labeled data. This allows one to use the optimized neural network model to predict the battery capacity for future m time steps and estimate the RUL given a fixed-length battery capacity time series x.

3 Methodology

In this section, we introduce a deep learning framework called MLP-Mixer with Mixture of Expert (MMMe) for predicting RUL. The MMMe pipeline is illustrated in Fig. 1. The MMMe framework begins with a Bi-directional Gated Recurrent Unit (BiGRU) with Multi-Head Attention (MHA) block, which aims to extract the temporal features $X \in \mathcal{R}^{T\times C}$ from the input battery sequences $X \in \mathcal{R}^{T\times D}$. The learned temporal features $X \in \mathcal{R}^{T\times C}$ enable the subsequent Mixer block to capture the order of sequence data. Then, the MLP-Mixer block facilitates communication between channel and temporal information by employing two MLPs. Finally, the predicted RUL value is generated by a Mixture-of-Experts (MoE) layer with a gating function that combines the predictions from different subnetworks (experts) based on the learned features from the Mixer block.

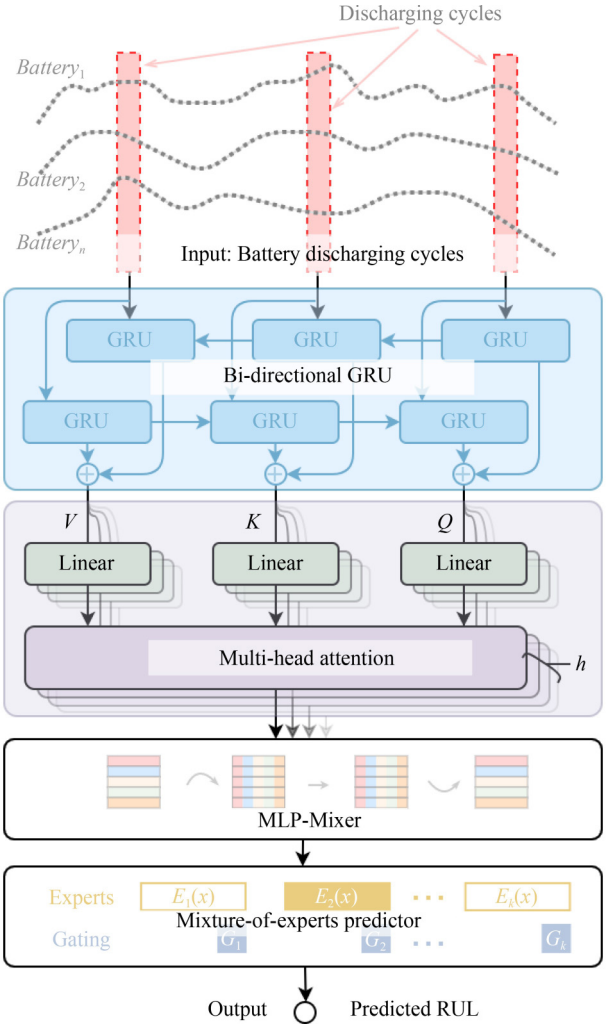


Fig. 1 Illustration of our proposed MMMe

3.1 Bi-GRU-MHA

A Bi-directional Gated Recurrent Unit with Multi-Head Attention (BiGRU-MHA) is employed to effectively extract features from battery sequence data. The BiGRU-MHA model consists of two layers of Bi-directional Gated Recurrent Units (BiGRU) and a multi-head attention mechanism. Given an M-dimensional battery sequence data $\{x_1^M, x_2^M, \ldots, x_t^M, \ldots, x_T^M\}$, the GRU takes the input vector x_t and the previous hidden state h_{t-1} to produce the current hidden state h_t. The following equations demonstrate the detailed process in the GRU.

$$\begin{cases} r_t = \sigma(W_r x_t + U_r h_{t-1}), \\ z_t = \sigma(W_z x_t + U_z h_{t-1}), \\ \tilde{h}_t = \tanh\left(W_{\tilde{h}} x_t + U_{\tilde{h}}(r_t \odot h_{t-1})\right), \\ h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, \end{cases} \tag{1}$$

where the r_t and z_t represent the output of the update gate and reset gate respectively. σ and tanh denotes the activation functions. BiGRU has the advantage of taking into account both past and future information by propagating in both directions, thus providing a better performance than the traditional GRU. The output of BiGRU is denoted as $O_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right]$, which $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ represent the hidden state from forward and backward directions.

It has been shown that not all features have an equal impact on the prediction of RUL [34]. To ensure the effectiveness and distinguishability of input features in the prediction model, Multi-Head Attention was employed to assign different weights to different features. To clarify, the scaled-dot product attention, as presented in Vaswani et al. [35], is defined as in the following manner:

$$\text{Attention}(Q, K, V) = \underbrace{\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)}_{attention\,matrix} V. \tag{2}$$

The scaled dot-product attention mechanism is capable of modeling the inter-dependencies within input battery data sequences. This attention mechanism maps a set of queries, Q, a set of keys, K, and a set of values, V, into an output vector.

In multi-head attention, the input features are projected into different subspaces to obtain different queries, keys, and values in parallel h times. The scaled-dot product attention is then applied to each of the obtained queries, keys, and values.

$$\begin{aligned} &\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O \\ &\text{where} \quad \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \end{aligned} \tag{3}$$

3.2 Rezero MLP-Mixer module

The MLP-Mixer was initially developed for image classification tasks but has since found applications in natural language processing and speech recognition. In this study, a novel variant of the MLP-Mixer is proposed for RUL prediction tasks, called the Rezero MLP-Mixer. This variant incorporates residual connections and zero initialization to enhance the model's performance.

Similar to the original MLP-Mixer architecture, the Rezero MLP-Mixer comprises two MLP blocks: the Channel-mixing MLP and the Temporal-mixing MLP, as shown in Fig. 2. The Channel-mixing MLP enables the communication between different feature channels within each time point, while the Temporal-mixing MLP operates on the time dimension. Suppose that $X \in \mathcal{R}^{T \times D}$ represents the learned features obtained by BiGRU-MHA. The Rezero MLP-Mixer can be formulated as follows:

$$Z = X + \alpha_1 W_2 \Phi(W_1 \text{Norm}(X)), \tag{4}$$

$$Y = Z + \alpha_2 W_4 \Phi(W_3 \text{Norm}(Z)). \tag{5}$$

Norm and Φ represent the LayerNorm and activate function. W_1, W_2, W_3 and W_4 denote the weights of the MLP. α_1 and α_2 are the learnable residual weight that rescales the contribution of the MLP layer with respect to the input.

3.3 MoE predictor

The MoE predictor is comprised of two main components: experts and gating. In this proposed approach, a single MLP is not relied upon to predict RUL. Instead, multiple experts, typically small MLPs, are used where each expert is trained to predict the RUL value using different subsets of the input features. The RUL predictions from the different experts are then combined using a gating mechanism that adjusts the contribution of each expert. The resulting formulation for the

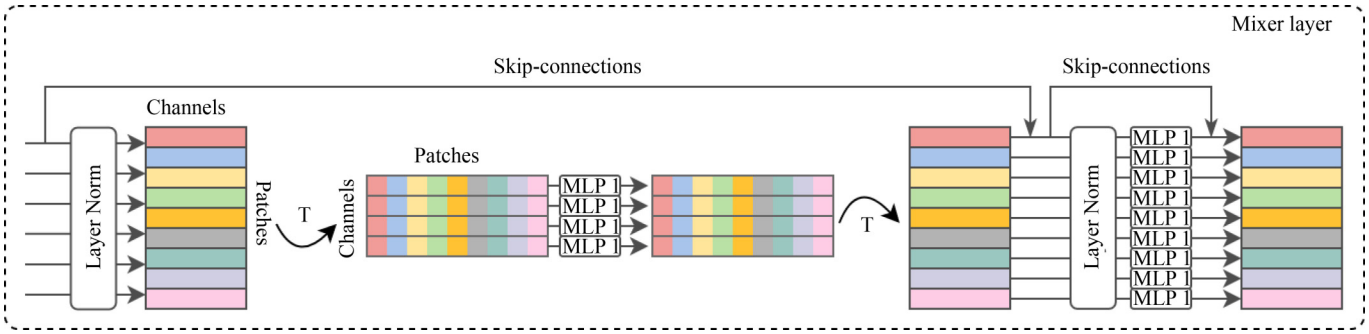


Fig. 2 An illustration of the MLP-mixer

MoE predictor is presented in Eq. (6), and a visual representation of the architecture is provided in Fig. 3. This approach offers several advantages, such as reducing the risk of overfitting and improving the generalization ability of the model. Moreover, it allows for a more interpretable model, as the contribution of each expert can be analyzed to gain insight into the importance of different input features in predicting RUL.

$$\hat{y} = \sum_{i=1}^{K} G_i(x) E_i(x), \tag{6}$$

where K is the number of experts; $G_i(\cdot): \mathbb{R}^d \mapsto \mathbb{R}$ is the gating function and $\sum_{i=1}^{K} G_i(x) = 1, 0 \leqslant g_m(x) \leqslant 1$. $E_i(\cdot): \mathbb{R}^d \mapsto \mathbb{R}^d$ is the i^{th} expert in learning RUL value based on the input features. In this study, each expert is a two-layer MLP.

3.4 Learning

As the RUL prediction is treated as a regression task, the Mean Squared Error (MSE) loss function is employed to guide the model training. The MSE loss function is defined as follows:

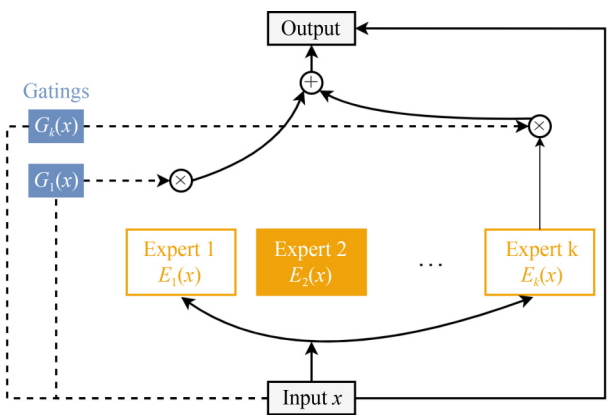


Fig. 3 Visualization of MoE predictor

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (P_i - T_i)^2, \tag{7}$$

where P is the prediction RUL value, and T corresponds to the true value. n indicates the number of samples.

4 Experiments and results

4.1 Experimental data

This paper presents the results of experiments conducted on two publicly available datasets: NASA and CALCE. The NASA dataset, which is available from the NASA Ames Research Center website, uses 18650 model lithium batteries as the subject of study, with accelerated aging tests used to collect data. The researchers controlled the conditions of temperature, discharge cut-off voltage, and discharge current, and collected records from four different groups of lithium batteries (B0005, B0006, B0007, and B0018), each of which was subjected to three repetitive operations of charging, discharging, and impedance measurements. The detailed specifications of the selected NASA lithium-ion batteries are shown in Table 1. Similarly, the CALCE dataset is available from the Center for Advanced Life Cycle Engineering (CALCE) at the University of Maryland, with CS2_33, CS2_34, CS2_35, CS2_36, CS2_37, and CS2_38 being the selected batteries. The detailed specifications of the selected CALCE Li-ion batteries are shown in Table 2. In this study, the EOL for each battery was set to 70% of its rated capacity, which is 1.40 Ah and 0.77 Ah for the NASA and CALCE datasets, respectively. At the same time, the leave-one-out method was used to partition the NASA dataset and the CALCE dataset: one battery was randomly selected, and the rest was used for training. Figures 4 and 5 present the capacity degradation curves of the NASA and CALCE datasets, respectively.

4.2 Baseline methods

The MMMe method used in this article was compared with the following deep models:

- **MLP** [36]: This approach is based on MLP to predict

Table 1 Detailed information of NASA dataset

Battery	Discharge current (A)	Rated capacity (Ah)	Charing/discharge cut-off voltage (V)	Minimal charge current (mA)	Failure threshold (Ah)
B0005	2	2	4.2/2.7	20	1.4
B0006	2	2	4.2/2.5	20	1.4
B0007	2	2	4.2/2.3	20	1.4
B0018	2	2	4.2/2.5	20	1.4

Table 2 Detailed information of CALCE dataset

Battery	Discharge current (A)	Rated capacity (Ah)	Charing/Discharge cut-off voltage (V)	Minimal charge current (mA)	Failure threshold (Ah)
CS2_35	1.1	1.1	4.2/2.7	50	0.77
CS2_36	1.1	1.1	4.2/2.7	50	0.77
CS2_37	1.1	1.1	4.2/2.7	50	0.77
CS2_38	1.1	1.1	4.2/2.7	50	0.77

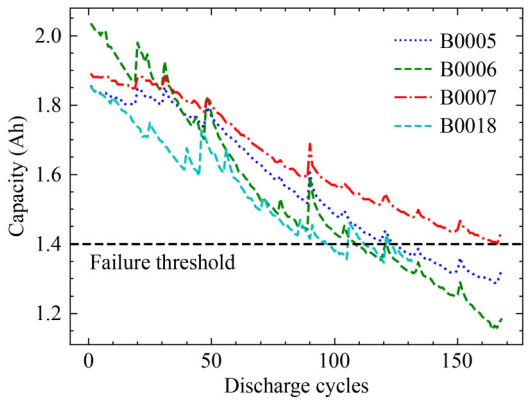


Fig. 4 NASA batteries capacity decay data

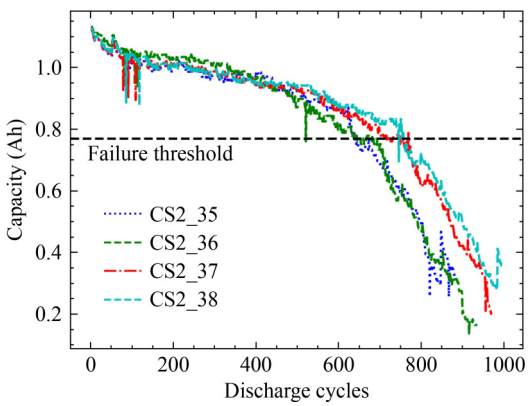


Fig. 5 CALCE batteries capacity decay data

the RUL.

- **RNN** [37]: This baseline constructs the adaptive/ recurrent neural network model to predict the RUL.
- **LSTM** [19]: This method trains the LSTM with the resilient mean square back-propagation method to predict the RUL task.
- **GRU** [38]: This baseline is originally designed for the SoC prediction task with the GRU model. Chen et.al [21] change it to an RUL prediction task.
- **Dual-LSTM** [39]: The Dual-LSTM connects the change point detection and RUL prediction with a newly proposed HI construction function [39].
- **DeTransformer** [21]: DeTransformer employs a Denoising Auto-Encoder to denoise the raw data and a Transformer to learn the feature for capacity fading.

4.3 Evaluation metrics

In this paper, to thoroughly assess the effectiveness of the selected model, three indicators have been chosen to evaluate the performance of RUL prediction. The three evaluation indicators are Relative Error (RE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics are calculated as follows:

$$\mathrm{RE} = \frac{\left|RUL^{\mathrm{pred}} - RUL^{\mathrm{true}}\right|}{RUL^{\mathrm{true}}}, \tag{8}$$

$$\mathrm{RMSE} == \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \widehat{x_t})^2}, \tag{9}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left\|x_t - \widehat{x_t}\right\|. \tag{10}$$

In these equations, n denotes the length of battery degradation trajectory; x_t and $\hat{x}_t$ are the corresponding capacity measured and prediction values at the tth cycle, respectively. RUL^{pred} and RUL^{true} indicate the predicted and true value of RUL.

4.4 Implementation details

The MMMe model proposed in this study was implemented using PyTorch 1.11.0, with Adam utilized as the optimizer and a learning rate of 1e–2 [21]. To prevent overfitting, the MMMe model was trained for a maximum of 1000 epochs and early stopping technology was employed.The hyper-parameters of the proposed model are presented in Table 3. All experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU, while the CPU model used was the Intel(R) Xeon(R) Gold 6226R CPU @ 2.90 GHz.

4.5 Comparison with the state-of-the-art approaches

Table 4 summarizes the comparison results of the proposed MMMe model with the aforementioned methods on NASA and CALCE datasets. It should be noted that we reported the published results of the compared methods as we used the same datasets and experimental settings. The best results are shown in bold. It can be observed that MMMe outperforms the other methods in all three metrics. For the NASA dataset, MMMe achieves RE, MAE, and RMSE of 0.005, 0.04, and 0.0515, respectively, which significantly outperforms the comparison methods under the same conditions. For the CALCE dataset, the proposed method reduces RE, MAE, and RMSE scores by at least 97.5%, 60.6%, and 56.6% compared

Table 3 Parameters used in MMMe

Parameters	Settings
No. of GRU layer	2
Hidden dim of GRU	16
No. of the head of MHA	2
No. of Experts	32
No. of MLP-Mixer layer	2
Learning rate	1e–2
Early stopping patience	200
Max No. of training epochs	1000

Table 4 Performances of deep learning models in NASA and CALCE datasets

Methods	NASA			CALCE		
	RE	MAE	RMSE	RE	MAE	RMSE
MLP	0.3851	0.1379	0.1541	0.4018	0.1557	0.2038
RNN	0.2851	0.0749	0.0848	0.1614	0.0938	0.1099
LSTM	0.2648	0.0829	0.0905	0.0902	0.0582	0.0736
GRU	0.3044	0.0806	0.0921	0.1319	0.0671	0.0946
Dual-LSTM	0.2557	0.0815	0.0879	0.0885	0.0636	0.0874
DeTransformer	0.2252	0.0713	0.0802	0.0764	0.0613	0.0705
MMMe	**0.005**	**0.04**	**0.0515**	**0.0019**	**0.0229**	**0.0306**

to the second-best method. These results demonstrate that the MMMe model provides a more accurate and robust RUL prediction for lithium-ion batteries.

Figure 6 presents the RUL prediction results of MMMe and DeTransformer on the NASA dataset, with the starting point of prediction at 17. The results in Fig. 6 demonstrate that the proposed MMMe's prediction capacity degradation curve is closer to the real one for each battery. Specifically, Figs. 6(a) and 6(b) show the predicted capacity degradation curve and the true curve for batteries B0005 and B0006, respectively. It is evident that as the cycle increases, the predicted curve of DeTransformer moves away from the real curve, whereas our proposed method demonstrates better consistency between the estimates and the actual capacity. This consistency leads to accurate predictions of the RUL for each cycle. However, for battery B0007, DeTransformer shows a complete divergence phenomenon, possibly due to its different deterioration characteristics from other batteries [40]. Nonetheless, the proposed MMMe still tracks capacity degradation well. Battery B0018 exhibits significant local regeneration during capacity degradation, bringing challenges to RUL prediction. In this case, MMMe still achieves higher accuracy than DeTransformer, as shown in Fig. 6(d).

Figure 7 compares the predicted capacity degradation curve of MMMe and DeTransformer for CALCE batteries with the starting point of prediction at 65. The predictive results on the CALCE dataset demonstrate that, apart from a few isolated instances, the proposed model provides a predicted curve that closely approximates the ground truth curve. Our method demonstrates a remarkable capacity for accurately tracking the degradation trend of battery capacity and for providing highly precise RUL predictions. Conversely, the capacity curve

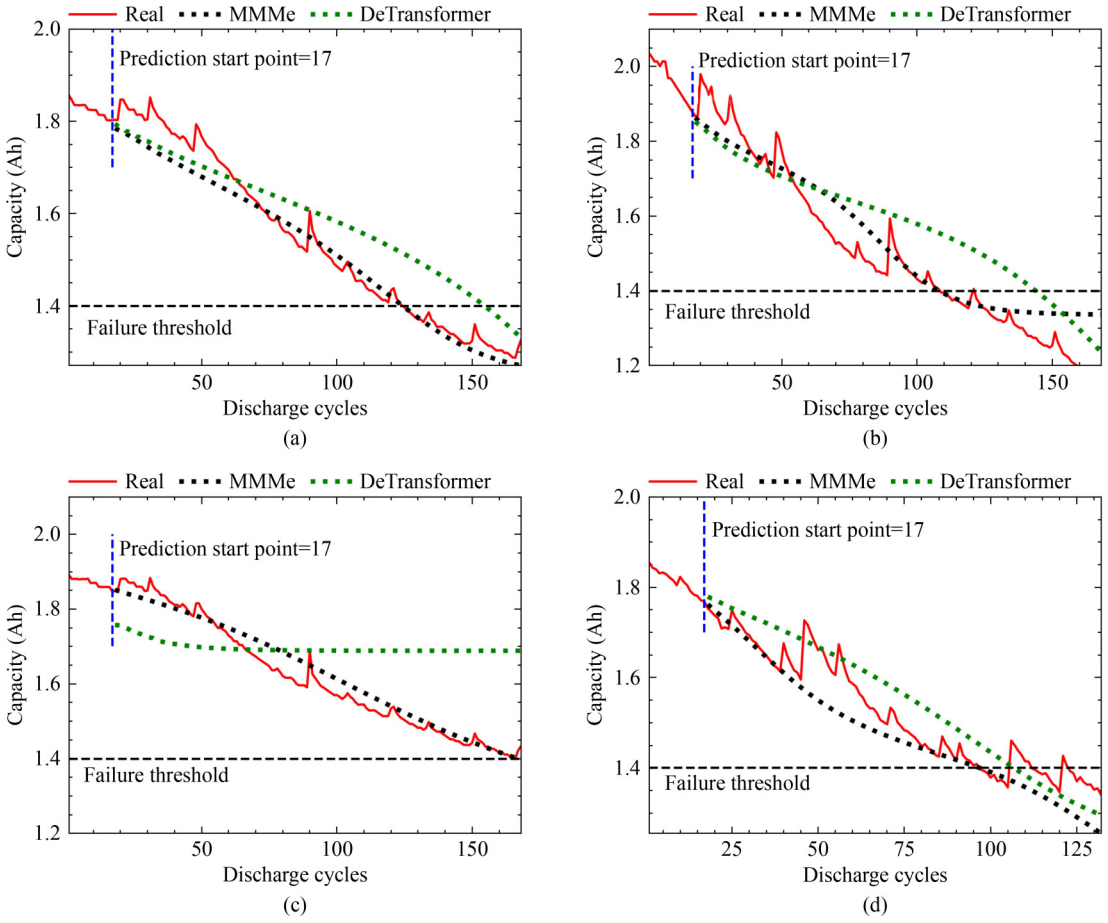


Fig. 6 RUL prediction results of the proposed method for NASA dataset. (a) Battery B0005; (b) Battery B0006; (c) Battery B0007; (d) Battery B0018

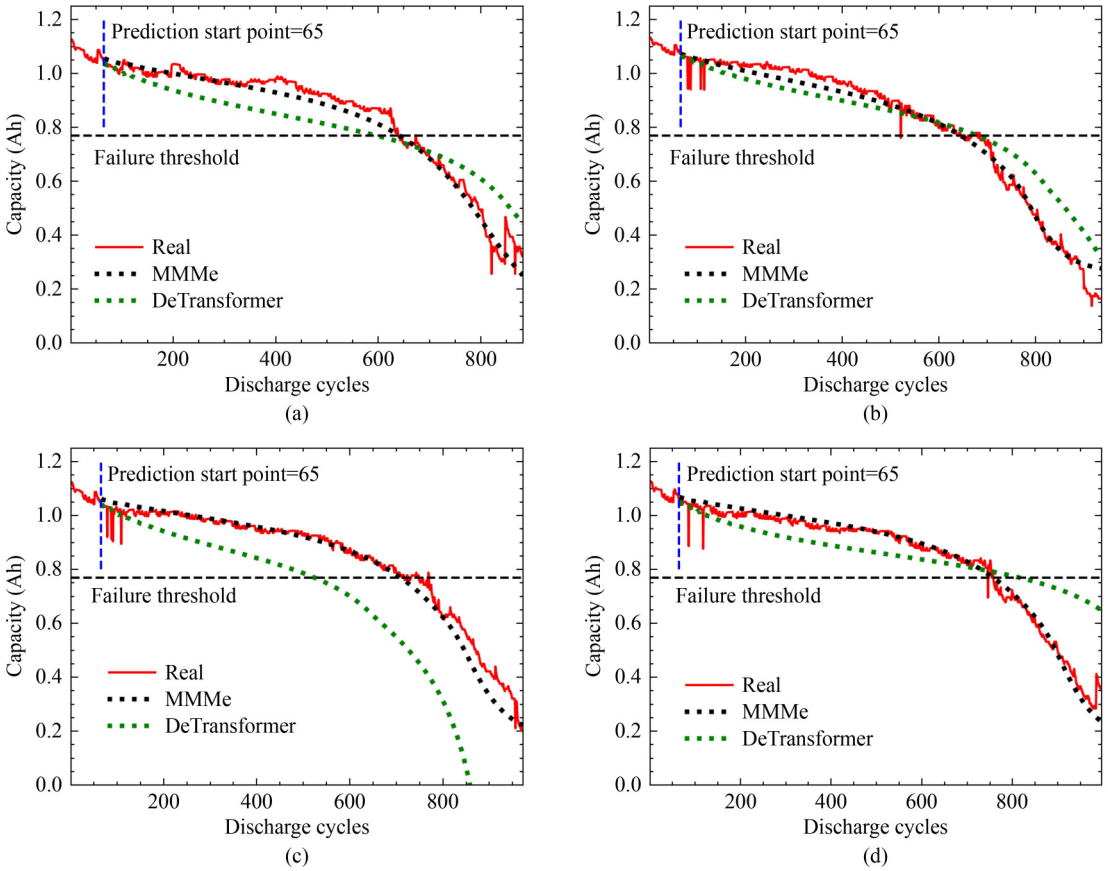


Fig. 7 RUL prediction results of the proposed method for CALCE dataset. (a) Battery CS2_35; (b) Battery CS2_36; (c) Battery CS2_37; (d) Battery CS2_38

tracking of the DeTransformer exhibits considerable deviation from the ground truth curve, especially for the battery CS2_37 (as shown in Fig. 7(c)), and its RUL prediction errors are likewise greater. This suggests that our deep learning model exhibits superior feature extraction capabilities for capacity time series compared to the comparative method. Furthermore, upon comparison with the predictions on the NASA dataset, it is evident that the MMMe model offers superior accuracy in tracking capacity degradation. This may be attributed to the fact that the batteries from the CALCE dataset seldom undergo regeneration, thereby reducing the prediction challenge.

4.6 Effect of number of experts

In this study, experiments were conducted on the NASA dataset to examine the impact of the number of experts in the Mixture-of-Experts (MoE) layer. Different models with varying numbers of experts were compared, and the average scores of RE, MAE, and RMSE were evaluated, as shown in Fig. 8. It is evident that the scores generally increase and then decrease as the number of experts increases. This pattern may be attributed to the fact that increasing the number of experts effectively increases the model capacity. However, when the number of experts becomes too large, the model parameters also increase, resulting in over-fitting.

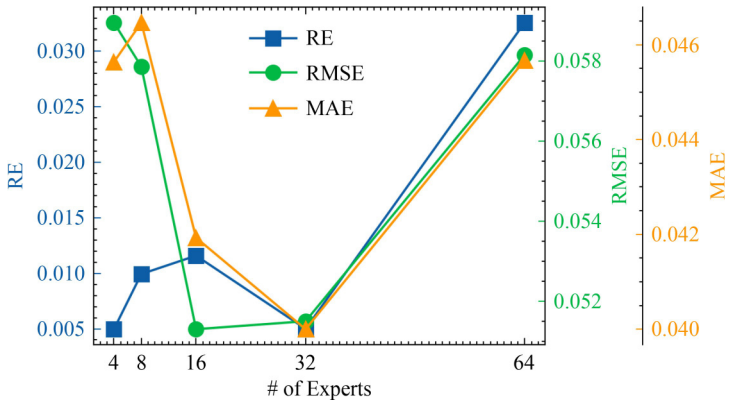


Fig. 8 The RE, RMSE, and MAE of MMMe with a different number of experts

Table 5 Results of ablation experiments

Methods	NASA			CALCE		
	RE	MAE	RMSE	RE	MAE	RMSE
BiGRU	0.3044	0.0806	0.0921	0.1319	0.0671	0.0946
BiGRU+MoE	0.021	0.065	0.081	0.111	0.068	0.095
BiGRU+Mixer-MLP+MoE	0.076	0.077	0.102	0.008	0.053	0.066
BiGRU+MHA+MoE	0.031	0.058	0.084	0.009	0.023	0.039
MMMe	**0.005**	**0.04**	**0.0515**	**0.0019**	**0.0229**	**0.0306**

4.7 Ablation experiments

In this ablation experiment, the performance of a combined model architecture that integrates the Gated Recurrent Unit (GRU), Multi-Head Attention, MLP-Mixer, and Mixture of Experts (MoE) predictor models are investigated. The objective is to evaluate the impact of each component on Lithium-ion battery RUL prediction tasks and assess the overall performance gain achieved through their combination. The evaluation will be based on the indicators of Relative Error (RE), Mean Error (ME), and Root Mean Squared Error (RMSE).

In order to prove the effectiveness of the strategy used in the algorithm of this paper, five sets of ablation experiments are conducted in this paper based on the proposed model, including 1) BiGRU model, 2) BiGRU+MoE model, 3) BiGRU+Mixer-MLP+MoE model, 4) BiGRU+MHA+MoE model, 5) BiGRU+MHA+Mixer-MLP+MoE model (this algorithm). The results of the ablation experiments are shown in Table 5.

The analysis of Table 5 reveals the better predictive capabilities of the proposed MMMe model, thereby validating the substantial contributions made by its constituent components: MHA, MoE, Mixer-MLP, and BiGRU. In addition, the comparison between the "BiGRU" and "BiGRU+MoE" configurations demonstrates significant improvements in RE, MAE, and RMSE across both datasets. Particularly noteworthy is the 0.28 enhancement in RE observed on the NASA dataset. Building upon the foundation of "BiGRU+MoE", the integration of Multi-Head Attention and MLP-Mixer models in "BiGRU+Mixer-MLP+MoE" and "BiGRU+MHA+MoE" respectively exhibits substantial improvements in terms of RE, MAE, and RMSE on the CALCE dataset. These findings affirm the effectiveness of the combined architecture within the MMMe model.

5 Conclusion

This paper proposes a novel deep-learning model for predicting the Remaining Useful Life (RUL) of Lithium-Ion Batteries (LIBs). The proposed method first constructs a time series matrix to preserve temporal information and projects the original input into high-dimensional space using the Bi-directional Gated Recurrent Unit with Multi-Head Attention (BiGRU-MHA) encoder. To learn abstract features for capacity fading, the ReZero MLP-Mixer is employed. Finally, the Mixture-of-Experts (MoE) mechanism is utilized to predict the RUL based on the learned features. Extensive experiments conducted on two publicly available LIB datasets demonstrate that the proposed MMMe method outperforms all baseline methods for RUL prediction.

Acknowledgements We are very grateful to the anonymous reviewers for their effort in evaluating our paper. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62102191, 61872114, and 61871020).

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

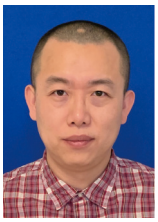
1. Tang X, Liu K, Li K, Widanage W D, Kendrick E, Gao F. Recovering large-scale battery aging dataset with machine learning. Patterns, 2021, 2(8): 100302
2. Wang Z, Liu N, Guo Y. Adaptive sliding window LSTM NN based RUL prediction for lithium-ion batteries integrating LTSA feature reconstruction. Neurocomputing, 2021, 466: 178–189
3. Ge M F, Liu Y, Jiang X, Liu J. A review on state of health estimations and remaining useful life prognostics of lithium-ion batteries. Measurement, 2021, 174: 109057
4. Rauf H, Khalid M, Arshad N. Machine learning in state of health and remaining useful life estimation: theoretical and technological development in battery degradation modelling. Renewable and Sustainable Energy Reviews, 2022, 156: 111903
5. Zhai Q, Ye Z S. RUL prediction of deteriorating products using an adaptive wiener process model. IEEE Transactions on Industrial Informatics, 2017, 13(6): 2911–2921
6. Wang Y, Tian J, Sun Z, Wang L, Xu R, Li M, Chen Z. A comprehensive review of battery modeling and state estimation approaches for advanced battery management systems. Renewable and Sustainable Energy Reviews, 2020, 131: 110015
7. Deng Z, Yang L, Deng H, Cai Y, Li D. Polynomial approximation pseudo-two-dimensional battery model for online application in embedded battery management system. Energy, 2018, 142: 838–850
8. Yang L, Cai Y, Yang Y, Deng Z. Supervisory long-term prediction of state of available power for lithium-ion batteries in electric vehicles. Applied Energy, 2020, 257: 114006
9. Son J, Zhou S, Sankavaram C, Du X, Zhang Y. Remaining useful life prediction based on noisy condition monitoring signals using constrained Kalman filter. Reliability Engineering & System Safety, 2016, 152: 38–50
10. Su X, Wang S, Pecht M, Zhao L, Ye Z. Interacting multiple model

particle filter for prognostics of lithium-ion batteries. Microelectronics Reliability, 2017, 70: 59–69
11. Tian J, Xu R, Wang Y, Chen Z. Capacity attenuation mechanism modeling and health assessment of lithium-ion batteries. Energy, 2021, 221: 119682
12. Li Y, Liu K, Foley A M, Zülke A, Berecibar M, Nanini-Maury E, Van Mierlo J, Hoster H E. Data-driven health estimation and lifetime prediction of lithium-ion batteries: a review. Renewable and Sustainable Energy Reviews, 2019, 113: 109254
13. Li S, Fang H, Shi B. Remaining useful life estimation of lithium-ion battery based on interacting multiple model particle filter and support vector regression. Reliability Engineering & System Safety, 2021, 210: 107542
14. Shu X, Li G, Shen J, Lei Z, Chen Z, Liu Y. A uniform estimation framework for state of health of lithium-ion batteries considering feature extraction and parameters optimization. Energy, 2020, 204: 117957
15. Liu Z, Cheng Y, Wang P, Yu Y, Long Y. A method for remaining useful life prediction of crystal oscillators using the Bayesian approach and extreme learning machine under uncertainty. Neurocomputing, 2018, 305: 27–38
16. Shen D, Wu L, Kang G, Guan Y, Peng Z. A novel online method for predicting the remaining useful life of lithium-ion batteries considering random variable discharge current. Energy, 2021, 218: 119490
17. Yang B, Liu R, Zio E. Remaining useful life prediction based on a double-convolutional neural network architecture. IEEE Transactions on Industrial Electronics, 2019, 66(12): 9521–9530
18. Ding P, Liu X, Li H, Huang Z, Zhang K, Shao L, Abedinia O. Useful life prediction based on wavelet packet decomposition and two-dimensional convolutional neural network for lithium-ion batteries. Renewable and Sustainable Energy Reviews, 2021, 148: 111287
19. Zhang Y, Xiong R, He H, Pecht M G. Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries. IEEE Transactions on Vehicular Technology, 2018, 67(7): 5695–5705
20. Zhao S, Zhang C, Wang Y. Lithium-ion battery capacity and remaining useful life prediction using board learning system and long short-term memory neural network. Journal of Energy Storage, 2022, 52: 104901
21. Chen D, Hong W, Zhou X. Transformer network for remaining useful life prediction of lithium-ion batteries. IEEE Access, 2022, 10: 19621–19628
22. Zheng L, He Y, Chen X, Pu X. Optimization of dilated convolution networks with application in remaining useful life prediction of induction motors. Measurement, 2022, 200: 111588
23. Ragab M, Chen Z, Wu M, Kwoh C K, Yan R, Li X. Attention-based sequence to sequence model for machine remaining useful life prediction. Neurocomputing, 2021, 466: 58–68
24. Wu J Y, Wu M, Chen Z, Li X L, Yan R. Degradation-aware remaining useful life prediction with LSTM autoencoder. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1–10
25. Jin R, Chen Z, Wu K, Wu M, Li X, Yan R. Bi-LSTM-based two-stream network for machine remaining useful life prediction. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1–10
26. Xiao L, Zhang L, Niu F, Su X, Song W. RETRACTED: remaining useful life prediction of wind turbine generator based on 1D-CNN and Bi-LSTM. International Journal of Fatigue, 2022, 163: 107051
27. Rouhi Ardeshiri R, Ma C. Multivariate gated recurrent unit for battery remaining useful life prediction: a deep learning approach. International Journal of Energy Research, 2021, 45(11): 16633–16648
28. Chen Z, Wu M, Zhao R, Guretno F, Yan R, Li X. Machine remaining useful life prediction via an attention-based deep learning approach. IEEE Transactions on Industrial Electronics, 2021, 68(3): 2521–2531
29. Tolstikhin I O, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J, Lucic M, Dosovitskiy A. MLP-Mixer: an all-MLP architecture for vision. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. 2021
30. Touvron H, Bojanowski P, Caron M, Cord M, El-Nouby A, Grave E, Izacard G, Joulin A, Synnaeve G, Verbeek J, Jegou H. ResMLP: feedforward networks for image classification with data-efficient training. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 5314–5321
31. Chen S, Xie E, Ge C, Chen R, Liang D, Luo P. CycleMLP: a MLP-like architecture for dense prediction. In: Proceedings of the 10th International Conference on Learning Representations. 2022
32. Yu T, Li X, Cai Y, Sun M, Li P. S^2-MLP: spatial-shift MLP architecture for vision. In: Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2022, 3615−3624
33. Jacobs R A, Jordan M I, Nowlan S J, Hinton G E. Adaptive mixtures of local experts. Neural Computation, 1991, 3(1): 79–87
34. Liu H, Liu Z, Jia W, Lin X. Remaining useful life prediction using a novel feature-attention-based end-to-end approach. IEEE Transactions on Industrial Informatics, 2021, 17(2): 1197–1207
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 6000−6010
36. Wu Y, Li W, Wang Y, Zhang K. Remaining useful life prediction of lithium-ion batteries using neural network and bat-based particle filter. IEEE Access, 2019, 7: 54843–54854
37. Liu J, Saxena A, Goebel K, Saha B, Wang W. An adaptive recurrent neural network for remaining useful life prediction of lithium-ion batteries. In: Proceedings of Annual Conference of the Prognostics and Health Management Society. 2010
38. Williard N, He W, Osterman M, Pecht M. Comparative analysis of features for determining state of health in lithium-ion batteries. International Journal of Prognostics and Health Management, 2013, 4(1): 1−7
39. Shi Z, Chehade A. A dual-LSTM framework combining change point detection and remaining useful life prediction. Reliability Engineering & System Safety, 2021, 205: 107257
40. Nagulapati V M, Lee H, Jung D, Brigljevic B, Choi Y, Lim H. Capacity estimation of batteries: influence of training dataset size and diversity on data driven prognostic models. Reliability Engineering & System Safety, 2021, 216: 108048

Lingling Zhao is an associate professor at Faculty of Computing, Harbin Institute of Technology, China. She received the PhD, MS, and BS degrees from Harbin Institute of Technology, China. Her current research interests include machine learning.

Shitao Song is currently a postgraduate student in the School of Electrical Engineering, Liaoning University of Technology, China. His research interest includes machine learning and micro-grid scheduling optimization.

Pengyan Wang received his BE degree in Computer Science and Technology from Northeast Electric Power University, China in 2021. Currently, he is pursuing a ME degree in Computer Science and Technology at Northeast Electric Power University, China. His research interest includes lithium-ion battery safety.



Chunyu Wang is a professor at the Faculty of Computing, Harbin Institute of Technology, China. He received his BS, MS, and PhD degrees in computer science and technology from Harbin Institute of Technology, China. His current research interests include bioinformatics and machine learning.

Junjie Wang received the BS degree in Information management and information system from Institute of Disaster Prevention in 2013, the MS degree in software engineering in 2015, and the PhD degree in Computer science and technology in 2020 from the Harbin Institute of Technology, China. He is a Lecturer with the School of Biomedical Engineering and Informatics, Nanjing Medical University, China. His current research interests include bioinformatics and deep learning.

Maozu Guo received the PhD degree from the School of Computer Science and Technology, Harbin Institute of Technology, China. He is currently a Professor with the School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, China. His current research interests include machine learning and artificial intelligence.