International Journal of Engineering and Technology Innovation, vol.16, no. 1, 2026, pp. 134-162

A Parallel Prediction Method for Battery Capacity Based on the Multiscale and Temporal-Spatial Feature Fusion

Wanzhen Wang[1,2,*], Ngu Sze Song[2], Qian Wang[1], Man Qiu[1], Xiaomei Ni[1,2], Miaomiao Xin[1,2], Xiao Wang[1], Na Li[1], Shengze Ma[1]

[1]School of Intelligent Manufacturing and Control Engineering, Qilu Institute of Technology, Jinan, China

[2]Faculty o Engineering, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia

Received 05 December 2025; received in revised form 22 January 2026; accepted 23 January 2026

DOI: https://doi.org/10.46604/ijeti.2026.15949

Abstract

This study proposes a parallel feature fusion-based method to accurately predict battery capacity for battery health management (BHM). Existing data-driven approaches suffer from ineffective extraction of multi-scale features and feature redundancy induced by sequential strategies. To address these gaps, a hybrid deep learning framework is proposed. Specifically, discharge voltage data are first standardized to unify sample dimensions. Then, parallel multi-scale branches are constructed to simultaneously capture the spatial and temporal features of battery discharge signals. A channel attention module is subsequently employed to adaptively filter redundant features and enhance the weight of degradation-related feature representations. Finally, a fully connected network maps the refined features to battery capacity values. The experimental results validate that the proposed method outperforms single-model and sequential baselines, with 0.00369 to 0.01331 RMSE on NASA battery, 0.0498 on CALCE batteries, and 1.8196 on Oxford batteries.

Keywords: battery, multi-scale, spatial features, temporal features, channel attention

1. Introduction

Energy storage is a vital intermediary for a variety of new energy power generation sources due to the necessity for stable power networks. Individual energy storage batteries are vulnerable to variations in charge and discharge loads under the supervision of battery management systems (BMS), which lead to degradation. For later scientific management and maintenance of these batteries, accurate capacity prediction is crucial. Proactive maintenance and a 20% increase in battery life have been made possible by sophisticated battery management systems that use machine learning algorithms to detect capacity deterioration patterns [1]. Similarly, intelligent algorithms built into battery systems may evaluate current variables like temperature and voltage to precisely estimate remaining capacity. By modifying charging and discharging plans, these algorithms can increase battery cycle life by more than 20% [2]. The development of reliable and accurate battery capacity prediction techniques is crucial for advancing energy storage technologies and their practical applications.

The traditional physics-based approach, the data-driven approach, and the physics-constrained hybrid data-driven approach are the three steps of battery capacity prediction. Some researchers employed statistical techniques, such as Gaussian process regression (GPR) [3], support vector regression (SVR) [4], and random forest (RF) [5]. By examining exterior battery characteristics like voltage, current, and temperature during cycles, they forecasted capacity deterioration. Nevertheless, these approaches struggle to properly handle nonlinear features in battery degradation since they rely on manual feature engineering.

* Corresponding author. E-mail address: 22010329@siswa.unimas.my

To address the insufficient mining of battery degradation information in conventional machine learning approaches, researchers concentrated on deep learning for battery capacity prediction, achieving modest success. Autoregressive models, such as the autoregressive (AR) model [6], the autoregressive moving average (ARMA) model [7], and the autoregressive integrated moving average (ARIMA) model [8], have yielded positive outcomes. However, they also have limitations in terms of linear assumptions, high data requirements, reliance on feature engineering, and limited long-term prediction.

In-depth data mining during battery charge or discharge to forecast capacity and track degradation in real-time is made possible by the advancement of deep learning. Battery capacity was predicted using traditional models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), as well as their variants, including long short-term memory (LSTM), gated recurrent units (GRUs), bidirectional LSTM (BiLSTM), and BiGRU. These techniques can produce end-to-end accurate predictions by automatically learning spatial-temporal dependencies from large amounts of battery cycle data [9].

Pre-trained models such as VGGNet, LeNet, and InceptionNet have been used by researchers to predict battery usage scenarios with specific outcomes. For instance, researchers converted and combined three feature curves into two-dimensional pictures and tested several CNNs, including LeNet, AlexNet, VGG16, ResNet, and TCNN, for battery lifetime prediction. They discovered that merging all curves, applying ReLU activation, and choosing particular voltage segments increased prediction accuracy from the initial voltage data of 100 cycles [10]. Traditional CNN techniques have intrinsic limits when handling the long-sequence time-series prediction problem of battery capacity fading, notwithstanding their outstanding performance in local feature extraction. In particular, it is difficult to accurately describe the cross-cycle dependencies in the battery aging process due to the translation-invariance assumption and fixed receptive field of traditional CNNs. This is particularly true when dealing with capacity regeneration phenomena in challenging operating environments, where CNN prediction ability dramatically declines [11].

To overcome these limitations, researchers have lately created time-series prediction techniques, such as LSTM [12], BiLSTM [13-14], GRU [15], BiGRU [16] architecture, etc. In lithium-ion battery capacity prediction research, single models offer the advantages of low deployment costs, high computational efficiency, and a simple structure, but their feature extraction capabilities are limited. To solve this problem, researchers have suggested hybrid approaches that combine spatiotemporal features. Currently, there are two hybrid strategies: sequential and parallel.

The parallel technique can efficiently combine features from multiple data sources or levels of the same data, thereby improving prediction performance. Researchers proposed a feature-fusion CNN model for estimating battery state of health (SOH) using QV, IC, and ΔQV features derived from partial voltage measurements. The model was evaluated and showed higher performance on three datasets [17]. Researchers proposed a methodology for identifying aging patterns via fast-rate, voltage-based charge/discharge cycles for accurate SOH assessment [18].

In addition to feature-level fusion, several studies combined raw data to obtain additional information on degradation during the charge and discharge processes. For example, researchers merged transformed 2D images from multi-channel signals, including QV, IC, and differential voltage curves, as inputs, and achieved exceptional results [19]. To address the difficulty of predicting lithium-ion battery temperature during the thermal runaway phase, a multimodal data-driven model was proposed that combines ResNet18 and a 1D CNN, resulting in high-precision early battery temperature prediction [20]. Furthermore, attention techniques such as self-attention [21-22] and informer models [23] have been used and have demonstrated the enormous potential of attention mechanisms in battery capacity prediction. Some related data-driven BHM research is summarized in Table 1.

Table 1 Data-driven approach for BHM

Research	Contributions in feature extraction	Merits	Demerits
GPR [3]	Variation coefficient extracted from partial charging voltage/current curves	Efficient and practical features	No multi-scale feature fusion mechanism
SVR [4]	Key health factors are screened by RF	Fast convergence speed	Manual screening
RF [5]	Voltage interval features	Balance accuracy and robustness	Manual screening
ResNet [9]	Feature engineering on imbalanced data	Real-world massive imbalanced data	No multi-scale feature fusion
CNN [10]	Feature extraction using CNN with Graphical feature curves	Use pre-trained CNN models to quickly extract features	Single model
LSTM-RNN [12]	Use raw voltage, current, and temperature as input	Temperature adaptability	No parallel structure
GASF-2DCNN-LSTM [19]	Use a model to extract features from raw multi-channel data	Examine the impact of image size on the outcomes, and use parallel structure	No multi-scale feature fusion

Although deep learning algorithms have made tremendous progress in predicting the lifespan of lithium-ion batteries, some gaps remain. These are primarily reflected in:

(1) The current approaches have restricted feature fusion scales and are unable to extract multi-scale degradation information from charge-discharge data. This results in poor forecast accuracy when addressing capacity regeneration concerns.

(2) In terms of spatiotemporal feature fusion, single models frequently perform poorly. Serial models often extract new information from features collected in the previous module, leading to a loss of some battery deterioration information and a decrease in forecast accuracy.

(3) Traditional feature selection methods necessitate substantial subject knowledge. When dealing with low-interpretable features in multi-scale spatiotemporal fusion, there is often erroneous or redundant information, which reduces the model's generalization capacity.

Notably, these limitations are not merely technical shortcomings but rather intrinsic flaws in the traditional architectural design that cannot be resolved solely by parameter tuning or algorithmic adjustments. Building on this reasoning, the core contributions of this work are outlined as follows:

(1) Data from many charge-discharge processes is used to extract batch features. A multi-scale CNN (MsCNN) model is proposed to extract spatial features from the battery charge-discharge process. This approach effectively extracts multi-scale spatial information from process data and detects battery degradation.

(2) A parallel multi-scale CNN with an LSTM network model (Parallel-MsCNN-LSTM) is proposed to extract spatiotemporal features from battery charge-discharge process data, resulting in rich spatiotemporal feature fusion. This parallel technique efficiently captures the temporal and local variation patterns of the battery charge-discharge process.

(3) A feature adaptive selection technique based on channel attention (AT) is proposed to overcome the feature redundancy problem in multi-scale spatiotemporal feature fusion. This method further screens the model's extracted multi-scale spatiotemporal data, suppressing variables unrelated to battery degradation, resulting in high-precision battery capacity prediction.

The rest of the paper is organized as follows: Section 2 describes the suggested Parallel approach in depth, including the multi-scale structure, LSTM, and channel attention mechanism. Section 3 summarizes the data used and the parameters for the ablation and comparison tests. The subsequent sections examine and discuss the outcomes of each experiment. The final section summarizes the study's work and discusses future research directions.

2. Methodology

This section discusses the proposed prediction architecture and then elaborates on each component, including the Multi-scale CNN, LSTM, and channel attention, to extract more in-depth characteristics relevant to battery degradation. The whole workflow is discussed initially.

2.1. Overall workflow

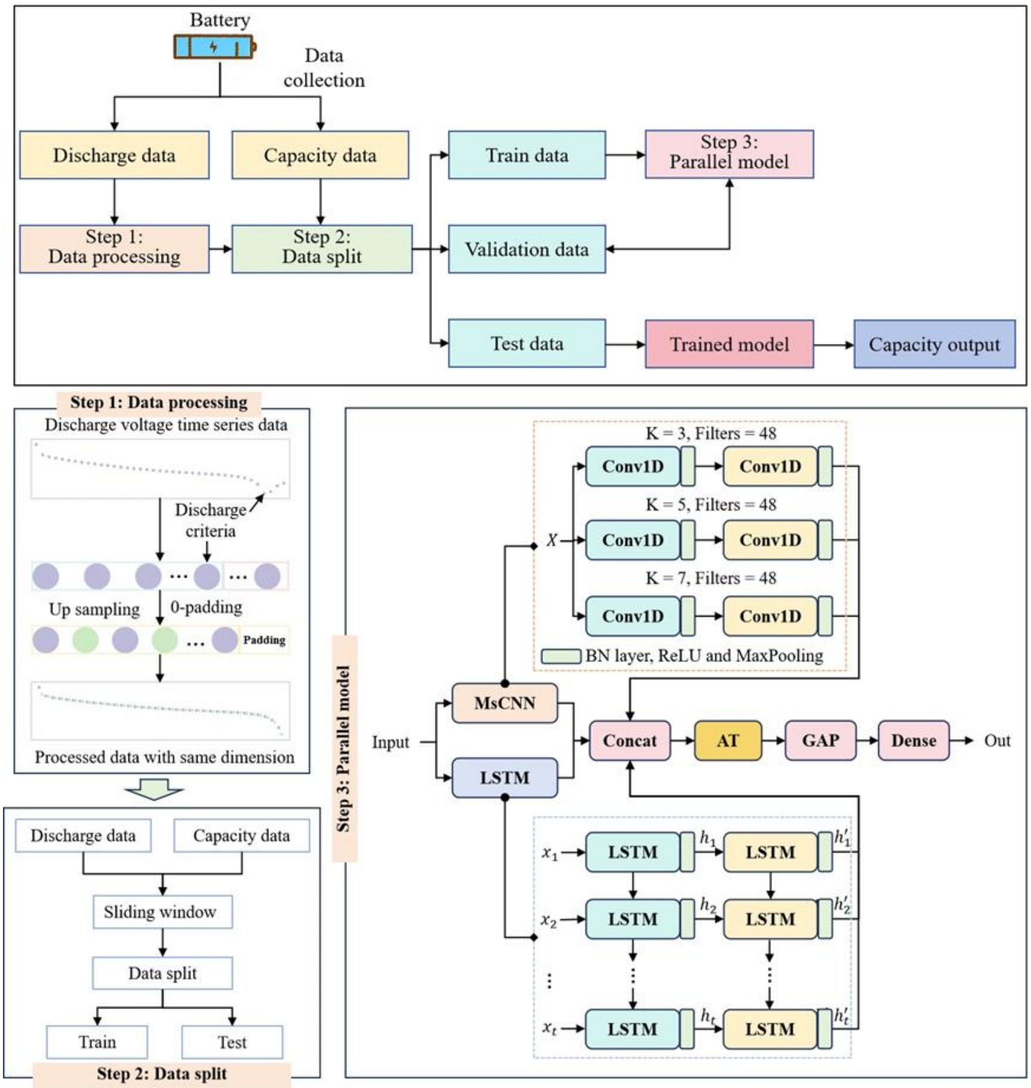


Fig. 1 The overall architecture of the proposed model

This research presents a battery capacity prediction framework based on a parallel multi-scale CNN and LSTM feature-fusion model with channel attention (Parallel-MsCNN-LSTM-AT). The complete workflow of the proposed parallel model for battery capacity prediction is depicted in Fig. 1. First, this article abandons the feature-extraction approaches for the raw charge-discharge process that are commonly utilized in the literature in favor of raw data. During the charge-discharge process of lithium-ion batteries, the discharge and charge voltage curves are the primary indications of battery degradation. At the same time, current and temperature data can serve as additional indicators to comprehensively evaluate battery health. This research uses raw discharge time-series data to estimate battery capacity. Following that, because both CNN and LSTM require consistent input dimensions, oversampling of the battery discharge data is used to address the issue of variable internal sampling rates. As the depth of discharge grows, the time to reach the preset discharge limit decreases, resulting in uneven durations of discharge sequence data. As a result, following oversampling, the shorter sequences are zero-padded.

The Parallel-MsCNN-LSTM-AT consists of three parts: the MsCNN module with multiple kernel sizes, the stack of LSTM layers, and the adaptive feature selection module that employs the channel attention. The capacity values are estimated using global average pooling and a dense layer.

2.2. Parallel MsCNN-LSTM module

The Parallel-MsCNN-LSTM module extracts features from input data by running numerous parallel CNN branches with varying convolutional kernel sizes. This collects properties at multiple temporal and spatial scales during battery depletion, hence increasing feature diversity and expressiveness.

The input data is a time series, and the sliding window technique is used to generate new subsequences. The number of subsequences can be approximated as Eq. (1):

$$C_{\text{in}} = \frac{n-L}{s} + 1 \tag{1}$$

where C_{in}, L and s denote the number of subsequences, the window size, and the stride, respectively. n is the length of the input data. The final feature dimension is converted to $C_{\text{in}} \times L$.

The proposed MsCNN architecture consists of three parallel 1D-CNN branches designed to extract features at different receptive fields. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_t, \ldots \boldsymbol{x}_{C_{in}}\} \in \mathbb{R}^{C_{\text{in}} \times L}$ denote the input feature map. For each branch i ranging from 1 to 3, the transformation output is defined as Eq. (2):

$$\boldsymbol{P}_i = f(\text{Conv1D}(f(\text{Conv1D}(\boldsymbol{X}; k_i, C_{\text{out}})), k_i, C'_{\text{out}})) \tag{2}$$

where Conv1D denotes a one-dimensional convolution with kernel size k and C_{out} or C'_{out} output channels, f represents sequential calculation, including BN, ReLU, and MaxPooling operations. k_i is the kernel size specific to the i-th branch. The final multi-scale representation is obtained by concatenating the three branch outputs along the channel dimension, as shown in Eq. (3):

$$\boldsymbol{Out}_{\text{MsCNN}} = \text{Concate}(\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3) \tag{3}$$

While the multi-scale 1D-CNN effectively captures local morphological patterns in voltage sequences during battery discharge, the gradual capacity degradation is inherently a long-term temporal process characterized by complex nonlinear dynamics. To model such sequential dependencies, a parallel LSTM branch is introduced to operate directly on the same input. Unlike the CNN branches, which focus on fixed-size receptive fields and emphasize spatial locality, the LSTM is designed to retain memory over extended sequences, thereby learning the evolving trends and contextual shifts associated with capacity fade across multiple charge-discharge cycles. Compared with other RNN variants, such as GRUs, the LSTM model is better at capturing the long-term cumulative dependencies inherent in sequences of battery capacity degradation. Its distinct gating structure effectively mitigates the vanishing gradient problem, enabling it to preserve critical long-term degradation information that simplified architectures may overlook. This structural advantage tends to become more pronounced as the size of the training sample increases.

Let the same $\boldsymbol{X}$ denote the input feature map. $\boldsymbol{x}_t$ denote the input feature vector at the *t*-th time step. The LSTM cell at *t*-th time step updates the state using Eq. (4):

$$\left(\boldsymbol{h}_t, \boldsymbol{C}_t\right) = \mathrm{LSTM}\left(\boldsymbol{x}_t, \boldsymbol{h}_{t\text{-}1}, \boldsymbol{C}_{t\text{-}1}\right) \tag{4}$$

where $\boldsymbol{h}_{t\text{-}1}$ denotes the LSTM hidden state at step *t*-1, encoding the cumulative historical degradation trend of the battery. $\boldsymbol{h}_t$ is the hidden state at step *t*. $\boldsymbol{C}_t$ is the cell state, storing the full trajectory of battery degradation from the initial state to step *t*. $\boldsymbol{C}_{t\text{-}1}$ denotes the cell state at the *t*-1 time step. The internal workings of an LSTM unit can be divided into the following steps, from Eq. (5) to Eq. (10):

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_i \bullet \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_i\right) \tag{5}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_f \bullet \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_f\right) \tag{6}$$

$$\tilde{\boldsymbol{C}}_t = \tanh\left(\boldsymbol{W}_c \bullet \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_c\right) \tag{7}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t \bullet \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \bullet \tilde{\boldsymbol{C}}_t \tag{8}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_o \bullet \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_o\right) \tag{9}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \bullet \tanh\left(\boldsymbol{C}_t\right) \tag{10}$$

where $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, $\tilde{\boldsymbol{C}}_t$, and $\boldsymbol{o}_t$ denote the input gate, the forget gate, the candidate cell state, and the output gate, respectively. $\boldsymbol{i}_t$ represents the switch controlling how much new degradation information enters the cell's memory. $\boldsymbol{f}_t$ represents the filter for discarding irrelevant historical degradation memory, such as early-cycle voltage noise, while retaining long-term trends. $\tilde{\boldsymbol{C}}_t$ represents the candidate degradation trend derived from current voltage features and historical memory. $\boldsymbol{o}_t$ represents the selector for extracting relevant memory content to output the current degradation state for capacity prediction. The weight matrices $\boldsymbol{W}_i$, $\boldsymbol{W}_f$, $\boldsymbol{W}_c$, $\boldsymbol{W}_o$ denote the learnable parameters that capture correlations between discharge voltage features and battery capacity fade, while the bias vectors $\boldsymbol{b}_i$, $\boldsymbol{b}_f$, $\boldsymbol{b}_c$, $\boldsymbol{b}_o$ denote the fine-tuning terms for the respective feature mappings. The sigmoid function $\sigma(\bullet)$ represents the activation that constrains gate outputs to 0-1, acting as switches for information flow.

Collectively, these gates operate in a coordinated manner to dynamically regulate the flow of degradation information through the LSTM unit: the forget gate first clears outdated or noisy historical data to ensure the cell state remains focused on meaningful long-term trends; the input gate then selectively integrates new degradation information into the candidate cell state; finally, the output gate extracts the most relevant memory content to produce the current degradation state, which directly supports capacity prediction. This synergistic filtering, updating, and selection process enables the LSTM to maintain a robust, temporally consistent representation of battery degradation, avoiding both information overload and the loss of critical long-term dependencies.

To capture battery degradation trends from discharge voltage features, the LSTM cell updates as follows: The input gate $\boldsymbol{i}_t$ filters useful degradation signals from the current voltage feature vector $\boldsymbol{x}_t$, the forget gate $\boldsymbol{f}_t$ prunes irrelevant historical memory, the candidate cell state $\tilde{\boldsymbol{C}}_t$ generates a new degradation trend candidate, and the cell state $\boldsymbol{C}_t$ fuses retained history and new information to store the full degradation trajectory. Finally, the output gate $\boldsymbol{o}_t$ selects the memory content to produce the current degradation state $\boldsymbol{h}_t$, which directly maps to the battery's remaining capacity or serves as the input to the next LSTM cell. It dynamically learns the non-stationary characteristics of battery degradation and incorporates contextual information to improve predictions at the current time step. After stacking the LSTM layers, the extracted features are retrieved using Eq. (11):

$$\boldsymbol{Out}_{LSTM} = f(LSTM(f(LSTM(\boldsymbol{X})))) \tag{11}$$

where $f(\bullet)$ represents sequential calculation, including BN, ReLU, and MaxPooling operations.

The feature-fusion module combines the extracted features using concatenation and channel attention to provide adaptive fusion of spatial-temporal features. This improves prediction resilience and accuracy by using feature complementarity and information enrichment. It should be noted that the fused features have potential issues: the concatenated features contain multi-source information with varying relevance to battery capacity degradation, and there is redundancy and interference among different feature components. Specifically, the channel dimension of the fused feature tensor corresponds to the feature maps from the two parallel branches. For the MsCNN branch, each channel captures local voltage fluctuation patterns at different time scales. Small-kernel filters generate channels encoding short-term voltage spikes related to instantaneous load changes, with weak correlation to long-term degradation, while large-kernel filters produce channels characterizing long-term voltage plateau variations, which are a direct physical indicator of gradual capacity degradation.

For the LSTM branch, each channel encodes the sequential dependencies of voltage changes across different time windows, thereby characterizing cumulative degradation trends in batteries across consecutive cycles. To address the aforementioned issues of feature redundancy and uneven relevance, the channel attention mechanism is essential. It can adaptively assign weights to different channels, enhancing the contributions of critical degradation-related features, such as long-term voltage plateau variations and cumulative degradation trends, while suppressing redundant noise-induced features, such as short-term random voltage fluctuations. Eq. (12) represents the process of spatiotemporal feature fusion and selection:

$$\text{Capacity} = Dense(GAP(AT(Concate(\boldsymbol{Out}_{\text{MsCNN}}, \boldsymbol{Out}_{\text{LSTM}})))) \tag{12}$$

where $AT(\bullet)$ denotes the channel attention mechanism. The calculation process is shown in Fig. 2. $GAP(\bullet)$, $GMP(\bullet)$ and $Dense(\bullet)$ represent the global average pooling, the global max pooling, and the fully connected layer. Finally, the predicted capacity is calculated through a linear activation function. The detailed calculation process is shown as follows.

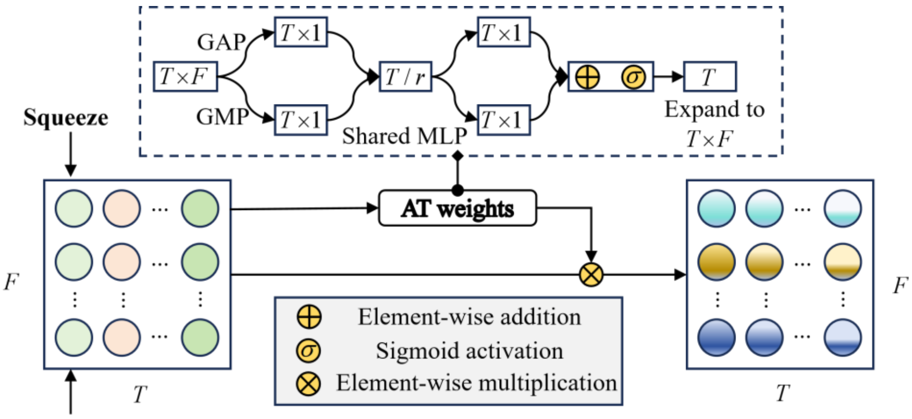


Fig. 2 The structure of the AT module

Let $\boldsymbol{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, ...\boldsymbol{I}_T\} \in \mathbb{R}^{T \times F}$ be the concatenated result of the parallel branches, where T and F are the numbers of temporal and feature dimensions, respectively. Adopting the concept of squeeze-and-excitation (SE) [24], compression is performed along the F direction. Two compression methods, max-pooling and average-pooling, are employed to produce compressed representations at different levels. The two operations at channel t are calculated in Eq. (13) and Eq. (14):

$$\boldsymbol{GAP}_t = \frac{1}{F}\sum_{f=1}^{F} \boldsymbol{I}_{t,f} \in \mathbb{R}^{T \times 1},\ t = 1, 2, ..., T \tag{13}$$

$$\boldsymbol{GMP}_t = \max_{f=1}^{F} \boldsymbol{I}_{t,f} \in \mathbb{R}^{T \times 1},\ t = 1, 2, ..., T \tag{14}$$

where $\boldsymbol{I}_{t,f}$ represents the f-th feature at the t-th channel. The two strategies compress the feature vector at each time step into a 1-dimensional representation.

Subsequently, the two types of feature representations are separately fed into a weight-shared multi-layer perceptron (MLP) to learn channel-wise weights as shown in Eq. (15) and Eq. (16):

$$\boldsymbol{Out}_{GAP} = \sigma\left(\boldsymbol{W}_2 \bullet \sigma\left(\boldsymbol{W}_1 \bullet \boldsymbol{GAP} + \boldsymbol{b}_1\right) + \boldsymbol{b}_2\right), \boldsymbol{W}_1 \in \mathbb{R}^{(T/r) \times T}, \boldsymbol{b}_1 \in \mathbb{R}^{(T/r) \times 1}, \boldsymbol{W}_2 \in \mathbb{R}^{T \times (T/r)}, \boldsymbol{b}_2 \in \mathbb{R}^{T \times 1} \tag{15}$$

$$\boldsymbol{Out}_{GMP} = \sigma\left(\boldsymbol{W}_2 \bullet \sigma\left(\boldsymbol{W}_1 \bullet \boldsymbol{GMP} + \boldsymbol{b}_1\right) + \boldsymbol{b}_2\right) \tag{16}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the weights of the shared parameters of the MLP network. $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are their bias. σ denotes the sigmoid activation function. r is the compression ratio. The number of hidden-layer neurons in the MLP is determined by the compression ratio r, which is set to half the number of input channels in this study. The outputs of the MLPs corresponding to the two feature representations are fused via an element-wise addition operation $\oplus$, followed by a sigmoid activation σ to generate channel weight scores as shown in Eq. (17):

$$\boldsymbol{Score} = \sigma\left(\boldsymbol{Out}_{GAP} \oplus \boldsymbol{Out}_{GMP}\right) \in \mathbb{R}^{T \times 1} \tag{17}$$

The attention scores are broadcast along the F direction using $B(\bullet)$ function, and the final attention matrix is calculated using the element-wise multiplication operation $\odot$ as shown in Eq. (18):

$$\boldsymbol{I}' = \boldsymbol{I} \odot B(\boldsymbol{Score}) \in \mathbb{R}^{T \times F} \tag{18}$$

where $\boldsymbol{I}'$ denotes the refined feature map. This output enables attention learning across different feature channels, thereby highlighting the feature channels that are strongly correlated with battery degradation information. After applying the attention mechanism, the resulting feature map is subjected to global average pooling to produce a vector GAP, which captures the average value across the temporal dimension for each channel. This vector is passed through a dense layer with a single neuron and a linear activation function to obtain the final prediction.

This model successfully captures the complex nonlinear dynamics of battery degradation by extracting multiscale spatial information, modelling temporal relationships, and fusing features. It offers a highly accurate technique for predicting lithium-ion battery capacity. To facilitate comprehension, key symbols used in MsCNN–LSTM feature fusion are recapitulated in Table 2:

Table 2 The key symbols used in MsCNN–LSTM feature fusion

Symbols	Definition	Symbols	Definition
C_{in}	Subsequences number	L	Window size
s	Stride of the sliding window	n	Input data length
$\boldsymbol{X}$	Input of the MsCNN and LSTM module	$\boldsymbol{P}_i$	Output of the i-th Conv1D branch
k_i	MsCNN kernels of the i-th branch	C_{out}	Output of the first MsCNN layer
C'_{out}	Output of the second MsCNN layer	$\boldsymbol{h}_t$	Hidden state at t-th step
$\boldsymbol{C}_t$	Cell state at t-th step	$\boldsymbol{i}_t$	Output of the input gate at the t-th step
$\sigma(\bullet)$	Sigmoid activation function	$\boldsymbol{W}$	Weights in the neuron networks
$\boldsymbol{b}$	Bias in the neuron networks	$\boldsymbol{f}_t$	Output of the forget gate at the t-th step
$\tilde{\boldsymbol{C}}_t$	Candidate cell state at t-th step	$\boldsymbol{o}_t$	Output gate result at the t-th step
$f(\bullet)$	BN, ReLU, and MaxPooling operations	$GAP(\bullet)$	Global average pooling operation
$GMP(\bullet)$	Global max pooling operation	$Dense(\bullet)$	Fully connected layer
$\boldsymbol{I}$	Result of the parallel branches	T	Neural units in the time-step dimension
F	Neural units in the feature dimension.	r	Compress ratio
$\oplus$	Element-wise addition operation	$B(\bullet)$	Broadcast operation
$\odot$	Element-wise multiplication operation	$\boldsymbol{I}'$	Output of the AT module

3. Algorithm Verification and Analysis

This section includes a complete description of the data used, including information on the data pretreatment techniques. It also describes the experimental design and evaluation measures for the upcoming ablation and comparative studies. The settings for model training and testing are also discussed.

3.1. Data description

The NASA dataset [25] provided by the NASA Prognostics Centre of Excellence (PCoE) is utilized to validate the effectiveness of the proposed model. The NASA dataset records the performance degradation data of various lithium-ion batteries under different charge-discharge cycles. It includes key indicators such as capacity fade and variations in internal resistance, which can comprehensively reflect the health status of the batteries. The experiment alternated charging and discharging according to the predetermined settings. All experiments with the 4 batteries were not stopped when the batteries reached the end-of-life (EOL) criteria. The EOL criteria here are a 30% fade in rated capacity (from 2 Ahr to 1.4 Ahr). Discharge data from four batteries named B0005, B0006, B0007, and B0018 are used in this work. The details of the NASA data are described in Table 3.

Table 3 The description of the NASA dataset

Battery Name	Discharge Samples	Discharge Criteria/ V	EOL Criteria/ Ah
B0005	168	2.7	1.4
B0006	168	2.5	
B0007	168	2.2	
B0018	132	2.5	

Using battery B0005 as an example, the battery gradually declines over the initial discharge cycles. As the number of charge-discharge cycles increases, the battery slowly declines—the discharge curves for cycles 50, 70, and 100 show a clear degradation trend. The time required to meet the discharge criterion gradually decreases.

The experiment assessed battery capacity after each discharge that met the predefined discharge criterion. When the capacity drops below 30%, the number of discharged-cycle samples varies by battery. Fig. 3 shows how the capacity of four batteries varies with the number of discharge cycles. Batteries B0005, B0006, B0007, and B0018 were discharged 124, 108, 168, and 132 times to reach their respective thresholds.

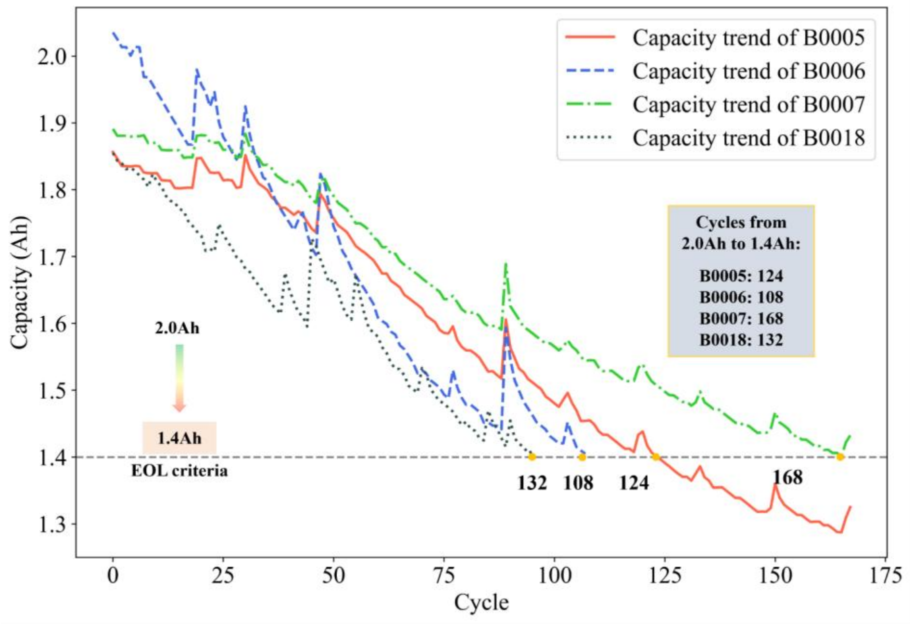


Fig. 3 Schematic diagram of capacity variation curves

3.2. Data processing

Discharge voltage directly reflects the effectiveness of the battery's internal chemical reactions and the health of the electrode materials, making it superior to other data for estimating battery capacity. Furthermore, to avoid time-consuming feature engineering, this study used raw discharge data as input for the model. Extracting statistical features related to battery degradation from multiple datasets using domain expertise is a standard and practical approach. Examples of such features include the duration to achieve the voltage plateau and various integral values. It is critical to note that the amount of data acquired from batteries in this field will gradually rise. As these datasets exhibit reciprocal influence and coupling, using deep learning algorithms to extract features from raw data directly will become more practical. This strategy has the disadvantage of allowing only partial interpretation.

Except for battery B0018, which maintains a consistent discharge data sampling rate, the other three batteries exhibit inconsistent sampling rates during the initial discharge stage and subsequent discharge cycles. This inconsistency leads to variable quantities of raw data points. Typical 1D-CNNs and LSTMs require constant input dimensions. Upsampling is used in this study to solve this issue. As previously stated, as the number of discharge cycles increases, the battery's usable time steadily reduces, resulting in fewer data points. After upsampling, the shorter sequences are padded with zeros at the end. The interpolation operation is exclusively applied to the NASA dataset to address the issue of inconsistent sampling rates in partial battery discharge voltage data.

To ensure the uniformity of data sampling frequency while maximizing the preservation of original signal characteristics, linear interpolation is adopted to upsample the voltage vector to a unified sampling rate of 0.1. Let $\boldsymbol{t} = [t_1, t_2, ...t_n]$ denote the time vector of the discharge voltage sequence, and $\boldsymbol{y} = [y_1, y_2, ...y_n]$ denote the corresponding discharge voltage vector, where y_i is the voltage at time t_i. The sampling frequency is initially calculated using the time intervals between consecutive data points, as shown in Eq. (19):

$$f = \frac{1}{t_{i+1} - t_i} \tag{19}$$

where t_i and t_{i+1} represent consecutive time points at time i and i+1 in the dataset. f is the sample rate of the voltage data. The trend in sampling rate across the NASA datasets is shown in Fig. 4. The new time and data values are computed using Eqs. (20) and (21).

$$t_{new} = t_i + \frac{t_{i+1} - t_i}{2} \tag{20}$$

$$y_{new} = \frac{y_{i+1} + y_i}{2} \tag{21}$$

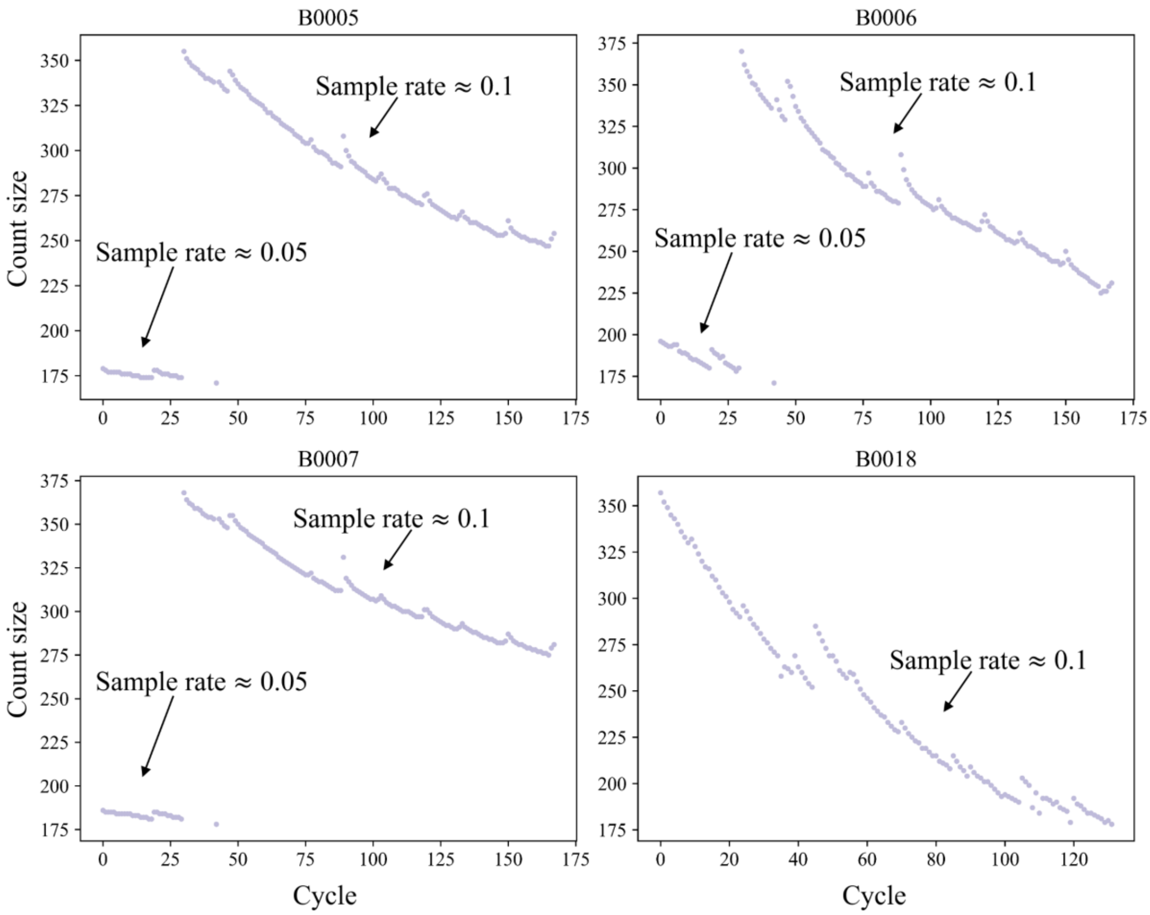


Fig. 4 Sampling rate trend in the NASA dataset

Fig. 4 shows that among the NASA battery dataset, only the battery cell B0018 maintains a uniform sampling rate, while the other three battery cells exhibit inconsistent sampling rates. When using deep learning models for feature extraction from raw data, it is critical to ensure temporal alignment across all datasets.

The voltage plateau is a core characteristic of the discharge voltage curve, which directly reflects the stability of electrochemical reactions and the consistency of energy release during the discharge process. To verify whether the interpolation operation preserves the physical characteristics of the original voltage signal. For example, on B0005, the voltage plateau of the voltage data before and after interpolation is calculated. The resulting MAE of 0.0031 indicates that the interpolation process caused negligible changes to the voltage plateau, confirming the validity of the interpolation method for subsequent data analysis. The trend of the voltage plateau is plotted as shown in Fig. 5.

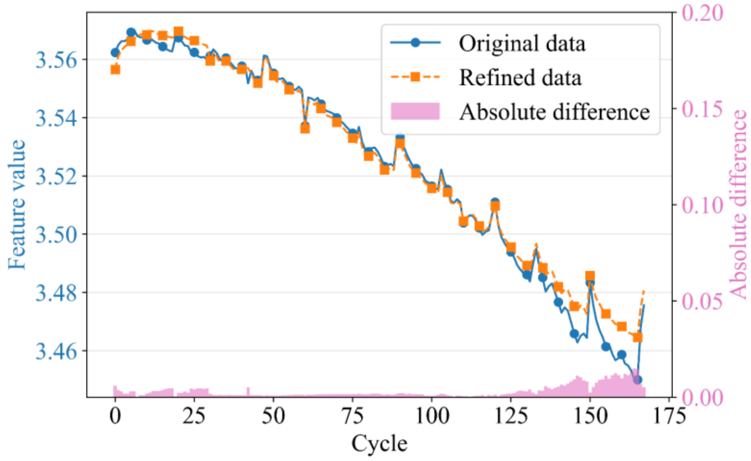


Fig. 5 Voltage plateau trend before and after interpolation in the B0005 dataset

Interpolation is not required for the CALCE and Oxford datasets, as the sampling rates remain consistent throughout the entire data sequence. After upsampling, if the length of the resulting data sequence is shorter than the maximum length $L_{\max}$, the zero-padding operation will be applied to the end of the sequence to ensure uniformity in the data dimensions. The padded sequence is constructed as Eq. (22).

$$Padding = \{y_1, y_2, ...y_n, 0, 0, ..., 0\} \tag{22}$$

where the number of zeros appended was $L_{\max} - n$. This approach assured that all data sequences had the exact dimensions. This trailing zero-padding strategy is equivalent to adopting a masking mechanism under the premise that the data is not standardized subsequently. The data is processed using the sliding window method to ensure interoperability with subsequent machine learning models, such as 1D-CNNs and LSTMs. The length of the voltage data determines the window size. To better adapt to various convolution and pooling operations across a specific number of layers in the model, small windows are used with short voltage data as input, and oversized windows with long voltage data. The sliding step is set to half the window size, resulting in an overlap rate of 0.5.

In this study, only data about capacity decline that meet the EOL criterion are used for model training and testing. The total number of samples per battery was 124, 108, 168, and 96, respectively. Table 4 presents a comprehensive segmentation of the NASA dataset with a test ratio of 0.3.

Table 4 The results of the data split for a test ratio of 0.3

Battery name	**Sample numbers**	**Train size**	**Test size**
B0005	124	87	37
B0006	108	76	32
B0007	168	118	50
B0018	96	68	28

3.3. Settings for model training, evaluation

All experiments are conducted on a Windows 10 (64-bit) computer equipped with an Intel Core i5-12400 CPU and 16 GB of RAM. No GPU acceleration is used during the experiments. The models are implemented and executed using Jupyter Notebook with the TensorFlow framework. During the ablation studies, the compression ratio *r* in the AT module is set to 2, and the number of hidden-layer units is set to half the number of time steps output by the preceding module. Ablation experiments are conducted on each model module to evaluate the proposed model's performance on NASA battery datasets. The settings for the ablation experiments are shown in Table 5.

Table 5 The setup of ablation experiments

Ablation Case	Model Name
Case 1	MsCNN
Case 2	MsCNN-AT
Case 3	LSTM
Case 4	LSTM-AT
Case 5	Parallel-MsCNN-LSTM
Case 6	Parallel-MsCNN-LSTM-AT

There are various essential factors in the model training process that help to maximize performance and ensure rigorous evaluation. A learning rate scheduler dynamically adjusts the learning rate every 200 epochs, reducing it by 0.05 to promote adaptive learning during long-term training. The model is built using the Adam optimizer with an initial learning rate of 0.005, mean squared error as the loss function, and MAE as the evaluation measure, both of which are usual for regression tasks. Training is supplemented with two callback functions: an early stopping mechanism that halts training if the validation loss does not improve after 20 consecutive epochs, preventing overfitting and ensuring the model's generalizability. The model is trained for a maximum of 1,500 epochs with a batch size of 8, and a 10% validation split is used to monitor performance during training. Table 6 presents the model's precise parameter settings.

Table 6 The parameters used in the proposed model

Module name	Parameter name	Parameter value
MsCNN	Input layer	(Times steps, Window size)
	Padding	same
	MaxPooling size	2
	Kernel regularizer	L2(0.001)
	Number of filters	[12,48], step = 6
	1st branch: Kernel size	3
	2nd branch: Kernel size	5
	3rd branch: Kernel size	7
LSTM	Input layer	(Times steps, Window size)
	Number of filters	[12,48], step = 6
	MaxPooling size	2
	Kernel regularizer	L2(0.001)
AT	Compress ratio (r)	0.5
Dense Layer	Dropout	0.3
	Activation	Linear
	Kernel regularizer	0.001

The window size should be carefully chosen according to the length of the voltage data. In this paper, various window size values were examined, and the same window size was utilized to compare different models to ensure fairness. To determine the number of neurons in the convolutional and LSTM layers, this study investigated parameter combinations with values of 12 and 48 at a step size of 6. Following that, a grid search was performed on the validation set to determine the appropriate hyperparameters for training the best model, which was then evaluated on the test set. The number of neurons in other models was tweaked using the same technique, while the number of stacked LSTM and convolutional layers in each branch remained constant at 2. A model degradation test was devised to assess the proposed model's capacity to forecast events early. In this experiment, the proportion of the test set is varied from 0.1 to 0.7 in 0.1-step increments to evaluate the model's performance.

To demonstrate the superiority of the proposed model, comparative experiments are conducted against several deep-learning-based models, including LSTM, BiLSTM, GRU, BiGRU, sequential MsCNN-LSTM, and sequential MsCNN-LSTM with channel attention. The experiments evaluate the performance using root mean squared error (RMSE), MAE, mean absolute percentage error (MAPE), and the coefficient of determination (R^2). Their expressions are as Eq. (23), Eq. (24), Eq. (25), and Eq. (26):

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{\mathrm{true},i}-y_{\mathrm{pred},i}\right)^2} \tag{23}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_{\mathrm{true},i}-y_{\mathrm{pred},i}\right| \tag{24}$$

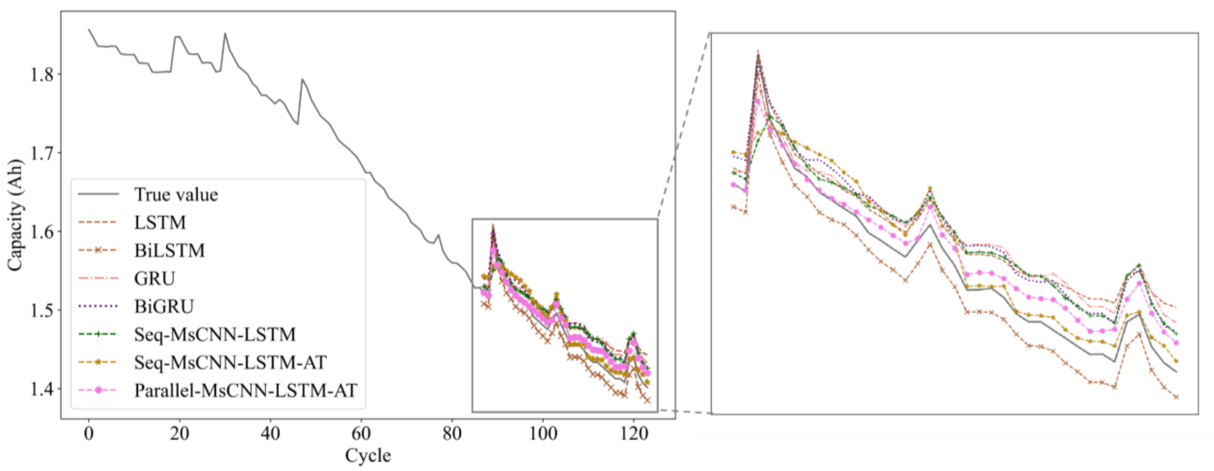
$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_{\mathrm{true},i}-y_{\mathrm{pred},i}}{y_{\mathrm{true},i}}\right|\times 100 \tag{25}$$

$$\mathrm{R}^2=1-\frac{\sum_{i=1}^{n}\left(y_{\mathrm{true},i}-y_{\mathrm{pred},i}\right)^2}{\sum_{i=1}^{n}\left(y_{\mathrm{true},i}-\bar{y}_{\mathrm{true}}\right)^2} \tag{26}$$

where n is the number of test samples. $y_{\mathrm{true},i}$ and $y_{\mathrm{pred},i}$ are the actual and predicted values for the i-th sample, respectively. $\bar{y}_{\mathrm{true}}$ is the mean of the actual values.

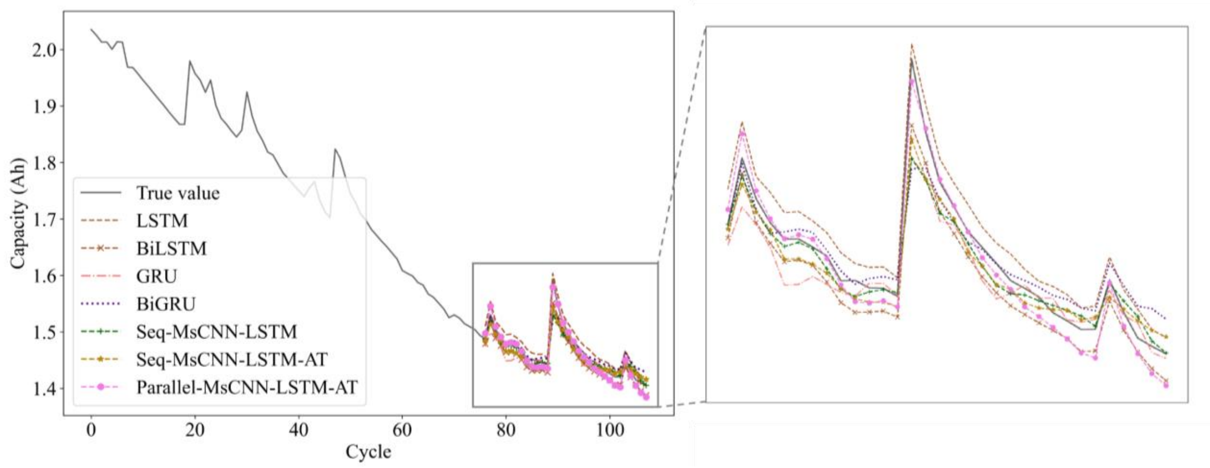
3.4. Comparative experiment

To demonstrate the proposed model's superiority on NASA battery datasets, Fig. 6, Tables 7, 8, 9, and 10 compare it to classic models, including RMSE, MAE, and MAPE values for the four batteries. In all comparative studies, due to the limited overall training data, a 7:3 training-to-validation split is used to validate model performance. Of course, model degradation experiments are conducted in the following sections to demonstrate the model's degradation as the training ratio decreases. The ratio refers to using the first 70% of the training samples to train the model and forecast the subsequent capacity change trends of the batteries, resulting in the prediction of individual battery performance. The sliding window method is used for processing input data. The NASA battery dataset's window size is set at 24, with a sliding coverage rate of 0.5.

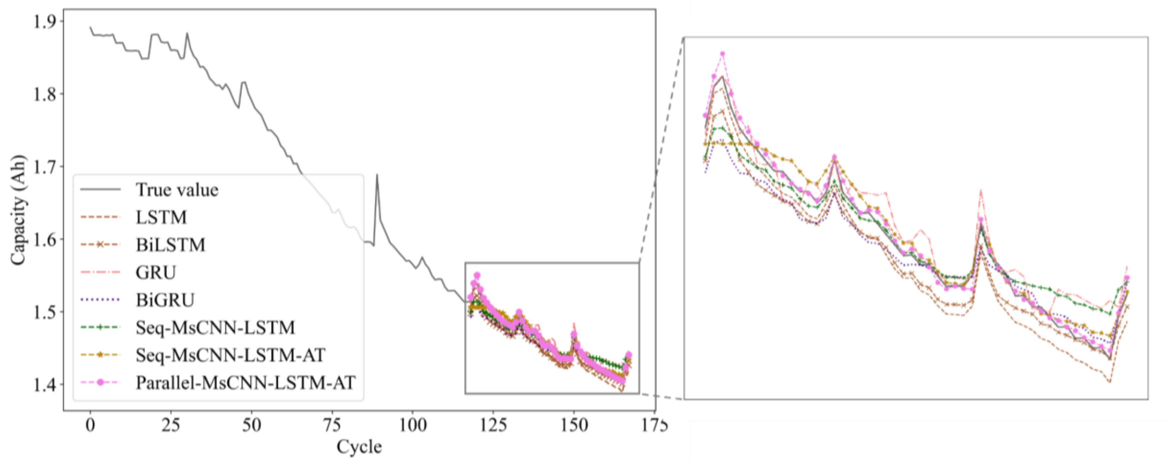


(a) Comparative results of metrics for B0005

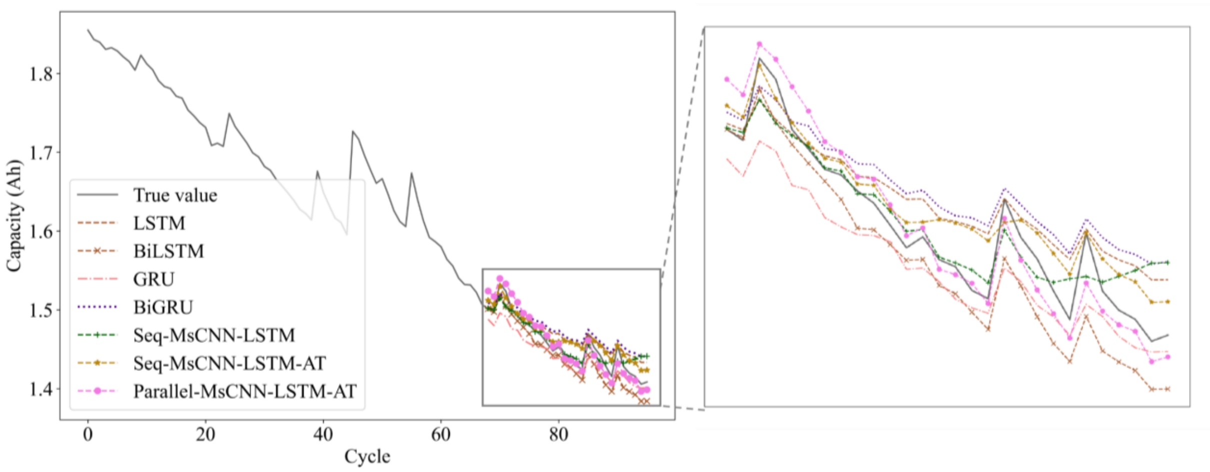
Fig. 6 Comparative results of the NASA battery datasets



(b) Comparative results of metrics for B0006



(c) Comparative results of metrics for B0007



(d) Comparative results of metrics for B0018

Fig. 6 Comparative results of the NASA battery datasets (continued)

As shown in Fig. 6, LSTM, GRU, and their variants BiLSTM and BiGRU, which are particularly effective for time-series modelling, all achieve good predictive performance at a test proportion of 0.3. Among them, there is little difference in prediction results between BiLSTM, BiGRU, ordinary LSTM, and GRU. Compared with the result of B0005, the prediction data of the other three batteries exhibit substantial fluctuations. In such cases, within the scope of single-model comparisons, LSTM performs slightly worse than GRU in handling such volatility. GRU achieves a lower MAE on volatile predictions, with an RMSE comparable to LSTM, thanks to its simplified single update gate that smooths short-term fluctuations while retaining the core temporal dependency modelling to fit long-term trends. Although the Seq-MsCNN-LSTM model and the sequential model with an attention mechanism show some improvement compared to the single model, the improvement is limited, and the attention mechanism does not perform well in the sequential model.

Table 7 shows the comparative results on battery B0005. As can be seen from the table, compared with both the single-model counterparts and the serial-model architectures, the proposed parallel model achieves the most accurate prediction results at approximately the exact prediction starting point, demonstrating that the parallel strategy offers distinct advantages in the battery capacity prediction task. The proposed parallel strategy achieves the lowest prediction error, with reductions of 43.24% and 18.24% in RMSE, and 48.30% and 14.32% in MAE compared to the Seq-MsCNN-LSTM and Seq-MsCNN-LSTM-AT models. Specifically, compared to the Seq-MsCNN-LSTM and Seq-MsCNN-LSTM-AT models, the RMSE and MAE are reduced by 32.35% and 29.75% on B0006, and by 63.50% and 57.34% on B0007, respectively, and by 8.66% and 25.64% on B0007, respectively. In the B0018 dataset with fewer samples, the RMSE is reduced by 15.31% and 15.25%, while the MAE shows relatively minor changes. The impact of limited training samples on the model is significant.

From the prediction results of B0006, B0007, and B0018, which exhibit substantial fluctuations in Fig. 6, it can be observed that the feature fusion and attention mechanism-based filtering in the proposed parallel model enhance the model's ability to identify fluctuation patterns. Compared with other models, the proposed model demonstrates superior predictive performance for capacity regeneration. In addition, the model with the attention mechanism achieves improved prediction stability compared with its counterpart without the attention mechanism. These results verify the superiority of the proposed parallel strategy in enhancing prediction accuracy across different battery datasets.

Given the disparities in forecast starting points, the table supplements the proposed model's outcomes at comparable forecast starting points. For example, the estimated beginning points of 101, 88, and 76 correspond to model performance at test ratios of 0.2, 0.3, and 0.4, respectively. It can be observed that the prediction error is the smallest when the forecast starting point is 100. Yao et al. [19] employed combined curve segments for prediction on the NASA dataset and also achieves favorable prediction performance. Guo et al. [26] adopted a serial combination of a CNN and a BiLSTM network. Liu et al. [27] employed a BiGRU model, while Hu et al. [28] utilized both GRU and LSTM models. All these studies adopted different prediction starting points. However, the proposed model exhibits distinct advantages in extracting multi-scale features from single discharge data and achieves lower errors in RMSE and MAE.

Table 7 Comparative results for the B0005

Test ratio	Estimate starting point	RMSE	MAE
LSTM	88	0.02386	0.02143
BiLSTM	88	0.01417	0.01404
GRU	88	0.02464	0.02256
BiGRU	88	0.02280	0.02195
Seq-MsCNN-LSTM	88	0.02345	0.02141
Seq-MsCNN-LSTM-AT	88	0.01628	0.01292
CNN-BiLSTM [26]	80	0.0476	0.0307
	100	0.0254	0.013
BiGRU [27]	100	0.00901	0.00632
	84	0.01025	0.00692
GRU & LSTM [28]	85	0.0298	0.0336
Reference [19]	-	0.1109	0.0944
Proposed	101	0.00343	0.00291
	88	0.01331	0.01107
	76	0.00930	0.00834

Tables 8, 9, and 10 show the comparative results on B0006, B0007, and B0018. The results demonstrate that at the identical forecast starting point, the sequential model outperforms the single model. This highlights the potential to incorporate spatiotemporal features into battery capacity prediction. After substituting the regular CNN architecture with a multi-scale architecture, the error is significantly reduced, demonstrating that the multi-scale architecture is critical for enriching the derived battery degradation information. The addition of an attention mechanism to the sequential model causes little change in error, implying that the attention mechanism has a limited influence on the sequential model's tiny number of final characteristics.

At the same forecast starting point of 88, the proposed model reduces the RMSE and MAE by 29.75%, 25.64% for B0006, 57.34%, 52.69% for B0007, and 15.25%, 13.68% for B0018, respectively, compared to the sequential model with attention mechanism, as shown in the results for B0006, B0007, and B0018. The parallel process improves the extraction of useful deterioration features, leading to more accurate predictions. The improvement is relatively minimal for B0018 due to a lack of samples, and this issue generally reduces the model's predictive performance.

Table 8 Comparative results for the B0006

Model	Estimate starting point	RMSE	MAE
LSTM	77	0.01693	0.01530
BiLSTM	77	0.01656	0.01556
GRU	77	0.01874	0.01364
BiGRU	77	0.01663	0.01114
Seq-MsCNN-LSTM	77	0.01459	0.00889
Seq-MsCNN-LSTM-AT	77	0.01405	0.01092
CNN-BiLSTM [26]	80	0.0274	0.0101
	100	0.0183	0.009
BiGRU [27]	100	0.01568	0.01098
	84	0.01946	0.01173
Reference [19]	-	0.2094	0.1562
Proposed	88	0.00819	0.00658
	77	0.00987	0.00812
	66	0.01482	0.01401

Table 9 Comparative results for the B0007

Model	Estimate starting point	RMSE	MAE
LSTM	119	0.01133	0.01081
BiLSTM	119	0.01059	0.00957
GRU	119	0.01103	0.00862
BiGRU	119	0.01151	0.00904
Seq-MsCNN-LSTM	119	0.01011	0.00764
Seq-MsCNN-LSTM-AT	119	0.00865	0.00613
BiGRU [27]	100	0.00826	0.00613
	84	0.00923	0.00638
Reference [19]	-	0.0566	0.0443
Proposed	134	0.00640	0.00528
	119	0.00369	0.00290
	102	0.01518	0.01448

Table 10 Comparative results for the B0018

Model	Estimate starting point	RMSE	MAE
LSTM	69	0.01765	0.01505
BiLSTM	69	0.01803	0.01570
GRU	69	0.01869	0.01657
BiGRU	69	0.02108	0.01882
Seq-MsCNN-LSTM	69	0.01358	0.00974
Seq-MsCNN-LSTM-AT	69	0.01357	0.01162
BiGRU [27]	100	0.01958	0.01260
	84	0.01977	0.01235
Reference [19]	-	0.1299	0.1019
Proposed	78	0.00872	0.00737
	69	0.01150	0.01003
	59	0.04144	0.03649

Furthermore, comparisons have been made to specific recent investigations. Based on the comparison results in the table with similar prediction starting points, the proposed model can achieve fewer errors. Specifically, when the prediction starting point is approximately 88, the suggested parallel method outperforms the serial RNN variants in the table on RMSE and MAE. Looking at the comparison findings with a prediction starting point of 100, the proposed model outperforms both the single BiGRU and the sequential CNN-BiLSTM, demonstrating the parallel strategy's predictive power.

3.5. Model degradation test

When the number of training samples is reduced, the ability to predict capacity declines, making early predictions more challenging. Tables 11, 12, 13, and 14 show the model test results for four datasets with test sample proportions ranging from 0.1 to 0.7 in increments of 0.1. It shows that as the proportion of test data increases, the volume of training data gradually decreases. Consequently, the model performance degrades to varying degrees across all four battery cells. Despite this degradation, the results demonstrate that the model still achieves highly accurate predictions at relatively early prediction points, thus verifying its competence for early-stage prediction.

Fig. 7 shows the trends in RMSE, MAE, and MAPE for the four batteries at various test set proportions. As seen in the tables and figure, the prediction error increases as the test proportion rises. The B0005, B0006, and B0007 batteries retain good predictive capacities at test proportions of 0.5, 0.4, and 0.4, respectively. The B0018 dataset, with fewer samples, can nonetheless provide a level of predictive potential at a test proportion of 0.3. The results show that the suggested model is particularly good at predicting early capacity. It can deeply mine multi-level characteristics in discharge data by fusing multi-scale spatiotemporal features and incorporating an attention module, achieving strong predictive performance even with few training samples.

Table 11 Model-degradation-test results of metrics for B0005 dataset

Test ratio	RMSE	MAE	MAPE(%)	R^2
0.1	0.00284	0.00219	0.97480	0.9380
0.2	0.00343	0.00291	2.14346	0.9843
0.3	0.01331	0.01107	3.51564	0.9268
0.4	0.00930	0.00834	4.38344	0.9747
0.5	0.01811	0.01648	5.47588	0.9458
0.6	0.05363	0.04807	6.18993	0.7202
0.7	0.12047	0.09655	12.9903	-

Table 12 Model-degradation-test results of metrics for B0006 dataset

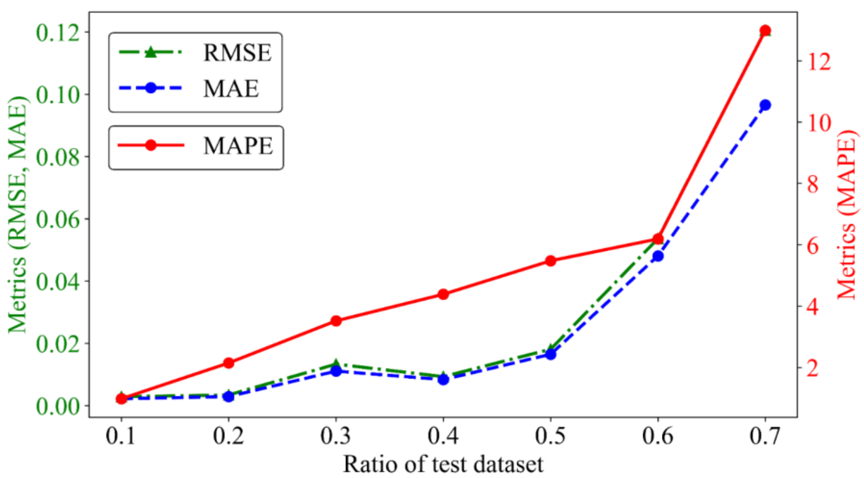
Test ratio	RMSE	MAE	MAPE(%)	R^2
0.1	0.00585	0.00448	1.30804	0.8165
0.2	0.00819	0.00658	3.56238	0.9688
0.3	0.00987	0.00812	3.31627	0.9424
0.4	0.01482	0.01401	3.69167	0.8993
0.5	0.07995	0.07338	8.89232	-
0.6	0.23552	0.21381	14.2854	-
0.7	0.40713	0.40334	25.2799	-

Table 13 Model-degradation-test results of metrics for B0007 dataset

Test ratio	RMSE	MAE	MAPE(%)	R^2
0.1	0.00234	0.00207	1.02159	0.9665
0.2	0.00640	0.00528	1.51609	0.9040
0.3	0.00369	0.00290	2.83574	0.9893
0.4	0.01518	0.01448	3.70064	0.8997
0.5	0.05353	0.04969	5.04243	-
0.6	0.08356	0.06798	5.43788	-
0.7	0.09285	0.07383	6.39670	-

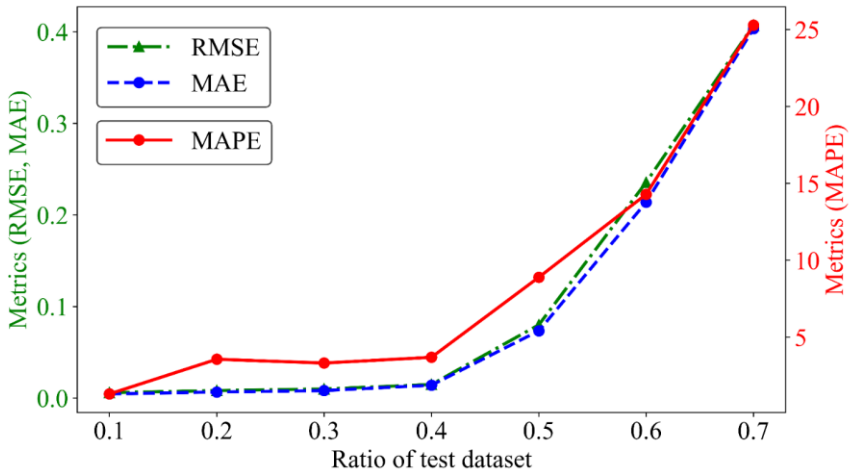
Table 14 Model-degradation-test results of metrics for B0018 dataset

Test ratio	RMSE	MAE	MAPE(%)	R^2
0.1	0.00319	0.00234	1.14457	0.9558
0.2	0.00872	0.00737	1.58444	0.7893
0.3	0.01150	0.01003	3.06279	0.8877
0.4	0.04144	0.03649	4.66569	-
0.5	0.05662	0.04500	4.71303	0.5147
0.6	0.09927	0.08703	6.61568	-
0.7	0.09297	0.08225	5.58500	-

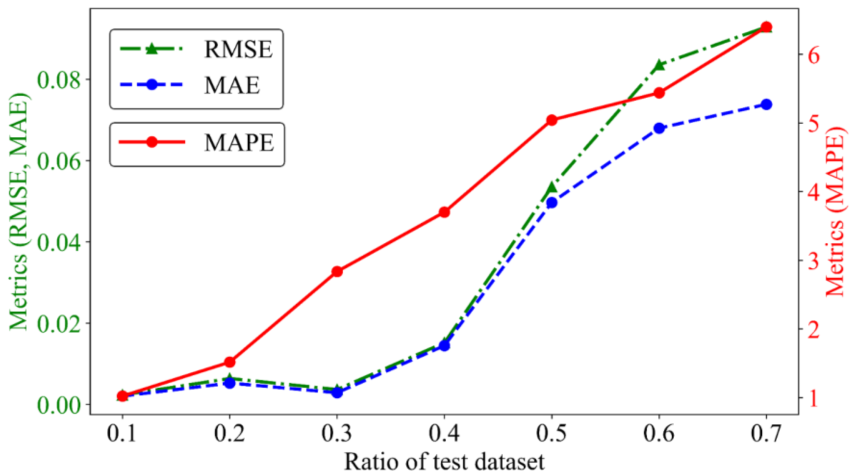


(a) The results of the B0005 battery

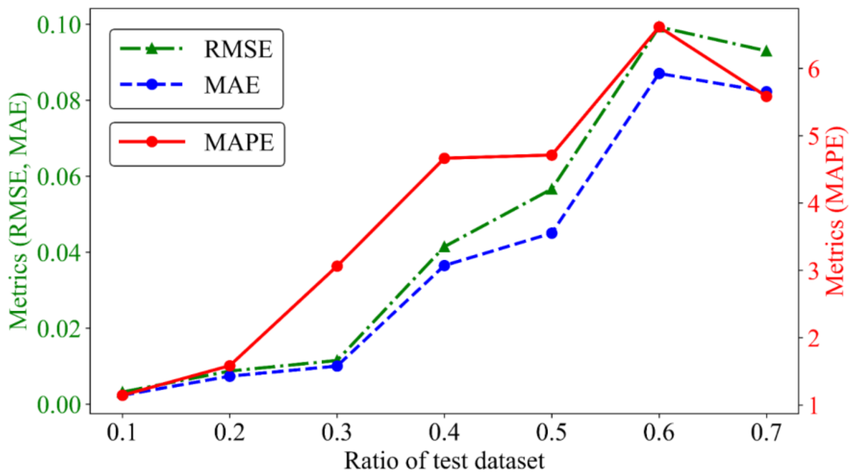
Fig. 7 Results under different test set proportions



(b) The results of the B0006 battery



(c) The results of the B0007 battery



(d) The results of the B0018 battery

Fig. 7 Results under different test set proportions (continued)

Taking B0005 as a representative example, which has mild capacity fluctuations, the performance degradation of the single LSTM model is tested, and the corresponding results are illustrated in Table 15 and Fig. 8. It can be observed that the single LSTM model undergoes rapid performance degradation when the test ratio ranges from 0.2 to 0.3. Specifically, under the condition of insufficient training samples, the single model suffers from overfitting, leading to poor robustness of the extracted features.

Table 15 Model-degradation-test results of single LSTM for B0005 dataset

Test ratio	RMSE	MAE	MAPE(%)	R^2
0.1	0.00452	0.00414	0.94321	0.8431
0.2	0.01420	0.01371	2.21863	0.7318
0.3	0.02386	0.02143	3.67334	0.7647
0.4	0.08796	0.07957	5.80814	-
0.5	0.16503	0.14890	10.07616	-
0.6	0.22340	0.20519	13.69088	-
0.7	0.31045	0.26072	16.18352	-

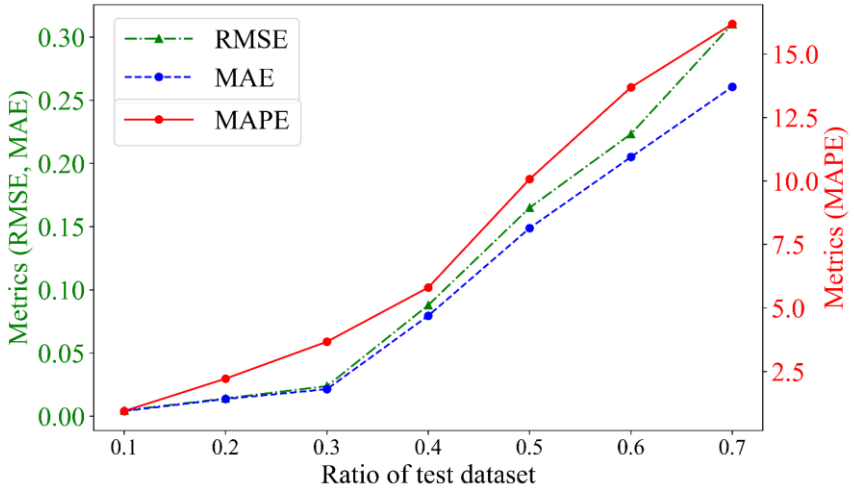


Fig. 8 Results of a single LSTM model under different test set proportions

3.6. Ablation experiment

Each module's efficiency is verified using ablation experiments. Following the experimental cases described in Table 5, model training and testing are performed on the B0005 dataset. The ablation results are shown in Table 16 and Fig. 9. The ablation studies are conducted with a test proportion of 0.3. This table quantifies the standalone performance of both the LSTM and MsCNN models, as well as the enhancement effect of the AT module on the single-model, sequential-model, and parallel-model architectures. Owing to the variations in sample sizes across different experimental groups, the results exhibit a certain degree of variability. As illustrated in Case 1 (using only the MsCNN model) and Case 3 (using only the LSTM model), the two models exhibit distinct predictive performance, with the LSTM model surpassing the MsCNN model. LSTM exhibits advantages in mining battery degradation data with strong temporal dependencies.

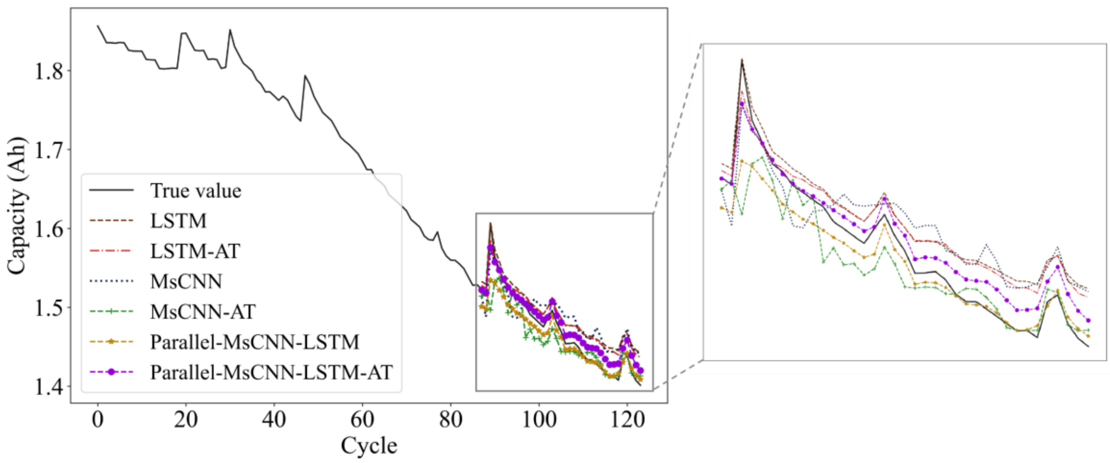
Compared with MsCNN, LSTM captures the time-varying characteristics of the battery degradation process, which are highly effective for mapping to capacity. For single models, prediction errors decrease with the addition of attention, but this trend varies across battery degradation patterns. For example, for B0005 with mild capacity fluctuations, the integration of the attention mechanism only yields a marginal improvement in predictive performance. For B0006, B0007, and B0018, which exhibit significant capacity fluctuations, the attention mechanism fails to improve performance. This phenomenon arises because single models extract insufficiently rich features. Further filtering by the attention mechanism may reduce the model's ability to capture degradation-related information.

In contrast, after improving feature richness via the parallel model, predictive performance is significantly better than in single- and sequential-architecture models. Using the B0005 dataset as an example, the parallel method reduced the RMSE and MAE by 26.45% and 43.77%, and by 35.02% and 50.76%, respectively, compared to the single LSTM and the single

MsCNN models. The parallel model with the AT module produced additional reductions of 24.16% and 8.13% in these measures when compared to the parallel model without the AT module, resulting in the lowest errors. When the attention mechanism is incorporated into the parallel model, targeted filtering of the rich feature set drastically boosts predictive performance and yields the optimal results. This outcome demonstrates the superiority of combining the AT with parallel feature extraction.

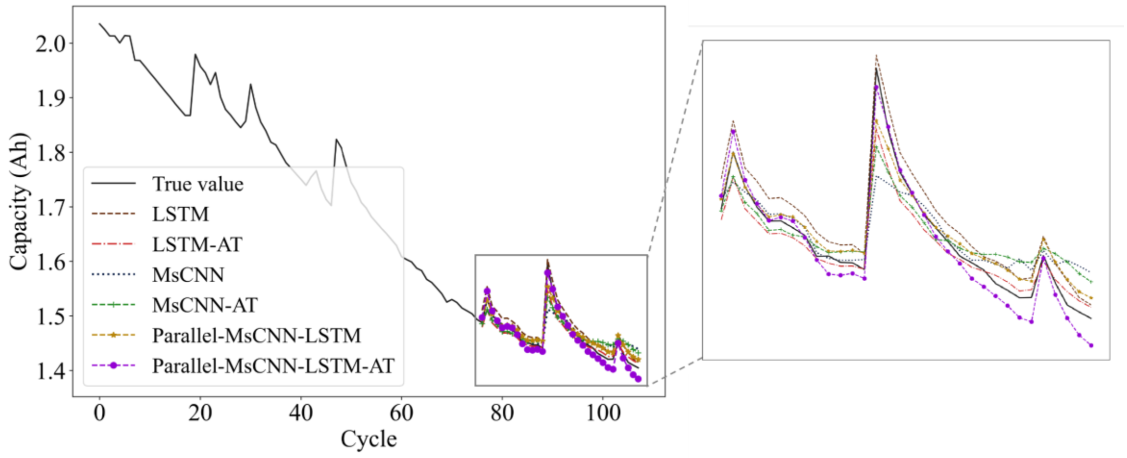
Table 16 Ablation results of metrics for the NASA dataset

Battery	Case	RMSE	MAE	MAPE(%)	R^2
B0005	1	0.02701	0.02447	3.39295	0.6985
	2	0.02410	0.01514	3.37350	0.7600
	3	0.02386	0.02143	3.67334	0.7647
	4	0.02188	0.01930	3.54123	0.8021
	5	0.01755	0.01205	3.32857	0.8727
	6	0.01331	0.01107	3.51564	0.9268
B0006	1	0.02156	0.01389	2.51581	0.7253
	2	0.01936	0.01479	2.46910	0.7783
	3	0.01548	0.01512	3.21010	0.8583
	4	0.01312	0.00957	2.64214	0.8982
	5	0.01213	0.00950	2.75775	0.9129
	6	0.00987	0.00812	3.31627	0.9424
B0007	1	0.01801	0.01492	2.39409	0.7442
	2	0.00983	0.00819	2.57843	0.9238
	3	0.01133	0.01081	2.93796	0.8988
	4	0.01034	0.00938	2.70081	0.9157
	5	0.00843	0.00781	2.7880	0.9439
	6	0.00369	0.00290	2.83574	0.9893
B0018	1	0.01900	0.01497	2.47814	0.6939
	2	0.01790	0.01538	3.04265	0.7281
	3	0.01765	0.01505	2.44314	0.7356
	4	0.01417	0.01300	2.67934	0.8297
	5	0.01180	0.01084	2.65850	0.8820
	6	0.01150	0.01003	3.06279	0.8877

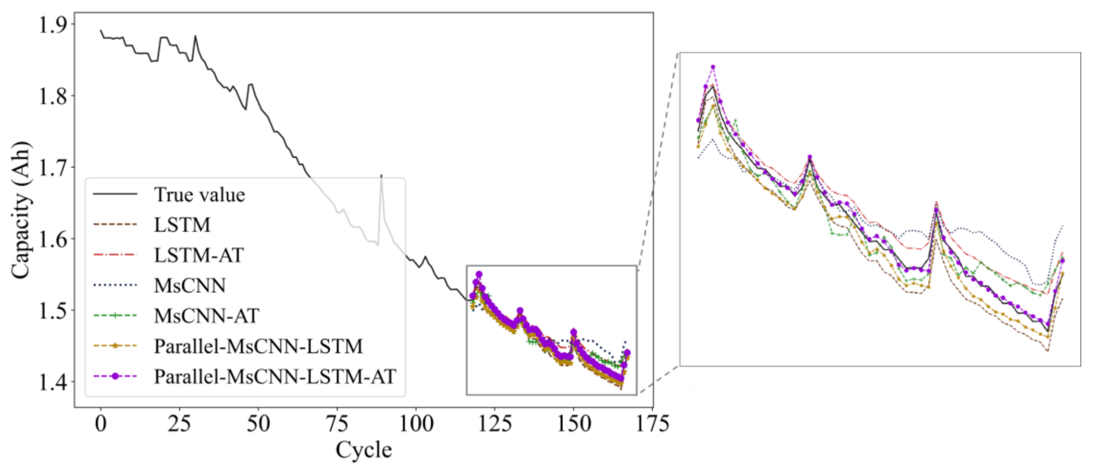


(a) Ablation result of the B0005 dataset

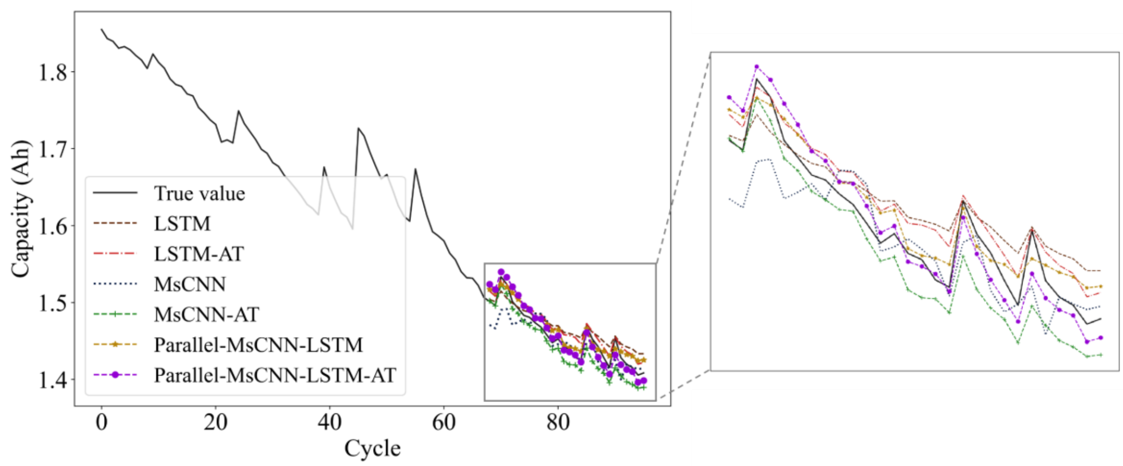
Fig. 9 Ablation results of the NASA datasets



(b) Ablation result of the B0006 dataset



(c) Ablation result of the B0007 dataset



(d) Ablation result of the B0018 dataset

Fig. 9 Ablation results of the NASA datasets (continued)

Further analysis of the results reveals that, among the single models, LSTM outperforms MsCNN due to LSTM's superior ability to capture the time-varying fluctuations in discharge voltage. The attention mechanism can further filter the features extracted by the standalone LSTM and MsCNN, yet the performance improvement is limited in this scenario. In contrast, when the parallel model extracts richer feature representations, the attention mechanism significantly enhances model performance. In summary, feature richness is critical for battery capacity prediction. The adaptive feature filtering effect of the attention mechanism is conditional: richer feature sets are more conducive to highlighting the effectiveness of attention-based filtering.

3.7. Model validation on other datasets

To further verify the validity of the model, two additional datasets are used, including the CS2-35 battery from the CALCE battery team [29] and Cell 1 from the Oxford Battery Degradation Dataset [30]. The detailed information of the two datasets is presented in Table 17.

Table 17 Detailed information of CS2-35 from the CALCE battery and Cell1 from the Oxford Battery Degradation Dataset

Battery name	Capacity Rating (mAh)	Charging Profile	Sample count
CS2-35	1100	Constant current charging at 0.5C until the voltage reaches 4.2V. Subsequent constant voltage charging at 4.2V until the charging current drops below 0.05A	893
Cell 1	740	Except first cycle: 2-C constant current charge at 1480 mA, conducted every 100 drive cycles, including 1-C cycles (740 mA) and pseudo-OCV cycles (40 mA)	78

The CS2-35 battery dataset has a large number of samples; however, due to the poor discharge-cycle sampling rate, there are only a few voltage data points. In contrast, the Cell 1 battery dataset has fewer samples but a higher sampling rate, resulting in more voltage data points. This influences the selection of various future parameters, such as the number of features and the window size. Finally, the window size for the CS2-35 battery is set to 6. Table 18, Table 19, Fig. 10, and Fig. 11 show the test results using the same test ratio of 0.3. The capacity labels of the CS2-35 battery are calculated based on the integral of discharge current over time, with the core formula and implementation details as Eq. (27):

$$Q = \sum_{t=1}^{n}\left(\Delta t_t \times I_t\right)/3600 \tag{27}$$

where Q denotes discharge capacity (in Ah), the final capacity label for each cycle. Δt_t denotes the time interval between consecutive sampling points (in seconds). I_t denote discharge current at time *t* (in A). *n* denotes the total number of sampling points in the discharge phase (Step_Index = 7 in the CALCE dataset). For the Oxford dataset, the capacity labels are defined as the maximum charging capacity for each cycle, calculated by extracting the final value of the charge capacity curve from the charging-phase data.

Table 18 Comparative results from the CALCE battery dataset (CS2-35 battery)

Model	Estimate starting point	RMSE	MAE
LSTM	626	0.0787	0.0604
BiLSTM		0.1026	0.0678
GRU		0.1146	0.1015
BiGRU		0.0787	0.0672
Seq-MsCNN-LSTM		0.0570	0.0487
Seq-MsCNN-LSTM-AT		0.0514	0.0488
Reference [19]		0.0135	0.0096
Proposed		0.0498	0.0370

Comparative results from the CALCE battery dataset (CS2-35) are shown in Table 18. The results demonstrate that the proposed model achieves the lowest prediction error with identical single-channel discharge-voltage data and exhibits superior ability to characterize voltage fluctuations. Nevertheless, the performance of the proposed model is inferior to that of the reference models that employ multi-channel data, indicating promising avenues for improving multi-channel parallel feature fusion in future research. Benefiting from the large volume of training samples, LSTM can fully leverage its structural merits. Single-model test results show that LSTM outperforms GRU. Although BiLSTM incorporates bidirectional information flow, it fails to outperform LSTM. When the training sample size increases, BiGRU achieves performance comparable to LSTM, yet its bidirectional architecture inevitably increases computational complexity.

A single model can estimate capacity to some extent, but its performance is significantly lower than that of the combined model. After incorporating the attention mechanism into the sequential model, the model's prediction errors (RMSE) dropped by 9.82%. Its MAE varies slightly, showing that the attention mechanism does not operate effectively in the sequential model. When the parallel mode is adopted, the RMSE and MAE decrease by 3.11% and 24.18%, respectively. Yao et al. [19] comprehensively utilize various types of data on the CS2-35 battery to achieve data fusion. It demonstrated strong predictive performance by integrating battery degradation information from multiple channels. The proposed model uses only discharge data; hence, there is an opportunity to improve model performance through multi-feature fusion further.

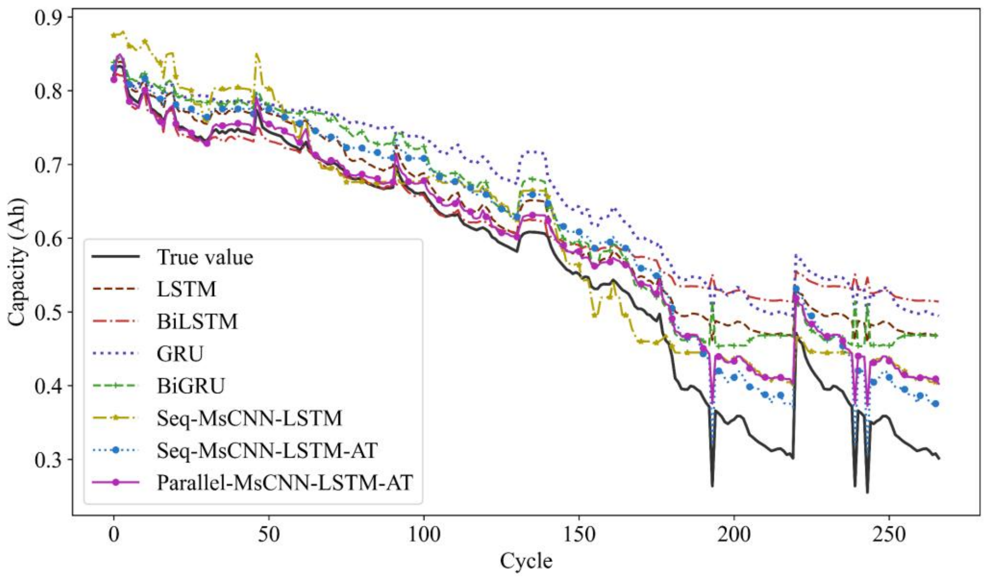


Fig. 10 The comparative results on the CS2-35 battery

As seen in Fig. 10, severe capacity fluctuations occur in the late stages of battery breakdown. This variation has a significant impact on the prediction process, ultimately leading to excessive prediction errors. Features generated solely from discharge-voltage data have limited ability to capture fluctuations, such as capacity renewal. As a result, the incorporation of new degradation-related features is considered as a feasible method to improve prediction performance.

Table 19 Comparative results from the Oxford Battery Aging Dataset (Cell 1)

Model	Forecast starting point	RMSE	MAE	R^2
LSTM	56	4.3478	4.0133	0.7879
BiLSTM		5.9814	5.3153	0.5986
GRU		4.4284	3.5585	0.7799
BiGRU		4.7326	3.7310	0.7487
Seq-MsCNN-LSTM		2.4404	2.2378	0.9332
Seq-MsCNN-LSTM-AT		2.9331	2.6665	0.9035
Proposed		1.8196	1.6103	0.9628

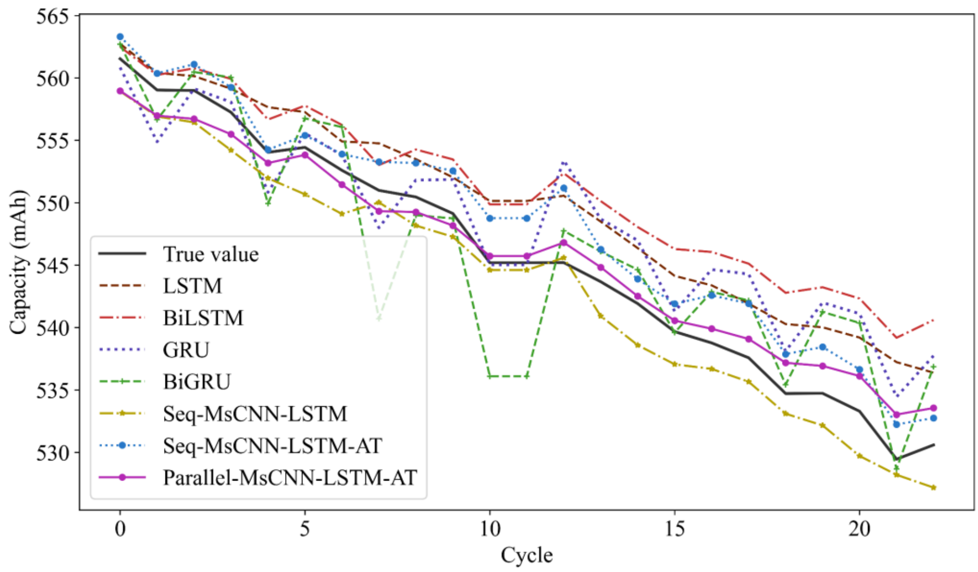


Fig. 11 The comparative results on the Cell 1 battery

The experimental results from the Oxford Battery Degradation Dataset (Cell 1) are compatible with the CS2-35 dataset. Comparative results from the Oxford Battery Degradation Dataset (Cell 1) are shown in Table 19 and Fig. 11. With a relatively small training sample size, the LSTM and BiLSTM models underperform GRU and BiGRU in terms of MAE. Additionally, BiGRU exhibits substantial fluctuations in prediction, indicating unstable predictive performance. The prediction error of a single model falls dramatically when integrated with multiple models. Compared to the sequential approach, the parallel model reduces the RMSE and MAE by 37.96% and 39.61%, respectively. The proposed model achieves the lowest error and best performance among the comparison models, demonstrating its validity.

In the experiment, it was found that the Oxford Battery Degradation Dataset contains more extended sequences of voltage-sampling data. In contrast, the NASA and CALCE datasets have fewer voltage data points. The selection of an appropriate window size has a substantial impact on model performance. As a result, experiments are conducted on Cell 1 using the proposed model to evaluate its performance across various window widths. During the tests, the window size ranges from 6 to 54, with a step size of 6. The window coverage rate is set to 0.5, while all other model parameters remain unchanged. The experimental results are shown in Table 20.

Table 20 Performance with different window sizes on Cell 1 of the Oxford Battery Degradation Dataset

Window size	RMSE	MAE	MAPE(%)	R^2
6	4.5728	4.0245	1.9117	0.7654
12	3.0941	2.5052	1.8384	0.8926
18	2.3332	2.0233	1.8401	0.9389
24	**1.8196**	**1.6103**	**1.8339**	**0.9628**
30	2.0721	1.6224	1.9848	0.9518
36	2.7462	2.2778	1.8515	0.9154
42	2.5720	2.1100	1.8897	0.9258
48	2.9160	2.4862	2.0022	0.9046
54	6.8492	5.9557	1.8847	0.4736

The voltage data points for each cycle of Cell 1 are processed until they surpass 3,000. When using the sliding window method, the window size affects the model. Table 20 shows the performance with different window sizes on Cell 1 of the Oxford Battery Degradation Dataset. It shows that both smaller and larger window sizes reduce model performance, and the ideal window size is 24. A small window size fails to capture the long-term degradation trends of the battery, leading to increased prediction errors, especially when confronted with severe fluctuations. In contrast, a large window size introduces redundant noise from irrelevant cycles, thereby impairing the model's focus on key degradation features.

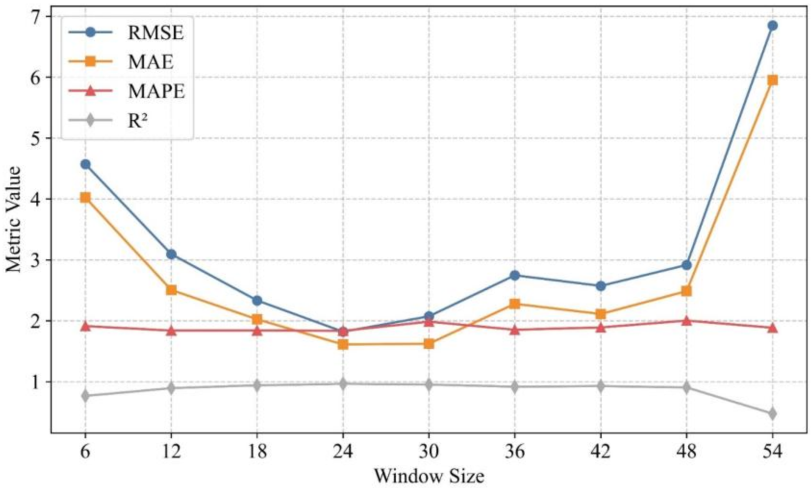


Fig. 12 Results on the Cell 1 battery with different window sizes

4. Conclusions

This paper proposes a parallel, multi-scale spatiotemporal feature fusion model integrated with an attention mechanism to enhance the accuracy of battery capacity prediction. Comprehensive experiments on NASA, CALCE, and Oxford Battery Degradation datasets verify that the proposed model outperforms some existing models, and the core findings and contributions are summarized as follows:

(1) The proposed hybrid parallel model outperforms single-component models (e.g., GRU, LSTM, MsCNN) on the NASA datasets. The multi-scale module enables the effective extraction of rich battery degradation features, and the parallel architecture demonstrates clear advantages over the serial architecture.

(2) In the NASA datasets, the attention mechanism shows a limited improvement on single models due to insufficiently rich degradation-related features extracted by them. In contrast, it significantly improves the performance of the parallel model, enabling it to capture more comprehensive features.

(3) Although both the proposed model and the single LSTM model degrade in performance as the number of training samples decreases, the proposed model degrades later, indicating better robustness to small sample sizes.

(4) The sliding window size has a significant impact on model performance, and the optimal window size is around 24, as verified on the Oxford dataset.

There are also some limitations in this research:

(1) The model only uses discharge-voltage time-series data, which is difficult to obtain in practical applications.

(2) Only single-channel data is used, and information loss may lead to performance bottlenecks.

(3) There are limitations in the interpretability of the attention mechanism, and the in-depth complexity analysis across different modules is insufficient.

Future research will focus on the following aspects: multi-channel information aggregation and computational complexity balance; interpretability of model mechanisms and uncertainty quantification; and prediction under fragmented data and small sample conditions.

Acknowledgments

This work was funded by the Shandong Provincial Key Research and Development Program (2024CXPT011), the National Key Research and Development Program of China (2024YFB3312302), the Qilu Institute of Technology for Talent Project (QIT24TP027), and the research program of Qilu Institute of Technology (QIT255NN019).

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] R. R. Kumar, C. Bharatiraja, K. Udhayakumar, S. Devakirubakaran, K. S. Sekar, and L. Mihet-Popa, "Advances in Batteries, Battery Modeling, Battery Management System, Battery Thermal Management, SOC, SOH, and Charge/Discharge Characteristics in EV Applications," IEEE Access, vol. 11, pp. 105761-105809, 2023.

[2] S. Ji, J. Zhu, Y. Yang, G. Dos Reis, and Z. Zhang, "Data-Driven Battery Characterization and Prognosis: Recent Progress, Challenges, and Prospects," Small Methods, vol. 8, no. 7, article no. 2301021, 2024.

[3] Y. He, W. Bai, L. Wang, H. Wu, and M. Ding, "SOH Estimation for Lithium-Ion Batteries: An Improved GPR Optimization Method Based on the Developed Feature Extraction," Journal of Energy Storage, vol. 83, article no. 110678, 2024.

[4] Y. Zhi, H. Wang, and L. Wang, "A State of Health Estimation Method for Electric Vehicle Li-Ion Batteries Using GA-PSO-SVR," Complex & Intelligent Systems, vol. 8, no. 3, pp. 2167–2182, 2022.

[5] X. Wang, B. Hu, X. Su, L. Xu, and D. Zhu, "State of Health Estimation for Lithium-Ion Batteries Using Random Forest and Gated Recurrent Unit," Journal of Energy Storage, vol. 76, article no. 109796, 2024.

[6] H. Feng and H. Yan, "State of Health Estimation of Large-Cycle Lithium-Ion Batteries Based on Error Compensation of Autoregressive Model," Journal of Energy Storage, vol. 52, Part B, article no. 104869, 2022.

[7] B. Ma, W. Guo, and Z. S. Li, "Prediction of Lithium-Ion Battery Capacity Based on the ARMA Method," Proceedings of IEEE Conference on Global Reliability and Prognostics and Health Management (PHM-Shanghai), pp. 1-7, 2020.

[8] S. Kim, P. Y. Lee, M. Lee, J. Kim, and W. Na, "Improved State-of-Health Prediction Based on Auto-Regressive Integrated Moving Average with Exogenous Variables Model in Overcoming Battery Degradation-Dependent Internal Parameter Variation," Journal of Energy Storage, vol. 46, article no. 103888, 2022.

[9] J. Wu, J. Wang, M. Lin, and J. Meng, "Retired Battery Capacity Screening Based on Deep Learning with Embedded Feature Smoothing Under Massive Imbalanced Data," Energy, vol. 318, article no. 134761, 2025.

[10] N. He, Q. Wang, Z. Lu, Y. Chai, and F. Yang, "Early Prediction of Battery Lifetime Based on Graphical Features and Convolutional Neural Networks," Applied Energy, vol. 353, Part A, article no. 122048, 2024.

[11] C. Li, H. Zhang, P. Ding, S. Yang, and Y. Bai, "Deep Feature Extraction in Lifetime Prognostics of Lithium-Ion Batteries: Advances, Challenges and Perspectives," Renewable and Sustainable Energy Reviews, vol. 184, article no. 113576, 2023.

[12] D. Chung, J. Ko, and K. Yoon, "State-of-Charge Estimation of Lithium-ion Batteries Using LSTM Deep Learning Method," Journal of Electrical Engineering & Technology, vol. 17, no. 3, pp. 1931–1945, 2022.

[13] J. Guo, M. Liu, P. Luo, X. Chen, H. Yu, and X. Wei, "Attention-Based BiLSTM for the Degradation Trend Prediction of Lithium Battery," Energy Reports, vol. 9, pp. 655-664, 2023.

[14] X. Wang, J. Wang, S. Zhang, Z. Zhang, Y. Li, and X. Liu, "BiLSTM-Akef Hybrid Driven Lithium-Ion Battery SOH Prediction Model Based on CEEMD with Different Distributions Noises," Journal of Energy Storage, vol. 111, article no. 115437, 2025.

[15] X. Wang, K. Dai, M. Hu, and N. Ni, "Lithium-Ion Battery Health State and Remaining Useful Life Prediction Based on Hybrid Model MFE-GRU-TCA," Journal of Energy Storage, vol. 95, article no. 112442, 2024.

[16] R. Sun, J. Chen, B. Li, and C. Piao, "State of Health Estimation for Lithium-Ion Batteries Based on Novel Feature Extraction and BiGRU-Attention Model," Energy, vol. 319, article no. 134756, 2025.

[17] Z. Lu, Z. Fei, B. Wang, and F. Yang, "A Feature Fusion-Based Convolutional Neural Network for Battery State-of-Health Estimation with Mining of Partial Voltage Curve," Energy, vol. 288, article no. 129690, 2024.

[18] P. J. Weddle, S. Kim, B. Chen, Z. Yi, P. Gasper, et al., "Battery State-of-Health Diagnostics During Fast Cycling Using Physics-Informed Deep-Learning," Journal of Power Sources, vol. 585, article no. 233582, 2023.

[19] L. Yao, C. Zhao, Y. Xiao, H. Dai, Z. Fei, and L. Zhang, "Battery Health Prediction Using Two-Dimensional Multi-Channel Ensemble Models," Journal of Energy Storage, vol. 86, article no. 111304, 2024.

[20] X. Li, C. Gwan, S. Zhao, X. Gao, and Y. Zhu, "Multimodal Temperature Prediction for Lithium-Ion Battery Thermal Runaway Using Multi-Scale Gated Fusion and Bidirectional Cross-Attention Mechanisms," Journal of Energy Storage, vol. 116, article no. 116098, 2025.

[21] Y. Jiang, Y. Chen, F. Yang, and W. Peng, "State of Health Estimation of Lithium-Ion Battery with Automatic Feature Extraction and Self-Attention Learning Mechanism," Journal of Power Sources, vol. 556, article no. 232466, 2023.

[22] J. Park, G. H. Lee, J. Kim, Y. Yoo, and I. Lee, "Detailed Architectural Design of a Multi-Head Self-Attention Model for Lithium-Ion Battery Capacity Forecasting," IEEE Access, vol. 13, pp. 48212-48225, 2025.

[23] Z. He, X. Ni, C. Pan, S. Hu, and S. Han, "Full-Process Electric Vehicles Battery State of Health Estimation Based on Informer Novel Model," Journal of Energy Storage, vol. 72, article no. 108626, 2023.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141, 2018.

[25] B. Sahaand and K. Goebel, "Battery Data Set." NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA.

[26] X. Guo, K. Wang, S. Yao, G. Fu, and Y. Ning, "RUL Prediction of Lithium Ion Battery Based on CEEMDAN-CNN BiLSTM Model," Energy Reports, vol. 9, pp. 1299-1306, 2023.

[27] L. Liu, W. Sun, C. Yue, Y. Zhu, and W. Xia, "Remaining Useful Life Estimation of Lithium-Ion Batteries Based on Small Sample Models," Energies, vol. 17, no. 19, article no. 4932, 2024.

[28] L. Hu, W. Wang, and G. Ding, "RUL Prediction for Lithium-Ion Batteries Based on Variational Mode Decomposition and Hybrid Network Model," Signal, Image and Video Processing, vol. 17, no. 6, pp. 3109-3117, 2023.

[29] 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," Journal of Power Sources, vol. 196, no. 23, pp. 10314–10321, 2011.

[30] [C. Birkl, "Diagnosis and Prognosis of Degradation in Lithium-Ion Batteries," Doctor of Philosophy, University of Oxford, 2017.