2026-01-0460 Published 07 Apr 2026

Enabling Sustainable E-Mobility: An Edge–Cloud Collaborative Framework for Battery Lifecycle Health Management in Electric Vehicles

Weimin Gao South China University of Technology

Zhilong Lv Hubei University of Arts and Science

Shiqi(Shawn) Ou South China University of Technology

Citation: Gao, W., Lv, Z., and Ou, S., "Enabling Sustainable E-Mobility: An Edge–Cloud Collaborative Framework for Battery Lifecycle Health Management in Electric Vehicles," SAE Technical Paper 2026-01-0460, 2026, doi:10.4271/2026-01-0460.

Received: 06 Nov 2025 Revised: 12 Dec 2025 Accepted: 26 Jan 2026

Abstract

The rapid adoption of electric vehicles (EVs) is a cornerstone of the transition to sustainable transportation. However, uncertainty regarding battery degradation remains a significant obstacle, hindering vehicle energy efficiency, operational safety, and the recovery of end-of-life value. Accurate estimation of the battery state of health (SOH) and prediction of the remaining useful life (RUL) are therefore critical for sustainable vehicle lifecycle management. This study proposes an edge–cloud collaborative intelligent framework for in-vehicle deployment that leverages a Transformer-based architecture to jointly model SOH and RUL. The cloud-side model retains the full configuration to capture long-term degradation trajectories for high-accuracy RUL prediction. A lightweight edge-side model, engineered via pruning and knowledge distillation, delivers millisecond-level inference for real-time SOH estimation onboard the vehicle. To ensure efficiency, only four core health indicators are extracted for end-to-end prediction. Experimental validation across 77 battery cells demonstrates that the framework achieves SOH estimation with a root mean square error (RMSE) of 1.41% and RUL prediction with an RMSE of 2.59% (78 cycles). Furthermore, a periodic cloud-side update and over-the-air deployment mechanism ensure long-term adaptability and cross-platform scalability without full local retraining. This intelligent prognostic framework directly enhances EV reliability and sustainability by providing health-informed decision support for optimal vehicle operation, maintenance scheduling, and the reuse of second-life batteries. Consequently, it serves as a vital tool for advancing resource optimization and circular economy principles within the E-mobility ecosystem.

Keywords

Battery lifecycle management, Deep learning, Edge–cloud collaboration, Remaining useful life, State-of-health, Sustainable electric mobility

Introduction

The rapid proliferation of electric vehicles (EVs) has become a major driver of a sustainable transformation of the transportation sector. As the core energy source of EVs, lithium-ion batteries (LiBs) play a crucial role in overall vehicle performance, operational safety, and lifecycle management. However, the degradation of LiBs is influenced by multiple factors, including temperature, charge/discharge rate, and operating conditions [1]. The highly nonlinear and coupled nature of these factors makes accurate assessment of battery health states particularly challenging. This issue has emerged as a key technical bottleneck restricting the reliable operation and large-scale industrialization of EVs. Therefore, accurately estimating the battery state of health (SOH) and reliably predicting the remaining useful life (RUL) of batteries has become a central task in battery management system (BMS) research and an important focus in the field of EVs.

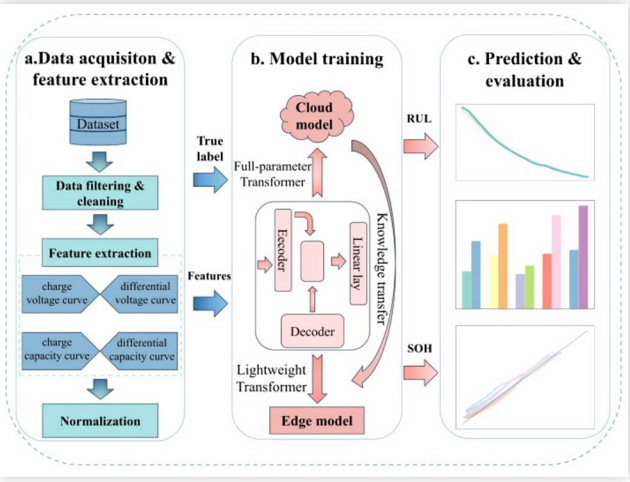
Traditional battery lifetime prediction methods mainly include mechanism-based models and empirical models

[2]. Mechanism- based approaches—such as electrochemical models or equivalent circuit–physical models (e.g., extended kinetic models [3] and particle diffusion models [4])—can capture the underlying physical degradation processes. However, they often involve complex modeling, high parameter sensitivity, and limited adaptability to diverse operating conditions [5]. Empirical or semi-empirical models (such as exponential decay models [6], regression models [7], and simplified state-space models [8]) are easier to implement but suffer from poor generalization, often performing inadequately across different battery types or operating conditions [9]. With advances in sensing and data acquisition technologies, data-driven approaches, particularly those based on deep learning, have gradually become mainstream in battery health prediction research [10, 11, 12, 13]. Recent studies have employed various architectures such as deep neural network swarms [14], convolutional neural networks (CNNs) [15], recurrent neural networks (RNNs) [16], and long short-term memory (LSTM) networks [17] to predict RUL or SOH, achieving promising results. However, traditional RNN- or CNN-based models still face limitations in long-sequence modeling, such as gradient vanishing and insufficient capture of dependencies, making it difficult to comprehensively characterize the long-term degradation behavior of batteries [18].

In recent years, the Transformer model has achieved remarkable success in fields such as natural language processing and time-series forecasting due to its powerful sequence modeling capability and global attention mechanism [19]. The Transformer is a network architecture entirely based on the attention mechanism. Its core lies in completely abandoning traditional recurrent or convolutional architectures, instead using multi-head attention to capture global dependencies within a sequence. The Transformer model adopts the classic encoder–decoder architecture, with both the encoder and the decoder composed of several identical stacked layers. Each encoder layer contains two sub-layers: the first is multi-head self-attention, and the second is a position-wise fully connected feed-forward network. The decoder additionally inserts a third sub-layer for performing multi-head attention over the encoder's outputs. The self-attention mechanism computes representations of a sequence by relating information from different positions within the same sequence, dynamically aggregating contextual information. Multi-head attention enables the model to focus on information from different representation subspaces and positions simultaneously. To retain the sequential order, positional encodings are injected into the input embeddings. Compared with RNN and CNN, the Transformer can more effectively capture long-term dependencies and enable parallel computation, demonstrating great potential for modeling the complex dynamic degradation processes of batteries. Consequently, Transformer-based data-driven prediction methods have been increasingly adopted for battery health management, demonstrating excellent predictive accuracy and generalization. For example, Li et al. proposed a data-driven framework combining a simplified physical model, subspace identification, and a Transformer for lithium-ion battery SOH prediction, achieving highly accurate and well-generalized health estimation across different cycling stages and battery types [20]. Chen et al. developed a regression network based on the Vision Transformer (ViT) for lithium-ion battery SOH estimation, which achieves high prediction accuracy and fast convergence through adaptive sampling selection and an improved ViT architecture [21]. However, existing methods still face several critical challenges. Traditional RNN and CNN architectures often suffer from gradient vanishing, difficulty in capturing long-term dependencies, and limited parallelism when modeling long time-series data, making it difficult to effectively characterize battery degradation trends over the entire lifecycle. Moreover, many existing approaches fail to balance prediction accuracy and deployment efficiency, lack a systematic design for integrating onboard real-time prediction with cloud-based remote updates, and thus struggle to meet the dual requirements of high reliability and resource constraints in electric vehicle applications.

To address these challenges, this study proposes an edge–cloud collaborative Transformer framework (E-CCT) for full-lifecycle battery health management in EVs, as illustrated in Figure 1. The proposed framework integrates the representational power of deep learning for complex, dynamic modeling with the systemic advantages of the edge–cloud architecture, establishing an intelligent, scalable battery health perception and prediction system that balances accuracy, efficiency, and extensibility. Specifically, the cloud model, built upon an enhanced Transformer architecture, leverages its global attention mechanism and long-sequence modeling capability to extract features and model degradation trajectories across various charging/discharging strategies and degradation stages, thereby achieving high-precision RUL prediction. To meet the stringent requirements of onboard systems for computational resources and latency, the E-CCT framework incorporates a lightweight edge model

FIGURE 1 The proposed edge-cloud collaborative architecture for estimating SOH and predicting RUL of electric vehicle batteries.



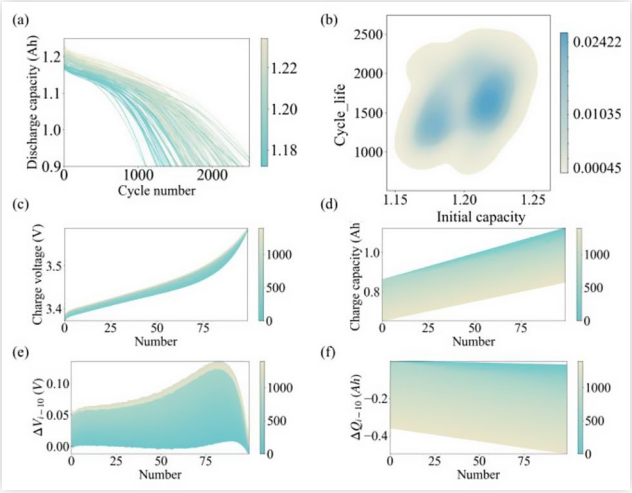
that compresses model size and parameters via knowledge distillation, enabling millisecond-level real-time SOH estimation. In addition, to ensure long-term stability and environmental adaptability, the E-CCT framework introduces a periodic cloud update and over-the-air (OTA) deployment mechanism that can be automatically triggered when model accuracy degrades or operating conditions drift, enabling dynamic model updates and cross-platform distribution. This design provides a unified, scalable, and deployment-flexible intelligent solution for health assessment and predictive maintenance of large-scale EV fleets.

The remainder of this paper is organized as follows: Section 2 introduces the battery dataset used in this study, including data sources, sampling features, and preprocessing methods. Section 3 details the overall design, core methodology, and training strategy of the proposed E-CCT framework. Section 4 presents and analyzes the experimental results, evaluating the real-time performance of the edge model in SOH estimation and the accuracy of the cloud model in RUL prediction, with comparisons against mainstream methods. Finally, Section 5 summarizes the key contributions of this work.

Dataset Description

The dataset used in this study is a publicly available battery degradation dataset released by Huazhong University of Science and Technology [22], which contains data from 77 cylindrical lithium-iron-phosphate (LFP)/graphite cells. All experiments were conducted in a temperature-controlled environment maintained at 30 °C. The nominal capacity and voltage of each cell are 1.1 Ah and 3.3 V, respectively. During testing, a unified fast-charging protocol was adopted. At the same time, the discharge process was performed under 77 distinct multi-stage constant-current (CC) schemes, thereby achieving diverse degradation trajectories under identical charging conditions. Experimental results show significant variation in cycle life among the cells, ranging from 1,100 to 2,700 cycles, reflecting substantial dispersion in electrochemical degradation behavior. The experimental platform indexed the batteries by groups, with each group containing eight channels, and a total of ten groups were tested. Therefore, the battery IDs follow the format 'group–channel' (e.g., 1-1). The charging process consists of three stages: first, a CC charge at 5C is applied to raise the battery from 0% to 80% state of charge (SOC); next, charging continues at 1C constant current until the terminal voltage reaches 3.6 V; finally, a constant-voltage (CV) charge is performed until the SOC reaches 100%, with the cut-off current set to C/20. The discharge process comprised four constant-current stages at 5C, 4C, 3C, and 1C, successively discharging the cell from full charge to 60%, 40%, 20%, and 0% SOC, with a discharge cutoff voltage of 2 V. In this experiment, 1C corresponds to a current of 1.1 Ampere (A). A 30-second rest period was set between each charge and discharge stage to simulate intermittent real-world

FIGURE 2 Trajectories and distributions of key features for battery degradation.



operating conditions and ensure electrochemical stability. For each cycle, key physical quantities, such as voltage, current, and capacity, were continuously recorded over time. End of Life (EOL) was defined as the point at which the cell's available capacity first declined to 80% of its initial rated capacity, representing the cell's lifetime threshold.

To ensure sample independence and improve the statistical robustness of model training, the dataset was split into training, testing, and validation sets at 7:2:1, with 53, 17, and 7 cells, respectively. Figure 2(a) illustrates the variation of discharge capacity with cycle number for all cells, where the color of each curve is scaled by cycle life (darker colors indicate longer lifetimes). Figure 2(b) shows the joint distribution of cycle life and initial capacity across all samples, both of which approximately follow wide normal-like distributions. This demonstrates that the dataset encompasses diverse degradation behaviors and sufficient sample variability, providing a solid foundation for validating the generalization performance of subsequent models.

Furthermore, to further verify the generalization capability of the proposed model across different chemistries, this study also employed the publicly available NCM battery dataset released by Zhu et al. [13]. The dataset contains 55 cells, with cycling data for each cell provided in CSV files named CYX#Y, where X denotes the temperature and Y the battery number. All experiments were conducted under strictly controlled isothermal conditions at 25°C, 35°C, and 45°C. Each battery cycle included charging, discharging, and resting processes. The charging process followed a CC–CV strategy: charging at a constant current to 4.2 V, then maintaining constant voltage until the cutoff current reached 0.05 C; this was followed by a 30-minute rest period (with an actual sampling interval of 120 s); finally, the battery was discharged at a constant current until the cutoff voltage of 2.5 V. The nominal capacity of the cells was 3.5 Ah, and the current rates were calculated based on this capacity;

i.e., 1C corresponds to 3.5 A. Each cycle fully recorded the time-series curves of key physical quantities, including voltage, current, and capacity. The EOL of a cell was defined as the point at which its maximum available capacity first fell to 71% of its initial rated capacity. Experimental results showed that at 25°C, the NCM cells reached 71% capacity after approximately 250–500 cycles; at 35°C, after 1250–1500 cycles; and at 45°C, after around 1000 cycles.

During data preprocessing, CY25#14 was identified as an abnormal sample due to its unusually low effective cycles and was therefore removed. The remaining 54 cells were split into training, testing, and validation sets at 8:1:1, with 37, 6, and 5 cells, respectively.

Methodology

Feature Engineering

In this study, four representative feature curves were extracted from the battery charging process: the charge voltage curve V (Figure 2 (c)), the charge capacity curve Q (Figure 2 (d)), the voltage difference curve between each cycle and the 10th cycle $\Delta V_{i-10} = V_i - V_{10}$ (Figure 2 (e)), and the capacity difference curve between each cycle and the 10th cycle $\Delta Q_{i-10} = Q_i - Q_{10}$ (Figure 2 (f)). Each feature curve was resampled to 100 points along the charging time axis using linear interpolation to ensure consistent data dimensionality and facilitate subsequent model processing (i.e., the number of horizontal axis points in Figures 2c–f). These four feature curves comprehensively capture the battery's behavioral characteristics during charging and the evolution of its electrochemical properties throughout the degradation process. Specifically, ΔV_{i-10} and ΔQ_{i-10} reflect the variation in charge voltage and charge capacity relative to the 10th cycle, respectively, characterizing the gradual capacity fade as cycling progresses. The 10th cycle was chosen as the reference to avoid the initial stage with minimal capacity change, thereby reducing noise interference and providing a more representative and informative signal.

For model input design, each cycle corresponds to four feature curves, which collectively form the model's input tensor. This input tensor consists of one sample, four feature channels, and 100 uniformly sampled points, representing the battery's dynamic behavior across different cycling stages. The model outputs include both SOH and RUL. The SOH is defined as the ratio of the discharge capacity at the current cycle to the nominal capacity of the battery, shown as Equation (1),

$$SOH_i = \frac{C_i}{C_{nom}} \tag{1}$$

where C_i is the discharge capacity at the i-th cycle, and C_{nom} is the nominal capacity of the battery. RUL is defined as the number of remaining cycles from the current cycle to the EOL, calculated by Equation (2),

$$RUL_i = N_{EOL} - N_i \tag{2}$$

where N_{EOL} denotes the cycle number when the SOH first decreases to 80%, and N_i represents the current cycle number.

Edge-Cloud Collaborative Modeling Framework

To achieve accurate and efficient battery SOH estimation and RUL prediction in heterogeneous deployment environments, this study proposes an edge–cloud collaborative modeling framework based on a Transformer architecture. The framework follows a task-decoupled knowledge distillation paradigm, with the main hyperparameter settings summarized in Table 1. In this design, the cloud supports large-scale sequence models trained on historical fleet data for complex representation learning, while the edge performs lightweight inference under latency and resource constraints. The proposed architecture adopts a dual-model design: a high-capacity teacher

TABLE 1 Main hyperparameter settings for the proposed framework.

Parameters	Value
Distillation temperature (T)	2.5
Teacher task weight (α)	0.7
Feature distillation weight (λ_f)	0.5
Output distillation weight (λ_o)	0.3
Learning rate	0.0001
Early stopping patience	7
Training epochs	50
Batch size	128
Loss function	MSE
Activation	GELU
Optimizer algorithm	Adam
Input sequence length	10
Start token length	10
Prediction sequence length	1
Encoder input size	4
Decoder input size	4
Output size	1
Cloud model dimension	256
Edge model dimension	64
Number of heads in the cloud model	8
Number of heads in the edge model	4
Encoder layer size of the cloud model	6
Encoder layers' size of the edge model	2
Decoder layer size of the cloud model	3
Decoder layers size of the edge model	1
Feed-forward network layer size of the cloud model	1024
Feed-forward network layer size of the cloud model	128
Dropout	0.1

model is trained in the cloud for long-term RUL prediction, while a lightweight student model is deployed on the vehicle edge for real-time SOH estimation. This separation between training and inference enables the system to maintain high prediction accuracy while ensuring scalability during deployment.

Both the cloud and edge models share a Transformer-based architecture specifically designed to process battery time-series data and perform regression prediction tasks. The model first maps the input sequences into a high-dimensional feature space via an embedding layer, and then generates learnable feature embeddings using positional and temporal encodings, enabling it to effectively capture temporal trends. The input features are then processed through a multi-head self-attention encoder that captures long-term degradation dependencies of the battery, formulated as Equation (3),

$$Attention\left(Q,K,V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Where Q, K, and V denote the query, key, and value matrices, respectively, and d_k is the key dimension used for scaling to prevent gradient explosion. The multi-head attention mechanism processes the inputs in parallel through multiple attention heads. It concatenates the results, enabling the model to jointly attend to information from different representational subspaces, as defined by Equations (4) and (5),

$$\text{MultiHead}\left(Q,K,V\right) = \text{Concat}\left(\text{head}_1,\ldots,\text{head}_h\right)W_O \tag{4}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{5}$$

where W_i^Q, W_i^K and W_i^V are learnable parameter matrices. The nonlinear feature representations extracted by the multi-head attention are then passed through a feed-forward network (FFN) consisting of two linear transformations and a GELU activation, as shown in Equation (6),

$$\text{FFN}\left(x\right) = \text{GELU}\left(xW_1 + b_1\right)W_2 + b_2 \tag{6}$$

where W_1 and W_2 are weight matrices and b_1, b_2 are bias terms. The resulting temporal representations are aggregated via global average pooling and passed to a regression head to generate RUL predictions. The model is trained in a fully supervised manner using historical battery degradation trajectories and stored in the cloud, serving as a knowledge base for extraction during the knowledge distillation process.

To enable real-time SOH estimation on edge BMS devices, a lightweight edge model was designed. While retaining the same overall architecture as the cloud model, the edge model significantly reduces the number of parameters by adopting a smaller embedding dimension, shallower encoder depth, and narrower feed-forward layers. Despite its simplified architecture, the edge model achieves robust performance through multi-level knowledge transfer from the cloud model. By decoupling model capacity from deployment constraints and systematically transferring hierarchical knowledge from the cloud predictor to a resource-efficient student model, the proposed framework establishes a scalable and adaptive approach for battery health monitoring. It provides a practical solution for deploying predictive maintenance algorithms across heterogeneous hardware platforms while maintaining consistency with the continuously updated, centrally learned knowledge in the cloud.

The Training Strategy of the Proposed Framework

To compensate for the accuracy loss caused by the lightweight edge model, the training process employs three forms of knowledge distillation: prediction-level, feature-level, and task-alignment distillation. Specifically, the student model's outputs are first aligned with the teacher model's softened outputs using temperature-scaled Kullback–Leibler (KL) divergence. This collaborative training mechanism is jointly optimized through a composite objective function, whose adjustable coefficients balance the contributions of supervised learning and knowledge distillation.

During the distillation stage, the teacher model is trained under supervision with RUL as the target. In contrast, the student model focuses on SOH estimation and learns the teacher's latent feature representations and predictive distributions via feature and output distillation. The total loss function used in training combines the teacher and student task losses, feature distillation loss, and output distillation loss, which is shown by Equation (7),

$$L_{total} = \alpha L_t^t + \left(1-\alpha\right)L_t^s + \lambda_f L_f + \lambda_o L_o \tag{7}$$

where α denotes the teacher task weight controlling the contribution of the teacher task loss; L_t^s represents the student task loss, reflecting the error of the student model on SOH estimation; λ_f is the feature distillation weight controlling the strength of feature alignment; and λ_o is the output distillation weight adjusting the influence of output-level knowledge transfer.

To ensure the accuracy of SOH estimation, the student model is anchored to the true SOH labels via a mean squared error (MSE) loss, defined by Equation (8),

$$L_t^s = MSE\left(\hat{y}_s, y_s\right) \tag{8}$$

where $\hat{y}_s$ is the predicted value and y_s is the ground truth. At the feature distillation level, alignment is achieved by minimizing the MSE between the student and teacher feature representations, enabling the student to capture high-order representations similar to the teacher despite

its constrained architecture. The feature distillation loss is expressed by Equation (9),

$$L_f = MSE\left(F_s, F_t\right) \tag{9}$$

At the output level, to enable the student model to better emulate the distributional characteristics of the teacher model's predictions, an output distillation loss based on KL divergence is introduced, defined as Equation (10),

$$L_o = T^2 \cdot KL\left(Softmax\left(\frac{Z_s}{T}\right) \| Softmax\left(\frac{Z_t}{T}\right)\right) \tag{10}$$

where T is the distillation temperature, which smooths the probability distribution to help the student learn the uncertainty and inter-class relationships encoded in the teacher's outputs. Z_s and Z_t represent the outputs of the student and teacher models, respectively. After model training converges, only the student model is deployed on the edge for inference, while the teacher model remains in the cloud for centralized training and periodic distillation updates. This asymmetric deployment allows the student model to benefit from continuous cloud learning without incurring the full computational cost of the teacher model. Moreover, the framework supports dynamic updating, enabling the edge model to be re-distilled under evolving operating conditions, thereby adapting to long-term domain shifts.

In summary, this collaborative framework offers an efficient and accurate method for assessing battery health in real-world applications. By principled knowledge distillation and decoupling the architecture across cloud and edge, it effectively balances model complexity and inference efficiency.

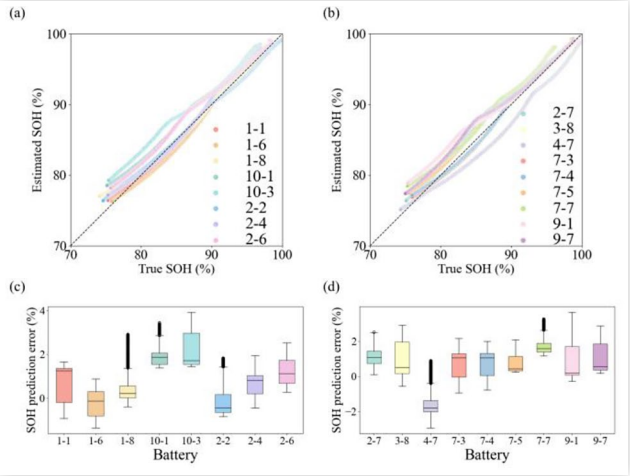
Results and Discussion

The SOH Estimation Results of the Edge Model

To validate the effectiveness of the designed lightweight edge model (student model) in the SOH estimation task, its performance on the test set was analyzed, with the results shown in Figure 3(a) and 3(b) visually illustrating the degree of fit between the predicted SOH values and the ground truth. It is clear that the predicted trajectories for all test cells closely follow the ideal diagonal line. This indicates that the edge model not only provides accurate point estimates but also tracks the SOH degradation trajectory with high fidelity throughout the entire battery lifecycle.

The error boxplots in Figures 3(c) and 3(d) further demonstrate the robustness of the edge model in SOH estimation. Analysis shows that the median error across all test cells remains low (approximately 10–30 mAh), corresponding to a median relative error below 2.7%.

FIGURE 3 Edge model SOH estimation performance. (a) and (b) present scatter plots comparing the predicted SOH values with the true values of different battery cells. (c) and (d) represent the estimated errors of SOH for the corresponding test batteries.



Meanwhile, the interquartile ranges are generally narrow, indicating high stability and consistency in the model predictions. Although a few outliers exist in certain samples, such as cells 1-8 and 7-7, the vast majority of prediction errors are strictly controlled within 40 mAh, with relative errors not exceeding 3.7%. In summary, despite its lightweight configuration, the student model achieves high accuracy and robustness in SOH estimation, as evidenced by strong agreement between predicted and true values and low magnitudes and variances of prediction errors. These results demonstrate that the knowledge-distilled, lightweight model has great potential for reliable deployment on resource-constrained edge devices.

To quantitatively evaluate the performance of the proposed model, three metrics were employed: Root mean squared error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2). RMSE measures the average deviation of model predictions from the ground truth, with values closer to zero indicating higher prediction accuracy, and is defined as Equation (11)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{11}$$

where $\hat{y}_i$ and y_i are the predicted and true values, respectively. MAPE quantifies the relative error between predictions and true labels, with smaller values indicating better predictive performance, and is defined as Equation (12),

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{12}$$

R^2 assesses the correlation between the predicted and observed values, providing a clear indication of the model's goodness-of-fit. Higher R^2 values, approaching 1, indicate better model fit, and can be expressed as Equation (13),

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}_i\right)^2} \tag{13}$$

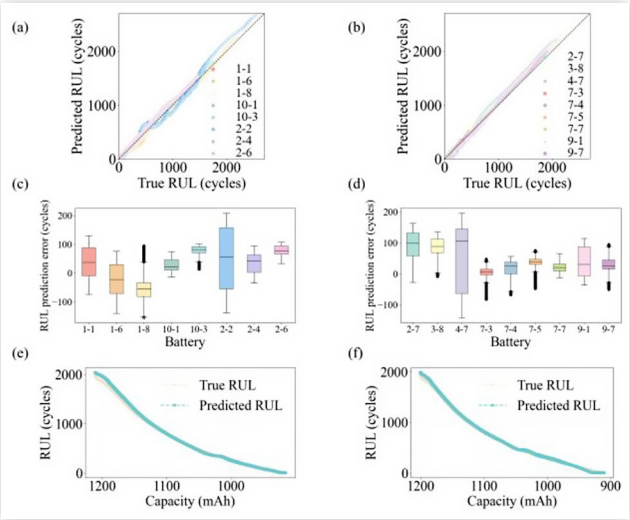
where $\overline{y}_i$ is the mean of the true values.

The RUL Prediction Results of the Cloud Model

To evaluate the performance of the proposed cloud model in long-term degradation prediction, a comprehensive assessment was conducted on the test set, and the model's RUL predictions are shown in Figure 4. Although RUL prediction is more challenging than SOH estimation, the model achieved an RMSE of 2.59% (78 cycles), a MAPE of 12.04%, and an R^2 of 0.98 across the entire test set, indicating highly consistent and accurate predictions throughout the battery lifecycle. Figure 4(a) and 4(b) illustrate the fit between predicted and true RUL values, confirming that the model accurately reconstructs the long-term degradation trajectories of multiple cells exhibiting different aging patterns. This near-linear correspondence demonstrates that the cloud model effectively captures the inherent nonlinear degradation trends of lithium-ion batteries while maintaining stable generalization across diverse degradation profiles. Figure 4 (c) and 4 (d) provide a quantitative analysis of the RUL prediction errors using boxplots. The results show that the median absolute errors for most cells remain close to zero, while a few cells exhibit higher dispersion, possibly due to local anomalies or measurement noise.

Figure 4(e) and 4(f) present the true degradation trajectories and predicted trajectories for two representative cells. The predicted curves nearly perfectly overlap with the measured data, indicating that the model not only provides accurate pointwise predictions but also faithfully captures the temporal evolution of the degradation trend. This trajectory-level fidelity highlights the model's ability to capture cross-scale relationships between electrochemical degradation and capacity fade, reflecting a physically consistent understanding of the underlying mechanisms. In summary, the high accuracy of RUL predictions, low error distribution, and precise reproduction of degradation trajectories demonstrate that the cloud model effectively learns the complex internal mechanisms of battery degradation. The high-fidelity predictions provide a reliable foundation for knowledge distillation, enabling the transfer of generalized degradation representations from the cloud-based teacher model to the lightweight student model deployed at the edge.

FIGURE 4 Cloud model RUL prediction performance. (a) and (b) show the comparison between the predicted RUL values and the true values of different battery cells. (c) and (d) represent the RUL prediction errors of the corresponding test cells. (e) and (f) present the predicted RUL trajectories and the actual decline trajectories of two representative cells.

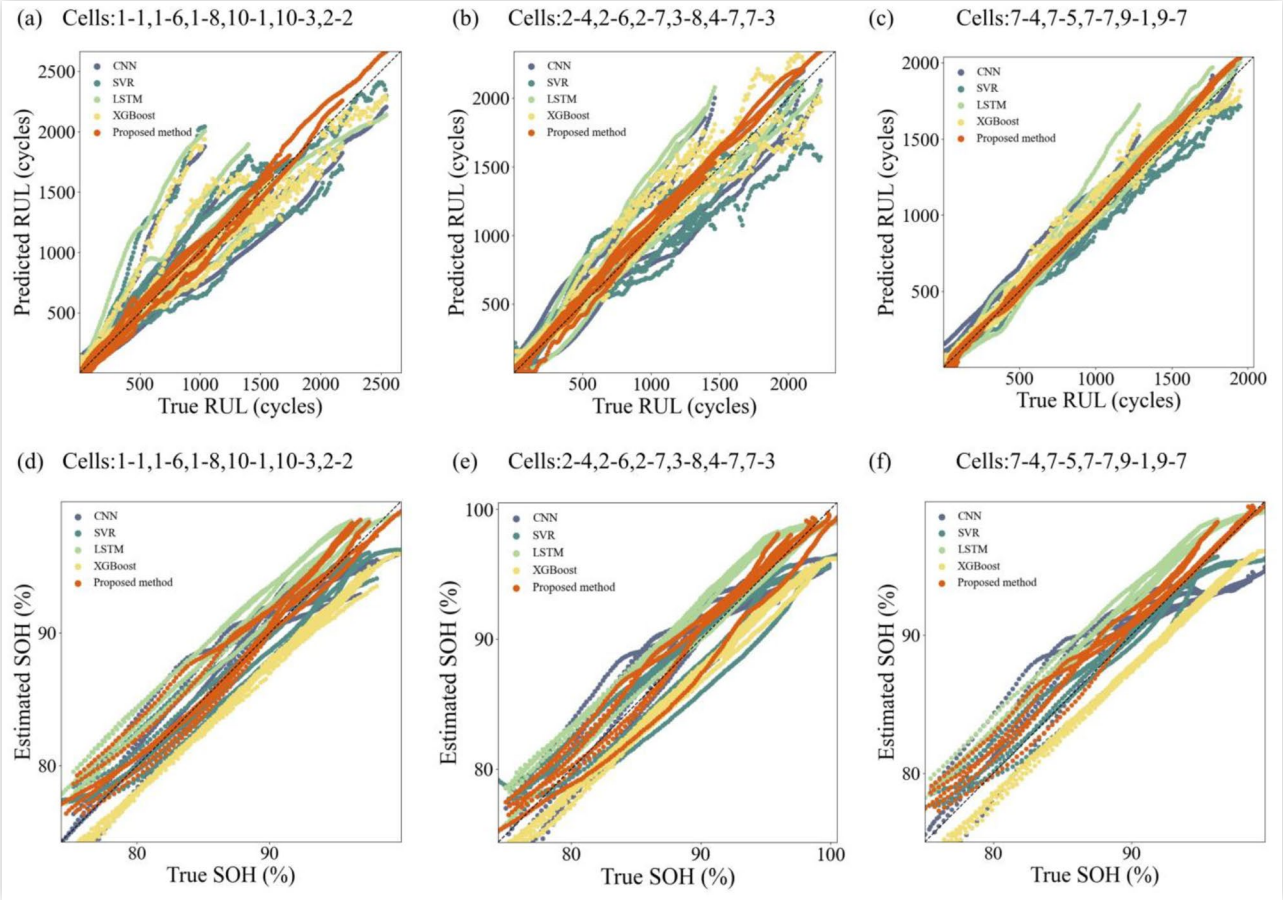


Comparison with Other Methods

Comparison with Representative Baseline Models To further validate the effectiveness and generalization capability of the proposed edge–cloud collaborative framework, a series of comparative experiments was conducted against representative baseline models, including CNN, LSTM, Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost). These models were selected for their established applications in battery prediction and for their use of diverse modeling paradigms, providing a comprehensive benchmark for both deep learning and traditional machine learning approaches. All models were implemented using the PyTorch deep learning framework (Python 3.12) and tested on a hardware platform equipped with an Intel Core i5-14600KF processor, 16 GB of RAM, and an NVIDIA GeForce RTX 5060 Ti GPU with 8 GB of dedicated memory. The comparative experiments covered both RUL prediction and SOH estimation tasks.

Figures 5(a)–(c) present the comparison between predicted and true RUL values on the test set. The baseline models exhibited varying degrees of bias and dispersion in their predictions. Specifically, CNN and LSTM tended to underestimate RUL during mid-to-late cycles, while XGBoost and SVR often produced unstable predictions beyond 1,500 cycles. In contrast, the proposed E-CCT framework fits the long-term degradation trajectories, demonstrating excellent generalization across both early and late degradation stages. Figures 5(d)–(f) further compare the SOH estimation performance across different models. Baseline models suffered from saturation effects and high variability, particularly when SOH

FIGURE 5 The proposed method is compared with the baseline model's predictions on the test set. (a)-(c) represent the RUL prediction results of different models. (d)-(f) represent the SOH estimation results of different models.



exceeded 90%, making accurate differentiation challenging. CNN and LSTM exhibited nonlinear distortions, whereas SVR and XGBoost tended to overfit local trends, resulting in suboptimal performance. By comparison, the E-CCT framework accurately estimated SOH with low bias and low variance across cells with diverse degradation behaviors.

To quantitatively evaluate performance differences among models, Table 2 presents analyses based on three evaluation metrics. As shown in Table 2, the E-CCT framework achieved the lowest RMSE in both tasks, with an SOH estimation RMSE of 1.41% and an RUL prediction RMSE of 2.59% (78 cycles). In contrast, SVR yielded RMSE values of 1.81% and 6.78% (204 cycles) for SOH and RUL, respectively; XGBoost achieved 2.58% and 5.94% (178 cycles); CNN produced 2.19% and 6.55% (197 cycles); LSTM resulted in 2.59% for SOH and the highest RUL RMSE of 7.22% (217 cycles). These results indicate that the E-CCT framework significantly reduces prediction bias across both short- and long-term scenarios, outperforming conventional machine learning and neural network baselines. Similarly, the MAPE results in Table 2 show that the E-CCT framework achieved relative errors of only 1.27% and 12.05% for SOH and RUL estimation, respectively, representing a substantial reduction compared to all baselines. Additionally, Table 2 demonstrates that the E-CCT framework achieved the highest R^2 values, with 0.95 for SOH and 0.98 for RUL, indicating excellent goodness-of-fit. Collectively, these results confirm the advantages of the E-CCT framework in capturing the complex nonlinear dynamics of battery degradation,

TABLE 2 Performance comparison between the proposed method and the baseline model on the test set.

	RUL			SOH		
Method	RMSE (%/cycle)	MAPE (%)	R^2	RMSE (%)	MAPE (%)	R^2
LSTM	7.22/217	29.94	0.86	2.59	2.64	0.84
CNN	6.55/197	41.51	0.88	2.19	2.01	0.88
XGBoost	5.94/178	35.25	0.90	2.58	2.73	0.84
SVR	6.78/204	30.72	0.88	1.81	1.56	0.92
Proposed method	2.59/78	12.05	0.98	1.41	1.27	0.95

TABLE 3 Comparison of model performance and computational cost.

Method	RMSE (%)	GPU memory (MB)	Parameters	Inference latency (ms)	FLOPS
Cloud model (RUL)	2.59	65.05	5531649	4.1732	48146944
Edge model (SOH)	1.41	21.4	84737	1.9299	694848
Baseline model (SOH)	6.39	21.4	84737	1.9299	694848

enabling high-precision SOH estimation and long-term RUL prediction. By combining Transformer-based edge–cloud collaborative modeling with knowledge distillation, the framework offers a highly generalizable and deployable solution for intelligent battery health management.

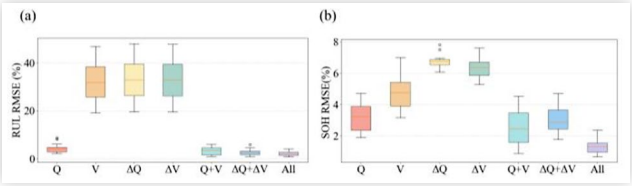
Knowledge Distillation Effectiveness Analysis To validate the effectiveness of the proposed knowledge distillation framework in terms of model lightweighting and knowledge transfer, the student model within the framework was compared with a baseline model (Transformer) that uses the same hyperparameters but was trained without distillation. The results are shown in Table 3. For a fair comparison, both the edge model and the baseline model use the same network architecture. Specifically, both models have 84,737 parameters, consume 21.4 MB of GPU memory, exhibit an inference latency of 1.9299 ms, and require 694,848 FLOPS. The only variable is whether knowledge distillation was employed during training. The performance difference is striking. The edge model trained with knowledge distillation achieves an RMSE of 1.41% for SOH estimation, whereas the undistilled baseline model has an RMSE of 6.39%. This indicates that, without introducing any additional computational overhead, the knowledge distillation mechanism reduces the relative error by approximately 77.9%, strongly demonstrating that the student model effectively absorbs generalized knowledge from the high-capacity cloud model, significantly improving prediction accuracy.

Furthermore, Table 3 quantifies the framework's effectiveness in lightweighting models. Compared to the cloud model, which has 5.53 million parameters and 48.1 MFLOPS, the edge model reduces the parameter count by approximately 65.3 and the computational load by roughly 69.3. In summary, the proposed framework successfully transfers the knowledge of a high-accuracy, complex model to a computationally efficient lightweight model, achieving a balanced trade-off between accuracy and efficiency.

Feature Sensitivity Analysis

To evaluate the contribution of each input feature to the model's performance, a sensitivity analysis was conducted on different feature combinations. The specific experimental settings were as follows: using only the charging capacity curve Q, only the charging voltage curve V, and only the capacity difference curve ΔQ_{i-10}, only the voltage difference curve ΔV_{i-10}, the combination of Q+V, the combination of $\Delta Q_{i-10} + \Delta V_{i-10}$, and the full feature set (All) employed in this study. All experiments were performed under the same dataset split and training strategy, and the corresponding RMSE metrics for RUL and SOH were calculated. Figures 6 (a) and 6 (b) show the RMSE boxplots for RUL prediction and SOH estimation using different feature combinations, respectively.

FIGURE 6 Feature sensitivity analysis. (a) represents the RMSE box plots of RUL prediction for different feature combinations, and (b) represents the RMSE box plots of SOH estimation for different feature combinations.

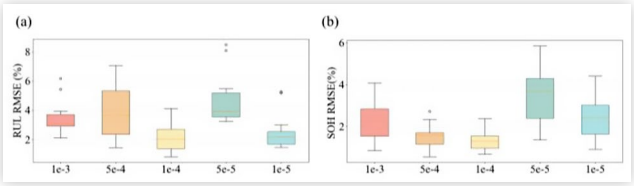


It can be observed that the two combined feature sets achieve relatively good performance in both RUL prediction and SOH estimation. This is because Q+V provides stable macroscopic information by complementing capacity and voltage, whereas $\Delta Q + \Delta V$ captures microscopic local variations. The All combination achieves the lowest median errors and the narrowest interquartile ranges for both RUL and SOH, indicating that the model can simultaneously leverage macroscopic degradation trends and microscopic variation information to make accurate and stable predictions across all test cells. Overall, single-feature sets perform worse, with higher medians and longer boxes, indicating that a single feature cannot fully represent the battery degradation information. Notably, the single feature Q, which inherently contains the primary information about battery cycle life, performs closely to the All combination in RUL prediction; however, to achieve the highest accuracy, the full feature set was still used as the model input in this study.

Hyperparameter Sensitivity Analysis

To evaluate the stability of the proposed model with respect to hyperparameter settings, a sensitivity analysis on the learning rate was conducted. Five different orders of magnitude of learning rates were tested: 1e-3, 5e-4, 1e-4, 5e-5, 1e-5. All experiments employed the same dataset split strategy and other hyperparameter settings, and the corresponding RMSE metrics for RUL and SOH

FIGURE 7 Hyperparameter (learning rate) sensitivity analysis. (a) represents the RMSE box plots of RUL prediction for different learning rates, and (b) represents the RMSE box plots of SOH estimation for different learning rates.



were calculated. Figures 7 (a) and 7 (b) show the RMSE boxplots for RUL prediction and SOH estimation under different learning rates, respectively.

It can be observed that for RUL prediction, the median RMSE is lowest at a learning rate of 1e-4, with a moderately sized box, indicating the highest and most stable prediction accuracy. In contrast, at 5e-4, the median is higher, and the box is longer, possibly due to oscillations during training that cause larger error fluctuations. Overall, the model's performance in RUL prediction is relatively stable across different learning rates, demonstrating robustness for this task.

For SOH estimation, 1e-4 also achieves the lowest median and a relatively narrow box. As shown in Figure 7 (b), larger learning rates (e.g., 1e-3, 5e-4) are more effective than smaller ones (e.g., 5e-5, 1e-5) at capturing the macroscopic trend of capacity degradation, enabling the model to converge quickly and yield stable predictions. Conversely, overly small learning rates lead to slow parameter updates, preventing the model from reaching an optimal state within a limited number of training epochs, resulting in higher errors and greater fluctuation.

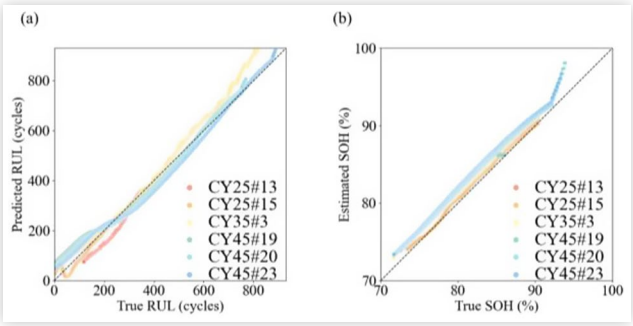
In summary, a learning rate of 1e-4 provides the best performance for both RUL and SOH prediction; therefore, it was selected as the final training learning rate in this study.

Evaluation on an Additional Chemistry Dataset

To further validate the generalization capability of the proposed model, an additional evaluation was conducted using the NCM battery dataset described in the Dataset description, which represents a different chemistry system.

Figures 8 (a) and 8 (b) show the regression results of RUL and SOH predictions on this NCM dataset. It can be observed that, despite the significant differences in chemistry compared with the original LFP dataset, the proposed knowledge distillation framework still achieves good prediction accuracy on this dataset, with only limited deviations occurring at a few points during the early life or near the end-of-life threshold. These minor deviations mainly arise from the higher uncertainty in NCM battery degradation behavior during these two stages: in the early stage, SOH fluctuations are small and signal variations are limited, making measurement noise more pronounced; near EOL, NCM batteries often exhibit accelerated degradation driven by structural instability, making the degradation rate harder to capture accurately with limited samples. Therefore, these local deviations are considered normal characteristics of the degradation behavior of this chemistry system and do not affect the model's ability to reliably track the overall life trend.

FIGURE 8 Evaluation of the proposed model on the NCM dataset. (a) and (b) present scatter plots comparing the predicted values with the true values of different battery cells: (a) for RUL prediction, and (b) for SOH estimation.



Limitations and Outlook

Although the proposed edge–cloud collaborative modeling framework demonstrates excellent performance on laboratory data, it still has certain limitations. First, temperature is a key factor affecting lithium-ion battery degradation and RUL, whereas the datasets used in this study were all collected under controlled laboratory temperatures and thus do not account for temperature fluctuations, extreme environments, or high/low-temperature charging conditions encountered in real-world vehicle operation. Second, although additional evaluations were conducted on different chemistry systems, all experiments were still performed under controlled laboratory conditions using a uniform fast-charging protocol. They did not include partial charge–discharge cycles, multi-rate charging, or load fluctuations caused by various driving patterns, which are typical in real-world scenarios. Therefore, the model's robustness under complex environmental conditions requires further validation.

To further enhance the practical applicability of the model, future work will focus on the following aspects: (1) incorporating real-world vehicle data across varying temperatures, multiple charge–discharge rates, different aging pathways, and additional chemistry systems to systematically evaluate the model's cross-domain generalization capability; (2) exploring knowledge transfer and domain adaptation strategies across battery platforms to further improve the model's adaptability to different battery form factors and usage scenarios.

Conclusions

This study proposes an edge–cloud collaborative modeling framework for battery SOH estimation and RUL prediction, integrating a lightweight student model deployed at the edge with a high-capacity teacher model residing in the cloud. Through knowledge distillation and transfer learning, the framework enables accurate, low-latency SOH estimation while significantly reducing the computational burden at the edge. The proposed framework was comprehensively evaluated on a dataset of 77 lithium-ion battery degradation profiles. Results show that the edge model achieves an SOH estimation RMSE of 1.41% and an RUL prediction RMSE of 2.59% (78 cycles), approaching the accuracy of the cloud-based teacher model while maintaining low inference latency and a compact model size. Moreover, comparative experiments against representative baseline models demonstrate that the proposed method achieves the highest accuracy across all metrics. Beyond performance, the proposed edge–cloud paradigm provides enhanced scalability, data privacy, and adaptability. By performing lightweight inference at the edge and periodically synchronizing knowledge with the cloud, the framework strikes a balance between predictive accuracy and resource efficiency. This approach offers significant potential for intelligent and distributed battery management in next-generation energy systems. Future work will explore dynamic update strategies in online learning environments, including continual adaptation to evolving degradation patterns and environmental conditions. In addition, extending the framework to integrate multi-modal sensor fusion and cross-domain transfer will further enhance its robustness and generalizability across different battery chemistries, form factors, and usage scenarios.

References

1. Hu, X., Xu, L., Lin, X., and Pecht, M., "Battery Lifetime Prognostics," *Joule* 4, no. 2 (2020): 310-346, doi:10.1016/j.joule.2019.11.018.
2. Ma, B., Zhang, L., Yu, H., Zou, B. et al., "End-Cloud Collaboration Method Enables Accurate State of Health and Remaining Useful Life Online Estimation in Lithium-Ion Batteries," *Journal of Energy Chemistry* 82 (2023): 1-17, doi:10.1016/j.jechem.2023.02.052.
3. Kuzhiyil, J.A., Damoulas, T., Planella, F.B., and Widanage, W.D., "Lithium-Ion Battery Degradation Modelling Using Universal Differential Equations: Development of a Cost-Effective Parameterisation Methodology," *Applied Energy* 382 (2025): 125221, doi:10.1016/j.apenergy.2024.125221.
4. Planella, F.B. and Widanage, W.D., "A Single Particle Model with Electrolyte and Side Reactions for Degradation of Lithium-Ion Batteries," *Applied Mathematical Modelling* 121 (2023): 586-610, doi:10.1016/j.apm.2022.12.009.
5. Liu, K., Gao, Y., Zhu, C., Li, K. et al., "Electrochemical Modeling and Parameterization Towards Control-Oriented Management of Lithium-Ion Batteries," *Control Engineering Practice* 124 (2022): 105176, doi:10.1016/j.conengprac.2022.105176.
6. He, W., Williard, N., Osterman, M., and Pecht, M., "Prognostics of Lithium-Ion Batteries Based on Dempster–Shafer Theory and the Bayesian Monte Carlo Method," *Journal of Power Sources* 196, no. 23 (2011): 10314-10321, doi:10.1016/j.jpowsour.2011.08.040.
7. Zhao, C., Andersen, P.B., Træholt, C., and Hashemi, S., "Data-Driven Battery Health Prognosis with Partial-Discharge Information," *Journal of Energy Storage* 65 (2023): 107151, doi:10.1016/j.est.2023.107151.
8. Meng, J., Stroe, D.I., Ricco, M., Luo, G. et al., "A Simplified Model Based State-of-Charge Estimation Approach for Lithium-Ion Battery with Dynamic Linear Model," *IEEE Transactions on Industrial Electronics* 66, no. 10 (2018): 7717-7727, doi:10.1109/TIE.2018.2880668.
9. Che, Y., Deng, Z., Tang, X., Lin, X. et al., "Lifetime and Aging Degradation Prognostics for Lithium-Ion Battery Packs Based on a Cell to Pack Method," *Chinese Journal of Mechanical Engineering* 35, no. 1 (2022): 4, doi:10.1186/s10033-021-00668-y.
10. Severson, K.A., Attia, P.M., Jin, N., Perkins, N. et al., "Data-Driven Prediction of Battery Cycle Life Before Capacity Degradation," *Nature Energy* 4, no. 5 (2019): 383-391.
11. Wang, F., Zhai, Z., Zhao, Z., Di, Y. et al., "Physics-Informed Neural Network for Lithium-Ion Battery Degradation Stable Modeling and Prognosis," *Nature Communications* 15, no. 1 (2024): 4332.
12. Zhang, H., Li, Y., Zheng, S., Lu, Z. et al., "Battery Lifetime Prediction Across Diverse Ageing Conditions with Inter-Cell Deep Learning," *Nature Machine Intelligence* 7, no. 2 (2025): 270-277.
13. Zhu, J., Wang, Y., Huang, Y., Bhushan Gopaluni, R. et al., "Data-Driven Capacity Estimation of Commercial Lithium-Ion Batteries from Voltage Relaxation," *Nature communications* 13, no. 1 (2022): 2261.
14. Lu, J., Xiong, R., Tian, J., Wang, C. et al., "Deep Learning to Estimate Lithium-Ion Battery State of Health Without Additional Degradation Experiments," *Nature Communications* 14, no. 1 (2023): 2760, doi:10.1038/s41467-023-38458-w.
15. Lee, G., Kwon, D., and Lee, C., "A Convolutional Neural Network Model for SOH Estimation of Li-Ion Batteries with Physical Interpretability," *Mechanical Systems and Signal Processing* 188 (2023): 110004, doi:10.1016/j.ymssp.2022.110004.
16. Chaoui, H. and Ibe-Ekeocha, C.C., "State of Charge and State of Health Estimation for Lithium Batteries Using Recurrent Neural Networks," *IEEE Transactions on Vehicular Technology* 66, no. 10 (2017): 8773-8783, doi:10.1109/TVT.2017.2715333.
17. Wang, Q., Wang, Z., Liu, P., Zhang, L. et al., "Large-Scale Field Data-Based Battery Aging Prediction Driven by Statistical Features and Machine Learning," *Cell Reports*

Physical Science 4, no. 12 (2023): 101720, doi:10.1016/j.xcrp.2023.101720.

18. Liu, H., Li, C., Hu, X., Li, J. et al., "Multi-Modal Framework for Battery State of Health Evaluation Using Open-Source Electric Vehicle Data," *Nature Communications* 16, no. 1 (2025): 1137, doi:10.1038/s41467-025-56485-7.
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J. et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems* 30, no. 1 (2017): 5999-6009.
20. Li, Y., Wang, H., Wang, C., Wang, L. et al., "Unified Physics-Informed Subspace Identification and Transformer Learning for Lithium-Ion Battery State-of-Health Estimation," *Journal of Energy Chemistry* 112 (2026): 350-369, doi:10.1016/j.jechem.2025.08.060.
21. Chen, L., Xie, S., Lopes, A.M., and Bao, X., "A Vision Transformer-Based Deep Neural Network for State of Health Estimation of Lithium-Ion Batteries," *International Journal of Electrical Power & Energy Systems* 152 (2023): 109233, doi:10.1016/j.ijepes.2023.109233.
22. Ma, G., Xu, S., Jang, B., Cheng, C. et al., "Real-Time Personalized Health Status Prediction of Lithium-Ion Batteries Using Deep Transfer Learning," *Energy & Environmental Science* 15, no. 10 (2022): 4083-4094, doi:10.17632/nsc7hnsg4s.2.

Contact Information

Shiqi (Shawn) Ou
sou@scut.edu.cn; oushiqi@pazhoulab.cn
Phone number: +86-020-81181684
Mailing address:
1. South China University of Technology, School of Future Technology, 777 Xingye Ave East, Panyu District, Guangzhou, Guangdong, 511442, China
2. Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), 70 Yuean Road, Haizhou District, Guangzhou, Guangdong 510335, China.n

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2024YFE0115800) and the Introduced Innovative R&D Team of Guangdong (2023ZT10L145). All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of sponsors.

ISSN 0148-7191