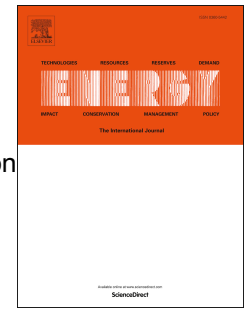
Journal Pre-proof



Self-Supervised Learning for Electric Vehicle Battery Remaining Useful Life Prediction Using Real-World Unlabeled Data

Zhilong Lv, Shiqi (Shawn) Ou, Hao Jing, Guoyuan Wu, Dapai Shi

PII: S0360-5442(26)01408-8

DOI: https://doi.org/10.1016/j.energy.2026.141302

Reference: EGY 141302

To appear in: *Energy*

Received Date: 18 March 2026

Revised Date: 22 April 2026

Accepted Date: 5 May 2026

Please cite this article as: Lv Z, Ou S(S), Jing H, Wu G, Shi D, Self-Supervised Learning for Electric Vehicle Battery Remaining Useful Life Prediction Using Real-World Unlabeled Data, *Energy*, https://doi.org/10.1016/j.energy.2026.141302.

Self-Supervised Learning for Electric Vehicle Battery Remaining Useful Life Prediction Using Real-World Unlabeled Data

Zhilong Lv[1], Shiqi (Shawn) Ou[1,2,*], Hao Jing[1,2], Guoyuan Wu[3], Dapai Shi[4]

[1] School of Future Technology, South China University of Technology, Guangzhou, Guangdong 511442, China

[2] Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), Guangzhou, Guangdong 510335, China

[3] Center for Environmental Research and Technology, Bourns College of Engineering, University of California at Riverside, Riverside, CA 92507, USA

[4] School of Mechanical Engineering, Hubei University of Arts and Science, Xiangyang 441053, China

* Corresponding author: sou@scut.edu.cn (S.O.)

Abstract

Accurate and cost-effective prediction of remaining useful life (RUL) is critical for reliable health management of electric vehicle (EV) batteries. However, most data-driven approaches for RUL prediction rely on fully supervised learning and extensive labeling, which is expensive and difficult to scale under heterogeneous operating conditions. This study proposes a contrastive-enhanced variational autoencoder–long short-term memory (VAE–LSTM) framework that leverages large-scale unlabeled charging data in a self-supervised learning paradigm. The framework is pretrained using joint reconstruction and contrastive objectives to learn monotonic degradation representations, and subsequently fine-tuned for RUL regression with limited labeled vehicles. The approach is evaluated on three real-world EV operational datasets, including one heterogeneous fleet dataset comprising passenger cars, taxis, and city buses, as well as two taxi fleets with different scales and operating characteristics. The proposed framework achieves a root-mean-square-error (RMSE) of 27 cycles, outperforming supervised and semi-supervised baselines. Label-efficiency and cross-fleet transfer studies further quantify robustness to domain shift and irregular sampling in field data. With labels from only 30% of vehicles, the pretrained model transfers to two target fleets with RMSE values of 43 and 50 cycles, respectively. A deployment-oriented cost analysis shows that the framework achieves an RMSE within 5% of the fully supervised model while reducing RUL labeling costs by about 70%. Latent factors learned during pretraining are correlated with physically meaningful voltage and energy-throughput signatures, improving interpretability. The proposed VAE–LSTM enables an accurate, interpretable, and economically scalable pathway for real-world EV battery RUL predictions.

Keywords: Electric vehicle; Real-world operational data; Lithium-ion battery; Remaining useful life; Self-supervised learning

1. Introduction

Driven by increasing concerns over global warming and reinforced by regulatory policies, electrified mobility is widely regarded as a key pathway toward carbon-neutral economic development. Lithium-ion batteries (LIBs) are extensively applied in various fields, particularly in electric vehicles (EVs), due to their high energy density and extended cycle life [1,2]. However, LIBs undergo inevitable capacity loss and power degradation during both operation and storage [3,4]. To address these challenges, accurate aging diagnosis of LIBs is regarded as a core function of the battery management system (BMS) in EV applications, as it directly supports optimized battery utilization, cost reduction, and safety enhancement [5,6]. Despite its importance, robust aging diagnosis under real-world on-road operation remains challenging, as degradation arises from coupled physical-electrochemical mechanisms and is strongly modulated by uncertain and heterogeneous usage behaviors [7-9]. Existing aging diagnosis approaches for LIBs are generally classified into three categories: model-based methods, data-driven methods, and hybrid approaches that integrate physical models with data-driven learning [10-17].

Data-driven methods have received growing attention for LIBs aging diagnosis as the volume of battery-related data continues to expand [18-20]. These methods establish quantitative relationships mapping measurable battery signals and temperature to the corresponding aging states of LIBs [21]. Convolutional neural network–based models have been proposed to extract degradation-related features from partial charging sequences in the time domain [22]. Recent studies have further improved modeling performance by incorporating advanced sequence learning and hybrid architectures. For example, long short-term memory (LSTM)-based models with enhanced feature-extraction mechanisms have been developed to capture temporal degradation patterns under partial-charging conditions [23]. Multi-task learning frameworks have also been explored to improve generalization across heterogeneous operating scenarios [24]. In addition, hybrid CNN–attention architectures have been proposed to jointly model local features and long-range dependencies in battery degradation signals [25]. In parallel, Gaussian process regression has been applied to incremental capacity curves to characterize aging behavior for battery capacity estimation [26]. Other studies have explored tree-based and kernel-based learning methods, including extreme gradient boosting and support vector regression to estimate capacity degradation [27,28]. Although these supervised data-driven methods demonstrate strong predictive capability, their performance generally depends on the

availability of large-scale labeled datasets for model training. However, acquiring labeled aging data through experimental testing is costly and time-consuming. For example, a widely used public degradation dataset was constructed using more than 100 commercial battery cells, where aging a single cell to the end of life required at least several months of continuous cycling [29]. Furthermore, degradation depends on interacting stressors-such as ambient temperature, charge/discharge rate, and depth of discharge-which can substantially erode supervised-model accuracy and robustness when labeled data do not sufficiently cover deployment conditions [30,31].

To alleviate the dependence of battery degradation diagnosis on large amounts of labeled data, transfer learning has been widely investigated at the cell level [32-34]. A study has demonstrated that degradation information learned from batteries with similar chemistries, but different capacities, can be transferred to enhance capacity estimation accuracy under diverse operating scenarios [35]. Although transfer learning reduces the demand for labeled data in the target domain, the source-domain data are still required to be collected under conditions that are comparable to real-world applications, which still limits its practical scalability [36]. Data augmentation has also been explored as an alternative strategy to mitigate the reliance of data-driven methods on labeled battery data [37]. This approach generates synthetic samples by learning the statistical characteristics of available labeled data using generative models, such as variational autoencoders (VAEs) [38] and generative adversarial networks [39]. In addition, physics-informed data generation methods have been proposed to incorporate domain knowledge of battery degradation into the synthesis process, enabling the construction of physically consistent degradation trajectories for model training [40]. Despite these efforts, the effectiveness of data augmentation is inherently constrained by the representativeness and diversity of the limited labeled data used for model training. More recently, semi-supervised learning has been well investigated to further reduce the dependence of battery degradation diagnosis and prognosis on labeled data by incorporating unlabeled samples into the learning process [41]. For example, confidence-weighted semi-supervised frameworks with label propagation have been proposed to generate pseudo-RUL labels for unlabeled samples and integrate them into supervised training [42]. Although improved predictive performance has been reported, the effectiveness of such methods largely depends on the quality and reliability of the generated pseudo-labels.

To narrow this lab-to-field gap, recent work has leveraged in-service EV operational data for aging diagnosis and remaining useful life (RUL) prediction. For example, SOH can be estimated, and RUL inferred

from real-world driving and charging records by learning longitudinal health-evolution patterns [43]. Another study reconstructs long-term capacity degradation trajectories from multi-year fleet data and applies stochastic processes to enable degradation trajectory forecasting with uncertainty quantification [44]. In addition, differences in EV operating characteristics are incorporated into prognostic frameworks to improve generalization across vehicles and usage conditions [45]. However, accurately determining the true aging state of a battery pack in the field generally requires check-up tests, which are significantly more expensive and difficult to conduct than cell-level tests under laboratory conditions. In addition, domain shifts across vehicles, chemistries, and usage patterns further challenge generalization. **Fig. 1** provides an illustrative overview of the multi-resolution system complexities and influential factors involved in real-world EV battery aging, spanning intra-battery pack characteristics, intra-vehicle features, usage patterns, and external operating conditions.

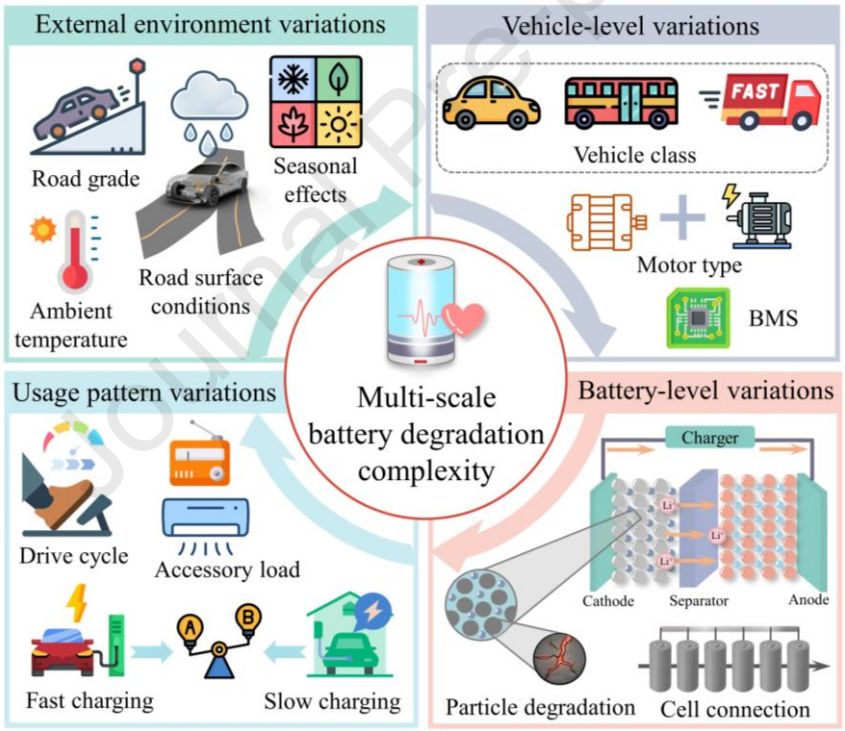


Fig. 1. Multi-resolution system complexities and influential factors affecting battery aging and RUL prediction in real-world EVs.

In summary, existing battery aging diagnosis and RUL prediction methods face three fundamental challenges when deployed in real-world EVs: (i) the high cost of obtaining reliable aging labels; (ii) limited generalization across heterogeneous operating conditions; and (iii) insufficient utilization of large-scale unlabeled field data. Existing approaches address the scarcity of labeled data from different perspectives.

Semi-supervised methods typically incorporate unlabeled samples by generating pseudo-labels and integrating them into supervised training, thereby making the model sensitive to the accuracy of these pseudo-labels. Data augmentation approaches, particularly physics-informed methods, enhance training data by constructing synthetic degradation trajectories based on domain knowledge. Still, their effectiveness depends on the validity of the underlying modeling assumptions. In contrast, self-supervised learning offers an alternative paradigm by directly exploiting unlabeled data for representation learning without converting them into labeled samples or relying on synthetic data generation. Therefore, this study proposes a contrastive-enhanced VAE–LSTM self-supervised learning framework specifically architected for RUL prediction using heterogeneous and unlabeled EV field data (**Fig. 2**). While traditional unsupervised models primarily focus on signal reconstruction, the proposed framework integrates a contrastive consistency objective within the VAE encoder to explicitly learn monotonic, degradation-aware latent representations from raw charging sequences. This design effectively decouples general representation learning from the need for large-scale labels, enabling high-precision RUL adaptation in the target domain using only a sparse set of labeled vehicles. The main contributions of this study are summarized as follows:

(1) A contrastive-enhanced VAE–LSTM framework is developed for battery RUL prediction. By leveraging large-scale unlabeled real-world battery data in a self-supervised learning paradigm and shifting from reconstruction-only learning to joint reconstruction and contrastive learning, the proposed method overcomes the limitations of conventional VAEs in handling highly irregular field data.

(2) The proposed framework addresses the high heterogeneity of real-world EV fleets, including diverse vehicle categories (passenger vehicles, taxis, and city buses), battery chemistries, pack configurations, and operating environments.

(3) A label-efficient downstream adaptation strategy is developed to support cross-fleet and cross-domain RUL prediction under limited supervision. Experimental results demonstrate that, when transferred to heterogeneous target fleets with sparse labels, the proposed framework achieves root-mean-squared-error (RMSE) values of 40–50 cycles, corresponding to a relative improvement of 42%–53% over direct-training baselines.

(4) An interpretability-oriented analysis is incorporated to investigate how the learned latent representations and temporal aggregation mechanisms capture battery degradation dynamics across different aging stages and operating regimes.

(5) By requiring labels from only a small fraction of vehicles, the total RUL labeling cost can be reduced by approximately 70%, while retaining about 95% of the predictive performance of supervised baselines trained with 100% labeled data.

The rest of this study is structured as follows. Section 2 covers the field data acquisition and preprocessing procedures. Section 3 details the methodology and the architecture of the proposed model. Section 4 presents the prediction results and performance evaluation. Section 5 concludes with a summary of the key findings and prospects for future work.

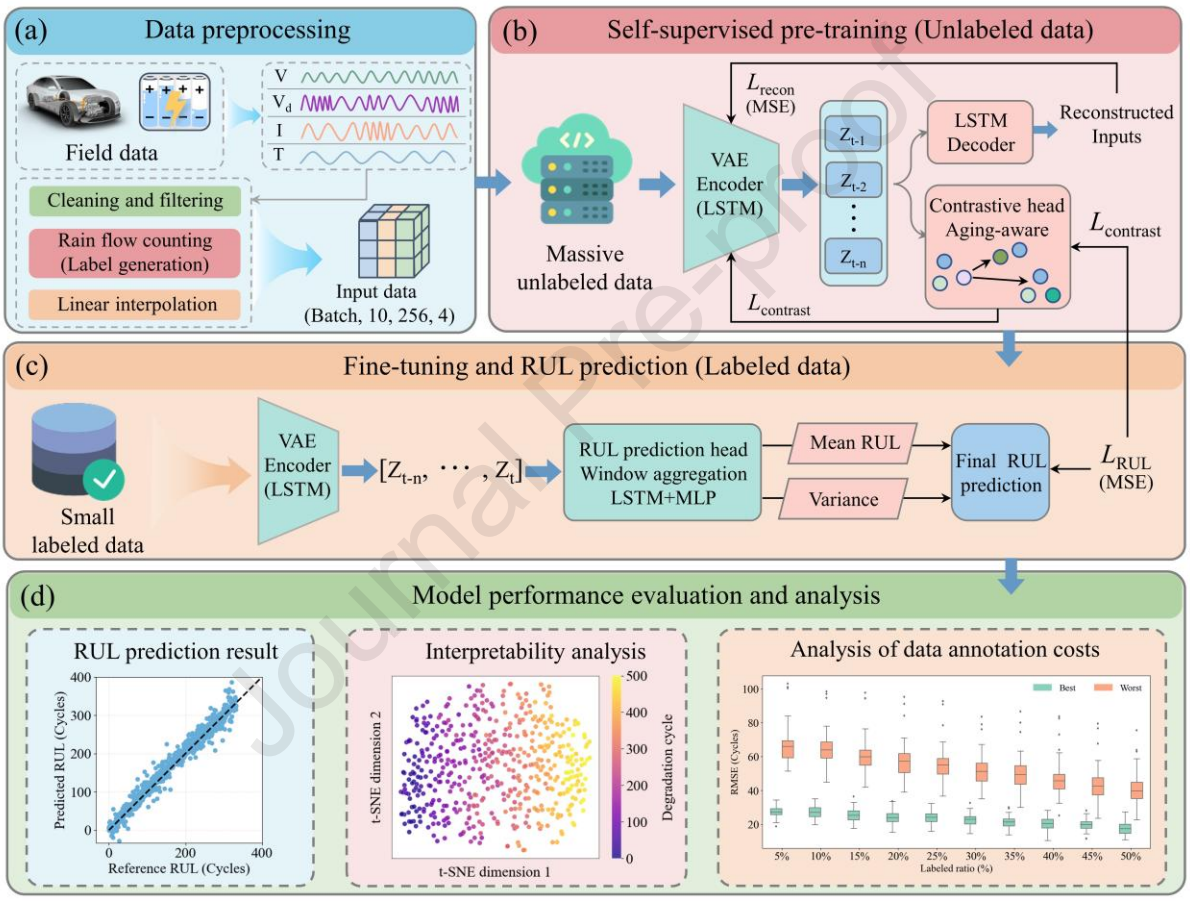


Fig. 2. A Self-supervised contrastive learning framework for data-efficient RUL prediction of EV Batteries.

2. Data Generation

2.1. Data source

This study employs three real-world EV field datasets to evaluate the proposed RUL prediction framework under diverse operating conditions and fleet scales. All three datasets are collected from onboard battery packs during practical vehicle operation, rather than from laboratory aging experiments. Dataset #1

is obtained from an operational monitoring platform, while Dataset #2 and Dataset #3 are derived from two publicly reported fleet datasets [46,47]. Since these datasets are collected during real-world vehicle operation, no predefined standard driving cycles, such as the urban dynamometer driving schedule or the worldwide harmonized light vehicles test cycle, are imposed. Instead, they capture real service profiles of passenger cars, taxis, and buses, covering heterogeneous battery chemistries, usage patterns, and service durations. This diversity supports a comprehensive assessment of model robustness and generalization in practical deployment scenarios. **Table 1** summarizes the key characteristics of the three datasets, including the battery chemistry, vehicle type, fleet size, sampling configuration, and operating conditions.

Dataset #1 is collected from an EV operational monitoring platform in China and contains long-term operational records of 20 EVs over approximately 2 years. It includes passenger vehicles, taxis, and city buses with multiple nominal battery capacities ranging from 126 Ah to 174 Ah. These vehicles operate in both southern and northern regions of China, introducing variability in climate, terrain, driving behavior, and charging patterns. The fleet further includes two battery chemistries, lithium iron phosphate (LFP) and lithium nickel cobalt manganese oxide (NCM), as well as different pack specifications, including rated capacity and series cell count. Therefore, the dataset reflects heterogeneous operating conditions, including urban commuting, taxi service, and bus stop-and-go operations, as well as diverse charging and thermal environments. To provide an intuitive illustration of this variability, **Fig. 3**(a)-(f) presents the statistical distributions and temporal characteristics of Dataset #1 as a representative case. Specifically, it includes the histograms of cumulative mileage and battery capacity (**Fig. 3**(a)-(b)), as well as the typical capacity degradation trajectories with respect to cycle number (**Fig. 3**(c)). Furthermore, boxplot distributions of key operating parameters across different months, including state of charge (SOC), pack current, and pack voltage (**Fig. 3**(d)-(f)), are provided to highlight the significant variability observed under real-world conditions.

Dataset #2 also consists of data from 20 taxis, but focuses on a more uniform configuration [46]. All vehicles are BAIC EU500 models equipped with CATL NCM battery systems and are monitored for 29 months with an 8-second (sec) sampling interval. This dataset represents a relatively homogeneous urban taxi duty cycle, characterized by repeated daily driving, frequent charging events, and similar vehicle platform configuration. Dataset #3 represents a large-scale fleet scenario and includes real-world operational data collected from 300 taxis that have completed their service life, with service durations ranging from 0.5

to 4 years [47]. This large-scale taxi fleet dataset mainly captures long-term urban service operations. The vehicles experience charging, discharging, and idle states during real-world use, reflecting practical operating patterns rather than standardized test cycles. Therefore, this dataset covers a comprehensive set of driving states, including charging, discharging, and idle conditions, and is recorded at a fixed sampling rate (per 10 sec), resulting in approximately 850 million time-series data frames. To illustrate cross-fleet heterogeneity and differences in label distributions across the three datasets, global comparisons are presented in **Fig. 3**(g)-(i). These include t-SNE feature space distributions (**Fig. 3**(g)), joint probability density contours of pack voltage and current (**Fig. 3**(h)), and capacity probability density distributions (**Fig. 3**(i)). Finally, to connect these data characteristics with the modeling objective, **Fig. 3**(j) provides a schematic overview of the prediction task. It defines the historical observation window, the current prediction point, and the target horizon associated with a predefined degradation threshold. The scale and heterogeneity of Dataset #3 are leveraged to examine the scalability and generalization capability of the proposed framework for RUL prediction in large and diverse EV fleets.

Table 1. Key characteristics of the three real-world EV field datasets used in this study.

Description	Dataset #1	Dataset #2	Dataset #3
Number of vehicles	20	20	300
Nominal capacity	126/128/140/145/174Ah	145 Ah	155 Ah
Number of cells	84/88/91/95/96/98	90	96
Vehicle type	Passenger car/taxi, and bus	Taxi	Taxi
Battery materials	LFP/NCM	NCM	NCM
Collecting period	2 years	29 months	0.5-4 years
Sampling interval	10 sec	8 sec	10 sec
Operating condition	Mixed real-world service conditions of passenger cars, taxis, and buses	Relatively homogeneous urban taxi operation	Large-scale urban taxi fleet operation

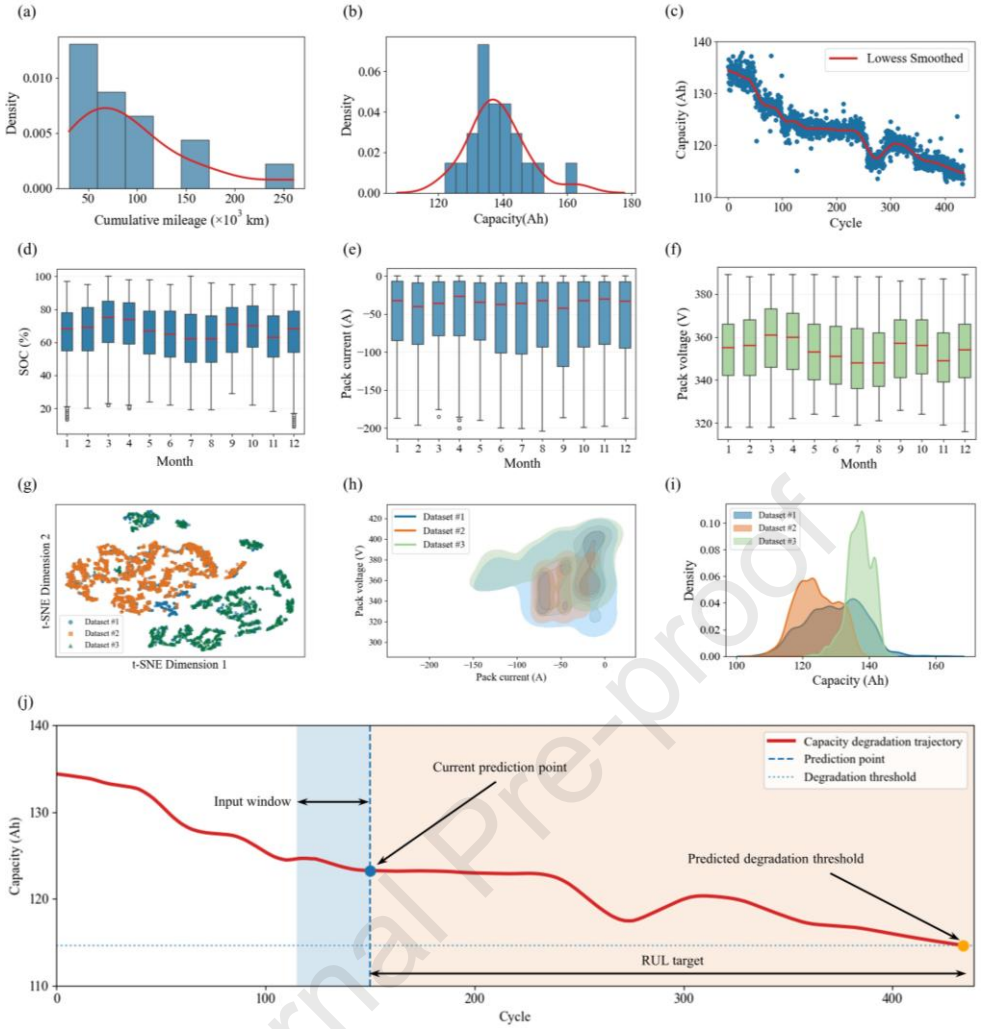


Fig. 3. Overview of dataset characteristics and the prediction task under real-world field conditions.

2.2. Data preprocessing

To ensure reliable representation learning from noisy and heterogeneous real-world EV battery data, a structured data preprocessing and feature engineering procedure is adopted before self-supervised model training. Raw battery measurements are first screened to remove abnormal values caused by sensor faults or communication errors through statistical outlier detection and physical constraint checking. Charging events are identified from continuous vehicle operation records, and only segments with a SOC increase of at least 30% are retained. Since the capacity estimation relies on a ratio involving ΔSOC, very short charging segments make the estimate highly sensitive to SOC measurement noise and quantization effects. On the other hand, enforcing a very large SOC window would remove most naturally occurring charging events in field operation. Therefore, a moderate constraint of $\Delta SOC \geqslant 30\%$ is adopted to ensure numerically stable estimation while retaining sufficient real-world samples for model training [48].

To eliminate variability introduced by different charging durations, all retained charging segments are temporally normalized by interpolation to a fixed length of 256 samples. Each normalized segment is further decomposed using a sliding window with a window length of 10 samples, generating overlapping subsequences that preserve local temporal continuity and enable fine-grained representation learning. For each window, four constructed features are extracted as model inputs. These include pack-level charging voltage, charging current, cell voltage imbalance, and average pack temperature. The cell voltage imbalance is defined as the difference between the maximum and minimum cell voltages within the battery pack. These features are not raw measurements but engineered descriptors derived from standard onboard signals, designed to jointly characterize electrical loading conditions, cell-level inconsistency, and thermal behavior during charging, all of which are closely associated with battery aging mechanisms [49-51]. After the aforementioned preprocessing and feature engineering, a rigorous data-splitting protocol is implemented at the vehicle level to construct the training and testing sets. This vehicle-wise splitting strategy is fundamental to preventing data leakage and ensuring that the model's generalization capability is evaluated on vehicles that are completely unseen.

2.3. Calculation of reference capacity and RUL labels

To provide a quantitative and physically interpretable health indicator for subsequent RUL labeling, a reference capacity is derived from field charging measurements. Because the SOC reported by the onboard BMS is an estimated state rather than a directly measured physical quantity, the derived capacity is subject to uncertainty. In this study, the reference capacity is computed as the integrated charge throughput during a charging event, normalized by the corresponding SOC increment. The reference capacity is presented by Eqn. (1),

$$C_t = \frac{-\int_{t_0}^{t_i} I(t)dt}{\Delta SOC} \tag{1}$$

where C_t denotes the estimated charging capacity, and I is the charging current. t_i and t_0 are the start and end times of the charging process, respectively. Here, ΔSOC denotes the increase in SOC over the retained charging event. Let the reported SOC be expressed as $SOC(t) = SOC(t) + e(t)$ where $e(t)$ denotes the SOC estimation error. Then, the reported SOC increment becomes $\Delta SOC(t) = \Delta SOC + e(t_i) - e(t_0)$. Accordingly, the capacity derived from Eqn. (1) is primarily influenced by the differential SOC estimation error over the

charging segment, whereas a constant offset in absolute SOC is partially canceled when the SOC increment is used. Under the assumption of small estimation errors, the relative capacity error can be approximated as $(C_t - C_t) / C_t \approx -[e(t_i) - e(t_0)] / \Delta SOC$. This relation indicates that charging segments with small SOC increments are particularly sensitive to SOC estimation uncertainty. It should be noted that this error propagation inevitably affects the overall RUL prediction. Since the derived capacity is subsequently used to construct the degradation trajectory and define RUL labels, any unmitigated residual uncertainty in SOC estimation introduces high-frequency label noise. If a data-driven model is trained directly on these noisy reference values, it is highly prone to overfitting to transient estimation disturbances caused by the production BMS algorithms, rather than learning the true physical battery degradation mechanisms, thereby severely degrading final prediction accuracy.

To mitigate this effect while retaining sufficient field samples, only charging segments with an SOC increase of at least 30% are retained for capacity calculation. In addition, raw records are screened by outlier detection and physical constraint checking, and the event level capacity sequence is further calibrated by locally weighted regression (LOWESS) smoothing [52,53]. The LOWESS smoothing procedure helps suppress local fluctuations due to sensor noise, incomplete charging behavior, transient disturbances, and residual uncertainty in SOC estimation, while preserving the long-term degradation trend. Accordingly, the obtained capacity values are treated as reference estimates from field operations rather than as direct ground-truth capacity measurements. Although some bias may remain, the processed capacity trajectories reliably retain the essential degradation information required for accurate downstream RUL prediction.

RUL labels are derived in the cycle domain because real-world datasets commonly contain heterogeneous degradation rates and incomplete end-of-life observations. The cycle index used for RUL labeling is computed by the rainflow counting after abnormal-value removal, counting is performed on the full operational record, including both charging and discharging data, rather than being restricted to deep charging segments [44]. Instead of adopting a static, single-value failure definition, a predefined capacity degradation milestone is defined to map the calibrated capacity trajectory to a target threshold-crossing cycle. The evaluation threshold is presented by Eqn. (2):

$$\begin{cases} C_{\text{th}} = \eta C_{\text{nom}} \\ N_{\text{th}} = \min\left\{ N : C(N) \le C_{\text{th}} \right\} \end{cases} \tag{2}$$

where C_{th} denotes the capacity evaluation threshold, C_{nom} is the nominal capacity, and η is the threshold ratio. To address the heavy right-censoring in real-world field data and comprehensively evaluate the generalizability of the proposed framework, a multi-threshold strategy is adopted across domains. Specifically, η is set to 0.92 for the source domain (Dataset #1), 0.85 for Target Dataset #2, and the industry-standard 0.80 for Target Dataset #3.

The setting of η = 0.92 for Dataset #1 acts as an early-stage degradation milestone rather than a physical end-of-life (EOL). In practical EV operations, collecting full-lifecycle data by waiting until batteries reach their standard physical EOL (e.g., 80% SOH) significantly hinders timely industrial deployment. Because the field data in the source domain (Dataset #1) span a relatively short observation period of 2 years, the capacities of these vehicles degrade only to 0.90-0.92. Therefore, 0.92 is strategically selected as an observable intermediate milestone to validate the model's feature extraction capability during early-stage capacity fade. Furthermore, applying 0.85 and 0.80 as evaluation thresholds in the target domains demonstrates a core advantage of this study: temporal and cross-domain transferability. Target Dataset #2 contains 29 months of operational data, while Target Dataset #3 spans 4 years, allowing the batteries to reach deeper levels of degradation. Evaluating the framework on these datasets demonstrates that latent representations learned from short-term (2-year), right-censored field data (Dataset #1) can be successfully transferred to predict mid- to late-stage degradation and the true physical EOL in unseen fleets operating over significantly longer periods. The corresponding threshold cycle N_{th} is defined as the first equivalent cycle at which the calibrated capacity trajectory reaches or falls below C_{th}. The RUL, in the context of these specific milestones, is defined in the cycle domain as the difference between the predicted threshold cycle and the current equivalent cycle count, presented by Eqn. (3):

$$RUL = N_{th} - N_{cur} \tag{3}$$

where N_{th} is defined as above, and N_{cur} is the current equivalent cycle index obtained via rainflow counting of the operational history.

3. Methodology

3.1. Self-supervised VAE–LSTM representation learning

Large-scale EV battery operation data usually contains rich temporal information but lacks reliable RUL labels. To exploit such unlabeled data, a self-supervised representation learning module is constructed

using a VAE combined with LSTM networks. As illustrated in **Fig. 4**, this module is designed to extract degradation-aware latent representations by jointly modeling temporal dependencies and stochastic uncertainty in battery aging signals. While the VAE-LSTM architecture has been previously explored for unsupervised prognostic tasks, its standard formulation often struggles with the severe domain shifts and noise inherent in large-scale EV operational data. To address this limitation, the proposed representation learning module introduces two critical enhancements to the standard paradigm. First, instead of optimizing solely for sequence reconstruction, a contrastive learning objective is embedded to force the latent space to capture ordered, aging-aware progressions. Second, the framework decouples feature extraction from downstream RUL regression, enabling zero-shot-like feature preservation during target domain adaptation. The key architectural settings of the proposed model are summarized in **Table 2**.

In this study, LSTM is adopted as the temporal modeling backbone due to its suitability for real-world EV battery data. First, the available labeled data are limited, and Transformer-based models typically require large-scale labeled datasets to train stably. In contrast, LSTM provides a more data-efficient solution and can be reliably trained under limited supervision. Second, the field data considered in this work are characterized by irregular sampling, measurement noise, and transient disturbances. Transformer architectures rely on attention mechanisms, which may be more sensitive to noise and can assign high weights to spurious fluctuations. In comparison, LSTM models perform implicit temporal smoothing through gated recurrent structures, making them more robust for modeling noisy and irregular time-series data. Third, Transformer-based models generally involve higher computational complexity and memory consumption, especially for long sequences, which may limit their practicality in large-scale industrial deployment. LSTM provides a more efficient and stable alternative under such constraints. Finally, it is important to note that the main contribution of this work lies in the proposed self-supervised representation learning framework rather than in the choice of a specific temporal backbone. The LSTM component is adopted as a practical implementation for temporal aggregation, and the framework can be extended to other architectures, such as Transformer-based models, in future work. This consideration also applies to other deep neural network architectures for time-series modeling, where increased model complexity and data requirements may limit their effectiveness in noisy, low-label settings.

Based on the above architectural configuration, the input representation and encoding process are formally defined as follows. A multivariate battery time series is first segmented into fixed-length sequences and denoted as a matrix. The encoding process is presented by Eqn. (4),

$$\mathbf{h} = \mathrm{Enc}_{\mathrm{LSTM}}(\mathbf{X}) \tag{4}$$

where, $\mathbf{X} \in \mathbb{R}^{T \times D}$ denotes the input sequence, with T being the temporal length and D the feature dimension; $\mathrm{Enc}_{\mathrm{LSTM}}(\cdot)$ is the shared LSTM encoder; and $\mathbf{h} \in \mathbb{R}^{H}$ denotes the latent representation with dimensionality H. The parameterization of the latent distribution is presented by Eqn. (5),

$$\mu = W_{\mu} h + b_{\mu}, \qquad \log \sigma^2 = W_{\sigma} h + b_{\sigma} \tag{5}$$

where, μ and $\log \sigma^2$ denote the mean and log-variance of the latent Gaussian distribution, respectively. The parameters are obtained via linear projections of the hidden representation h, with learnable weight matrices W_{μ}, W_{σ}, and bias terms b_{μ}, b_{σ}. To enable stochastic sampling while preserving differentiability, the latent variable is generated using the reparameterization trick. A Gaussian latent variable $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ is expressed as a deterministic transformation of a noise variable $\epsilon \sim \mathcal{N}(0, \mathrm{I})$, given by Eqn. (6),

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon, \ \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{6}$$

where μ and σ denote the predicted mean and standard deviation, respectively, and ϵ is sampled from a standard normal distribution. The latent representation is then decoded to reconstruct the original input sequence, which enforces the preservation of degradation-relevant information. The reconstruction process is presented by Eqn. (7),

$$\hat{\mathbf{X}} = \mathrm{Dec}_{\mathrm{LSTM}}(\mathbf{z}) \tag{7}$$

where, $\mathrm{Dec}_{\mathrm{LSTM}}(\cdot)$ denotes the LSTM-based decoder, and $\hat{\mathbf{X}} \in \mathbb{R}^{B \times T \times D}$ represents the reconstructed sequence. The self-supervised training objective is defined by jointly minimizing the reconstruction error and the Kullback–Leibler (KL) divergence [54]. The self-supervised loss function is presented by Eqn. (8),

$$\mathcal{L}_{\mathrm{SSL}} = \alpha \mathcal{L}_{\mathrm{rec}} + \beta \mathcal{L}_{\mathrm{KL}} \tag{8}$$

where, $\mathcal{L}_{\mathrm{rec}}$ denotes the mean squared reconstruction loss between $\mathbf{X}$ and $\hat{\mathbf{X}}$, $\mathcal{L}_{\mathrm{KL}}$ represents the KL divergence between the approximate posterior and a standard normal prior, α controls the contribution of the reconstruction term, and β regulates the strength of latent space regularization. For a diagonal Gaussian latent distribution, the KL divergence is computed in closed form and is presented by Eqn. (9),

$$L_{KL} = -\frac{1}{2}\sum_{i=1}^{L}\left(1+\log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right) \tag{9}$$

where $\mu \in \mathbb{R}^L$ and $\log\sigma^2 \in \mathbb{R}^L$ are produced by the encoder, $\sigma_i^2 = \exp\left(\log\sigma_i^2\right)$ and L is the latent dimension.

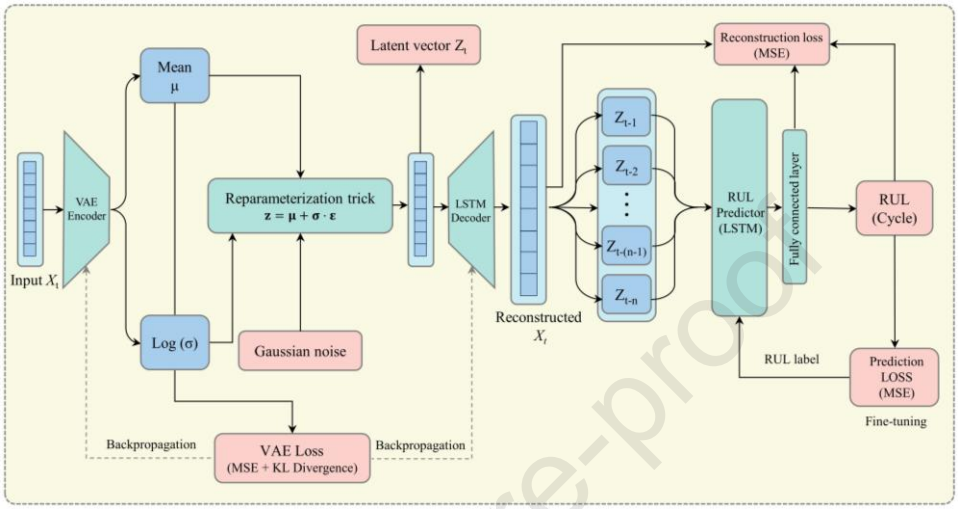


Fig. 4. Self-supervised VAE–LSTM framework for EV battery RUL prediction.

Table 2. Hyperparameters of the proposed self-supervised VAE–LSTM architectures.

Parameter	Meaning	Value
Window_size	Number of previous cycles	10
Seq_len	Length of a single cycle data sequence	256
Input_dim	Input feature dimension	4
Batch_size	Batch size for training and inference	64
patience	Early stopping patience	20
Pre_epochs	Number of epochs for self-supervised pre-training	150
Pre_lr	Learning rate for pre-training	1e-3
Alpha	Reconstruction loss weight	0.01
Beta	KL divergence weight (VAE regularization)	0.0 - 0.001 (Warmup)
contrast_w	Contrastive loss weight	0.5
Temp	Temperature parameter for contrastive loss	0.1
Ft_epochs	Total epochs for fine-tuning	65
Ft_lr	Learning rate during fine-tuning	2e-4
Hidden_dim	Hidden dimension size for encoder/decoder LSTM	256
Latent_dim	Dimension of the VAE latent space	8
Enc_layers	Number of LSTM layers in the encoder	3
Dec_layers	Number of LSTM layers in the decoder	3
Dropout	Dropout rate for encoder/decoder	0.1
RUL_LSTM_dim	Hidden size for RUL head's aggregator LSTM	64
RUL_MLP_dim	Hidden dimensions for RUL regression MLP	[64, 32]

3.2. Downstream RUL fine-tuning and prediction

After self-supervised representation learning, the learned latent features are transferred to the downstream RUL prediction task through a limited-label fine-tuning strategy. To capture local degradation trends and reduce sensitivity to short-term noise, a sliding window mechanism is applied to aggregate multiple consecutive latent vectors. The construction of the latent window is presented by Eqn. (10),

$$\mathbf{Z}_t = \{\mathbf{z}_{t-W+1}, \mathbf{z}_{t-W+2}, \ldots, \mathbf{z}_t\} \tag{10}$$

where, $\mathbf{z}_t \in \mathbb{R}^L$ denotes the latent representation at cycle *t*, *W* is the window size, and $\mathbf{Z}_t \in \mathbb{R}^{W \times L}$ represents the windowed latent sequence used for RUL prediction. The windowed latent sequence is then fed into an LSTM-based temporal aggregator to capture degradation progression over recent cycles. This temporal aggregation is defined in Eqn. (11),

$$\mathbf{h}_t^{\mathrm{RUL}} = \mathrm{LSTM}_{\mathrm{RUL}}\left(\mathbf{Z}_t\right) \tag{11}$$

where, $\mathrm{LSTM}_{\mathrm{RUL}}(\cdot)$ denotes the LSTM aggregator dedicated to the RUL task, and $\mathbf{h}_t^{\mathrm{RUL}}$ is the extracted degradation trend representation. Furthermore, a normalized life-progress indicator is used to represent the relative position within the battery lifetime. The final RUL prediction mapping is presented by Eqn. (12),

$$\hat{y}_t = f_{\mathrm{MLP}}\left(\left[\mathbf{h}_t^{\mathrm{RUL}}, \lambda_t\right]\right) \tag{12}$$

where, $\hat{y}_t$ denotes the predicted normalized RUL at cycle *t*, $f_{\mathrm{MLP}}(\cdot)$ represents a multi-layer perceptron (MLP) regressor , and $\lambda_t \in [0,1]$ is the normalized life-progress variable. The downstream fine-tuning objective is defined using a supervised regression loss. The RUL loss function is presented by Eqn. (13),

$$\mathcal{L}_{\mathrm{RUL}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{13}$$

where, y_i is the reference RUL label(derived based on Eqn. 3), $\hat{y}_t$ is the corresponding prediction, and *N* is the number of labeled samples used for fine-tuning.

To effectively transfer knowledge acquired during the self-supervised pre-training stage to downstream tasks, all experiments in this study adopt a fine-tuning strategy in which the encoder is frozen. At the same time, only the decoder and the task-specific RUL prediction head are fine-tuned. Specifically, the encoder weights initialized during the pre-training stage remain fixed, while the decoder and RUL prediction head are updated using labeled target data. This design choice is motivated by the need to bridge the domain gap

between the source and target datasets, allowing the decoder and RUL prediction head to refine their representations and adapt to the complex nonlinear degradation behaviors commonly observed in the later stages of battery life. Consequently, the downstream training objective relies solely on the supervised regression loss defined in Eqn. (13) to optimize all trainable parameters across the decoder and RUL prediction head. To evaluate RUL prediction performance, this study adopts a widely used regression metric. The RMSE is employed to measure the overall prediction accuracy and quantify the magnitude of prediction errors. The definition of RMSE is provided by Eqn. (14),

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{14}$$

where $\hat{y}_i$ is the predicted RUL for the *i*-th cycle, y_i is the corresponding reference RUL, and *n* is the total number of cycles for a single battery pack.

4. Results and Discussion

4.1. Self-supervised RUL prediction results

Fig. 5 presents the performance of the proposed self-supervised framework for RUL prediction under real-world EV operating conditions, based on Dataset #1, and compares it with supervised and semi-supervised baselines. **Fig. 5**(a)-(c) compare predicted RUL against reference RUL for the three training strategies. In the self-supervised setting, the predicted RUL shows a strong linear alignment with the reference values across the full life range, with most samples clustered near the diagonal. In contrast, the supervised and semi-supervised models show noticeably larger dispersion, particularly in the medium- and long-horizon RUL regions, indicating reduced robustness when trained with limited or noisy labels. The quantitative results reported in **Table 3** further highlight the superiority of the self-supervised approach in terms of both typical performance and worst-case robustness. Specifically, the self-supervised framework achieves a median RMSE of 15 cycles, substantially lower than that of the supervised (27 cycles) and semi-supervised (42 cycles) baselines. In addition, the 95th-percentile RMSE of the self-supervised model is limited to 56 cycles, compared with 125 cycles and 204 cycles for the supervised and semi-supervised methods, respectively, indicating a markedly reduced risk of large prediction errors.

Fig. 5(d)-(e) further provide case-level full-lifecycle trajectories for two representative individual vehicles (whose IDs are "V10" and "V16", respectively, belonging to the same vehicle class) selected from

the test set of Dataset #1, illustrating long-horizon prediction behavior and the evolution of uncertainty. Notably, the significant drop in the trajectory of V16 (**Fig. 5**(e)) reflects a typical nonlinear, sudden degradation phenomenon in real-world EVs, often triggered by BMS recalibrations, severe seasonal temperature shocks, or long-term parking. The predicted trajectories closely track the reference RUL evolution across the full degradation process, capturing even abrupt drops, while the uncertainty bounds remain tight over most of the service life. This behavior suggests that the learned representations capture the monotonic degradation characteristics of LIBs under real-world operation, rather than overfitting to localized capacity fluctuations or measurement noises.

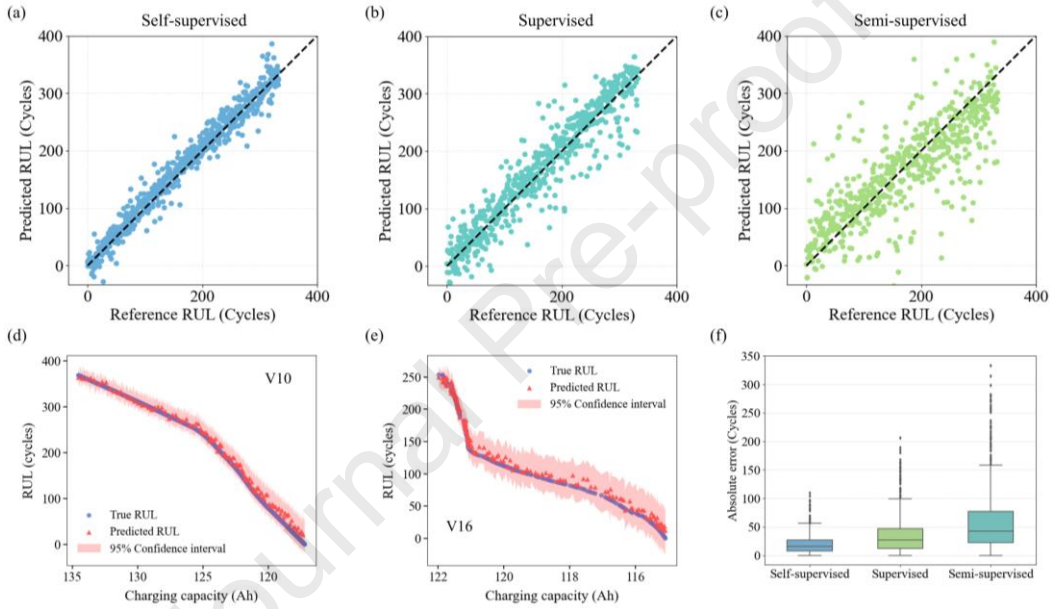


Fig. 5. Performance of the proposed framework for EV RUL prediction on Dataset #1.

Table 3. Robustness-oriented RMSE statistics of self-supervised, supervised, and semi-supervised methods evaluated on 10 test EVs from Dataset #1.

Methods	RMSE (Mean)	RMSE (Median)	RMSE (95th perc.)
Self-supervised	27	15	56
Supervised	51	27	125
Semi-supervised	87	42	204
Improvement (vs. Supervised)	47.1%	44.4%	55.2%
Improvement (vs. Semi-supervised)	69%	64.3%	72.5%

4.2. Explainable EV battery RUL prediction

To elucidate the decision-making mechanism of the proposed framework, a post-hoc interpretability analysis is performed using SHapley Additive exPlanations (SHAP) on the latent representations learned by

the encoder [55]. The feature-importance ranking in **Fig. 6**(a) shows a highly non-uniform distribution of contributions across dimensions. Specifically, Z_1 exhibits the largest impact on the predicted RUL (mean SHAP≈0.062), followed by Z_2 and Z_3, while the remaining dimensions contribute only marginally. These dominant components should be interpreted as principal predictive directions in the embedding space rather than predefined electrochemical variables. A high SHAP value implies that small variations along the corresponding representation axis lead to significant changes in the predicted lifetime. Therefore, the model decision is effectively controlled by a compact aging-related coordinate system learned from operational charging dynamics. Conversely, dimensions with low SHAP importance (e.g., Z_7 and Z_8) have minimal influence on the output, indicating that the regression head selectively relies on informative structures while suppressing less relevant variations present in the data.

The organization of these representations is further illustrated in **Fig. 6**(b) using t-SNE [56]. The samples form a continuous, ordered manifold in which the color-coded degradation cycles evolve smoothly from early life to failure. This structure demonstrates that the encoder maps battery operational states onto an ordered degradation progression within the latent space. Instead of performing direct curve matching, the model estimates RUL by determining the relative position of the current battery state along this learned progression trajectory. Together, the SHAP importance ranking and the monotonic embedding structure provide a behavior-level explanation of the prediction process. The framework first compresses complex charging dynamics into a low-dimensional degradation progression coordinate, and the regression head then infers RUL from the geometric location along this coordinate. This explains why the model generalizes across heterogeneous operating conditions: the prediction relies on the progression of degradation state rather than on specific waveform patterns.

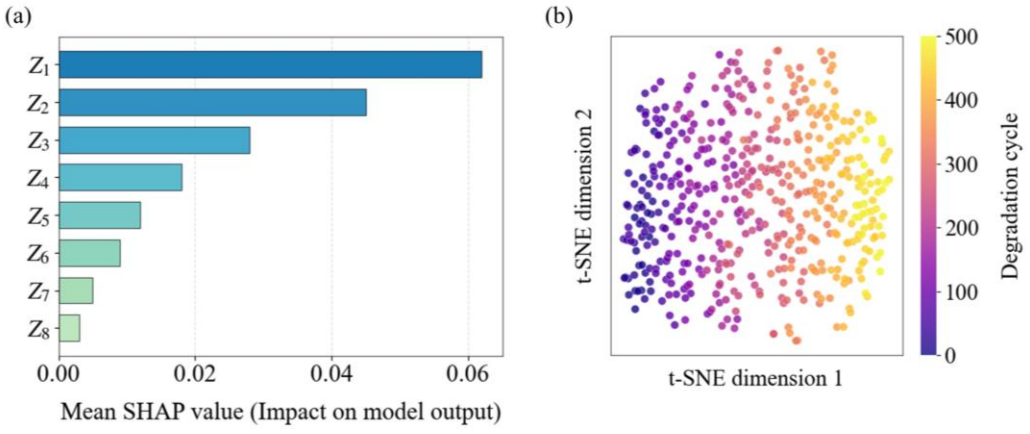


Fig. 6. Feature importance and evolution trajectory of latent variables.

4.3. Generalizability of the proposed framework

The generalizability of the proposed self-supervised framework is evaluated by transferring the pretrained model from the source dataset to two target datasets with different fleet sizes and operating characteristics. Specifically, Dataset #1 serves as the source domain and represents a heterogeneous real-world fleet (20 vehicles: passenger cars, taxis, and city buses) with diverse vehicle types, battery configurations, geographical regions, and operating conditions. Dataset #2 and Dataset #3 are selected as target domains to evaluate transferability in two complementary scenarios: a small-scale, relatively homogeneous fleet (20 taxis) and a large-scale, heterogeneous fleet (300 taxis). For a fair comparison under practical deployment conditions, model adaptation in the target domain is conducted using labeled data from 30% of the vehicles in each target dataset. In contrast, the remaining vehicles are used exclusively for testing. To ensure that the results are statistically significant and independent of specific data splits, a repeated random sub-sampling validation process is employed. Specifically, the 30% train, 70% test split is randomly performed 5 times using different random seeds (31, 10, 59, 112, and 8). The detailed trial-by-trial results and the corresponding standard deviations are summarized in **Table 4**, demonstrating the stability of the proposed framework across different vehicle selections.

Fig. 7(a)-(c) presents scatter plots comparing predicted and reference RUL for the source Dataset #1, target Dataset #2 (containing 20 vehicles), and target Dataset #3 (containing 300 vehicles), respectively. For the source Dataset #1, predicted RUL values remain tightly clustered around the diagonal, indicating high consistency between predictions and reference labels across the full RUL range. When transferred and fine-tuned on the target datasets, the model preserves a clear linear relationship between predicted and reference RUL. The quantitative results, summarized in **Table 5**, further support the transferability of the proposed framework with limited labeled supervision. On the source Dataset #1, the model achieves an RMSE (mean, median, and 95th percentile) of 27, 15, and 56 cycles, respectively. After fine-tuning with labeled data from 30% of the target-domain vehicles, the RMSE (mean, median, and 95th percentile) on target Dataset #2 becomes 43, 26, and 89 cycles, respectively. For the larger and more heterogeneous target Dataset #3, these values further increase to 50, 30, and 100 cycles, respectively. Although performance degrades under cross-dataset transfer, the median and 95th-percentile RMSE remain bounded, indicating that typical performance and tail-risk robustness are preserved to a considerable extent under limited-label adaptation.

Table 4. Detailed RMSE (Mean) results of repeated random sub-sampling validation across 5 independent trials with different random seeds.

Dataset	Trial 1 (Seed 31)	Trial 2 (Seed 10)	Trial 3 (Seed 59)	Trial 4 (Seed 112)	Trial 5 (Seed 8)	Average (Mean ± SD)
Target Dataset #2	43	40	46	41	45	43 ± 2.55
Target Dataset #3	50	46	55	48	51	50 ± 3.39

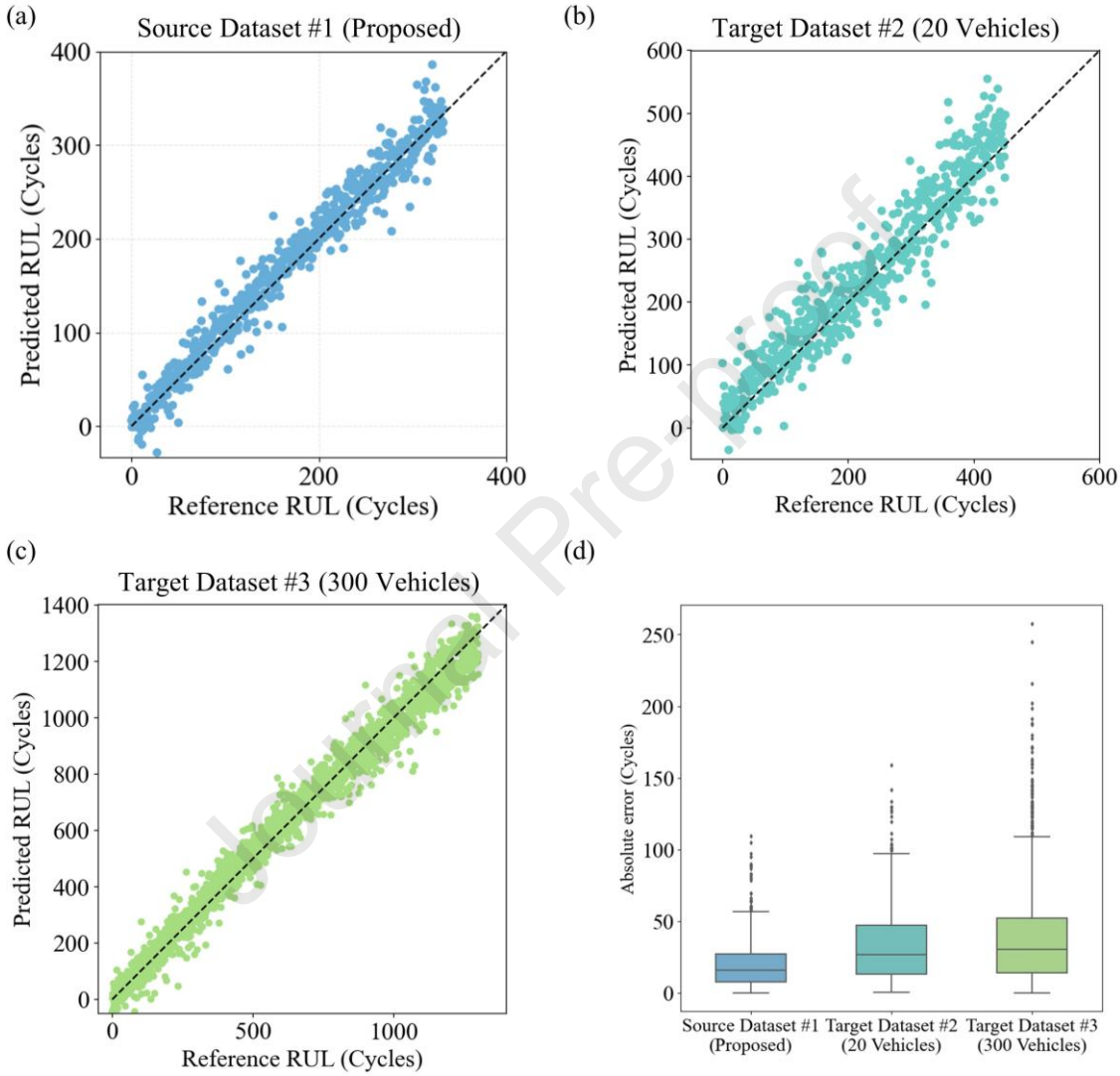


Fig. 7. Generalizability performance of the proposed framework across source and target datasets.

To explicitly evaluate the benefit of the proposed transfer strategy, the proposed framework is compared against a baseline supervised approach (Direct training). This baseline model is trained from scratch using only the available labeled data (30%) in the target domains, without leveraging any source-domain knowledge. As shown in **Table 5**, the baseline approach exhibits poor performance, with mean RMSEs of 75 and 91 cycles for Dataset #2 and Dataset #3, respectively. Furthermore, the high 95th-percentile errors (168 and 215 cycles) indicate that the baseline model is highly unstable and prone to overfitting when data

is scarce. In contrast, the proposed transfer learning framework significantly narrows this gap, reducing the mean RMSE by approximately 43% and 45%, respectively. Notably, the proposed transfer learning framework significantly reduces the mean RMSE by 42.7% for Dataset #2 and 45.1% for Dataset #3 compared to the baseline. These results confirm that the framework maintains stable performance as it scales to large, diverse vehicle fleets, even when only a small number of vehicles are labeled.

Table 5. Summary of generalizability performance metrics on the source and target datasets.

Dataset	Method	RMSE (Mean)	RMSE (Median)	RMSE (95th perc.)
Source Dataset #1	Proposed	27	15	56
Target Dataset #2	Proposed (transfer)	43	26	89
	Direct training (baseline)	75	48	168
	Improvement	42.7%	45.8%	47%
Target Dataset #3	Proposed (transfer)	50	30	100
	Direct training (baseline)	91	62	215
	Improvement	45.1%	51.6%	53.5%

To further examine the generalizability of the proposed framework using publicly available data resources, an additional cross-dataset experiment is conducted, using Dataset #2 as the source dataset for self-supervised pretraining, followed by downstream adaptation on Dataset #1 and Dataset #3. As shown in **Fig. 8**, the predicted RUL values on both target datasets remain closely aligned with the diagonal, indicating that the learned representation preserves strong transferability when the source domain is replaced by a public real-world dataset. The quantitative results in **Table 6** further support this observation. On the source Dataset #2, the proposed model achieves RMSE values of 25, 14, and 50 cycles for the mean, median, and 95th percentile, respectively. After transfer to Dataset #1, the corresponding RMSE values are 32, 20, and 63 cycles, indicating only a limited degradation in predictive accuracy. For Dataset #3, the RMSE values increase to 75, 39, and 158 cycles, which is mainly attributable to the larger fleet size and greater heterogeneity in operating conditions. Even so, the overall prediction trend remains stable, and the absolute error distribution remains concentrated within a reasonable range for most samples. These results provide additional evidence that the proposed framework generalizes across independent real-world fleets under different data availability settings, and strengthen the reproducibility of the experimental validation by using a publicly available dataset.

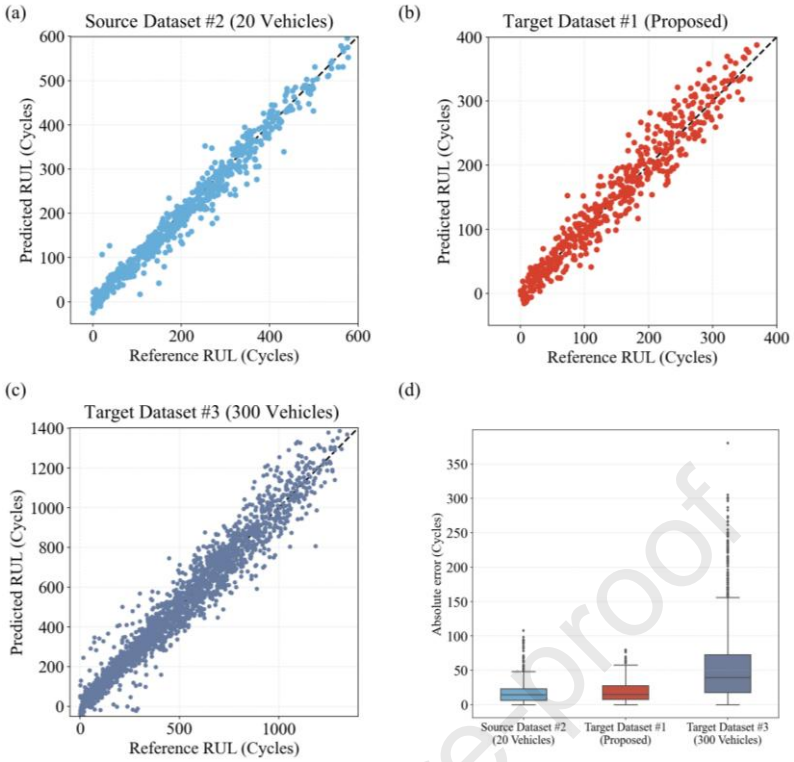


Fig. 8. Cross-dataset RUL prediction results and absolute error distribution of the proposed model using public Dataset #2 as the source dataset.

Table 6. Evaluation of cross-dataset generalization performance of the proposed model using public Dataset #2 as the source dataset.

Dataset	RMSE (Mean)	RMSE (Median)	RMSE (95th perc.)
Source Dataset #2	25	14	50
Target Dataset #1	32	20	63
Target Dataset #3	75	39	158

4.4. Ablation study of the proposed framework

An ablation study is conducted to quantify the contribution of key components in the proposed framework and to attribute the observed performance gains in real-world RUL prediction. **Fig. 9**(a)-(e) presents scatter plots of predicted versus reference RUL for the proposed method, an ablation variant without the contrastive objective, and three baseline models, including autoencoder (AE)-LSTM, VAE-MLP, and LSTM-Only. The proposed method exhibits the strongest alignment along the diagonal, indicating high consistency between predicted and reference RUL across the entire lifecycle. When the contrastive learning objective is removed, the dispersion around the diagonal increases noticeably, particularly in the medium- and long-horizon RUL regions. The baseline models show progressively larger deviations, with LSTM-Only

displaying the most pronounced scatter, suggesting limited capability to capture long-term degradation dynamics from raw temporal sequences alone.

The quantitative comparison in **Table 7** further substantiates these observations. The proposed method achieves the lowest RMSE across all robustness-oriented statistics, with an RMSE (mean) of 27 cycles, a median RMSE of 15 cycles, and a 95th-percentile RMSE of 56 cycles. When the contrastive objective is removed, the RMSE increases consistently to 34 cycles (mean), 21 cycles (median), and 72 cycles (95th percentile), indicating a clear degradation in both typical performance and tail-risk robustness. The reconstruction-based baselines, AE-LSTM and VAE-MLP, yield substantially higher RMSE levels, with RMSE (mean) values of 56 and 66 cycles, respectively. In addition, LSTM-Only achieves an RMSE (mean) of 88 cycles, with a median RMSE of 55 cycles and a 95th percentile RMSE of 175 cycles. Notably, integrating the contrastive learning objective yields a 20.6 percent reduction in mean RMSE compared to the pure reconstruction ablation variant. These results indicate that neither pure sequence modeling nor reconstruction-driven representation learning alone is sufficiently robust for accurate RUL prediction under complex real-world conditions. In contrast, the proposed method achieves more stable and reliable predictions. Furthermore, the observed performance gap between the proposed method and the standard VAE-MLP and AE-LSTM variants underscores the importance of incorporating both a contrastive learning objective and a temporal aggregation mechanism for real-world EV battery RUL prediction.

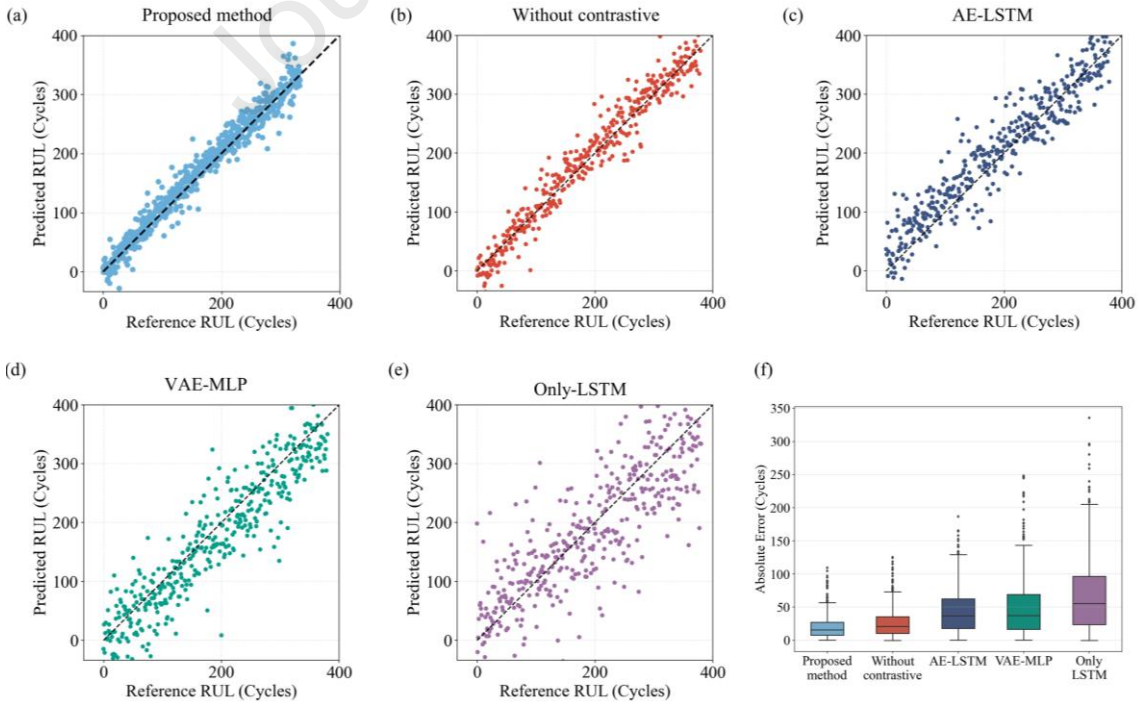


Fig. 9. Performance comparison with baseline methods and ablation study.

Table 7. Performance comparison of RUL prediction accuracy among the proposed method, ablation variant, and baseline models.

Methods	RMSE (Mean)	RMSE (Median)	RMSE (95th perc.)	Improvement (Mean)
Proposed method	27	15	56	-
Without contrastive	34	21	72	-20.6%
AE-LSTM	56	37	110	-51.8%
VAE-MLP	66	37	129	-59.1%
LSTM-Only	88	55	175	-69.3%

4.5. Method comparison and parameter sensitivity

To further validate the effectiveness of the proposed method, a comparative study is conducted on Dataset #1 against three state-of-the-art methods, namely battery masked autoencoders (BMAE), semi-supervised representation learning (SSRL), and domain adversarial learning (DAL) [57-59]. As shown in **Fig. 10**(a)-(d), the predictions of the proposed method are distributed more closely around the diagonal reference line than those of the three competing methods, indicating a higher consistency between the predicted RUL and the reference RUL over the entire prediction range. By comparison, BMAE exhibits greater dispersion and more outliers; SSRL shows a more pronounced deviation from the diagonal line, especially in the high RUL region; and DAL presents a clear systematic bias, particularly for samples with low and medium reference RUL values. These results indicate that the proposed method provides more accurate and more stable RUL prediction. The quantitative results in **Table 8** further support this observation. The proposed method achieves the lowest error across all three statistical criteria, with a mean RMSE of 27 cycles, a median RMSE of 15 cycles, and a 95th-percentile RMSE of 56 cycles. In contrast, BMAE yields 45, 23, and 92 cycles; SSRL yields 61, 58, and 88 cycles; and DAL yields 86, 69, and 157 cycles. Moreover, the boxplot in **Fig. 10**(e) shows that the proposed method has the narrowest error distribution and the lowest median absolute error among all compared methods, further demonstrating its greater robustness and better ability to control large prediction deviations. Overall, the results show that the proposed method achieves a more reliable balance between prediction accuracy and prediction stability than the three representative comparison methods.

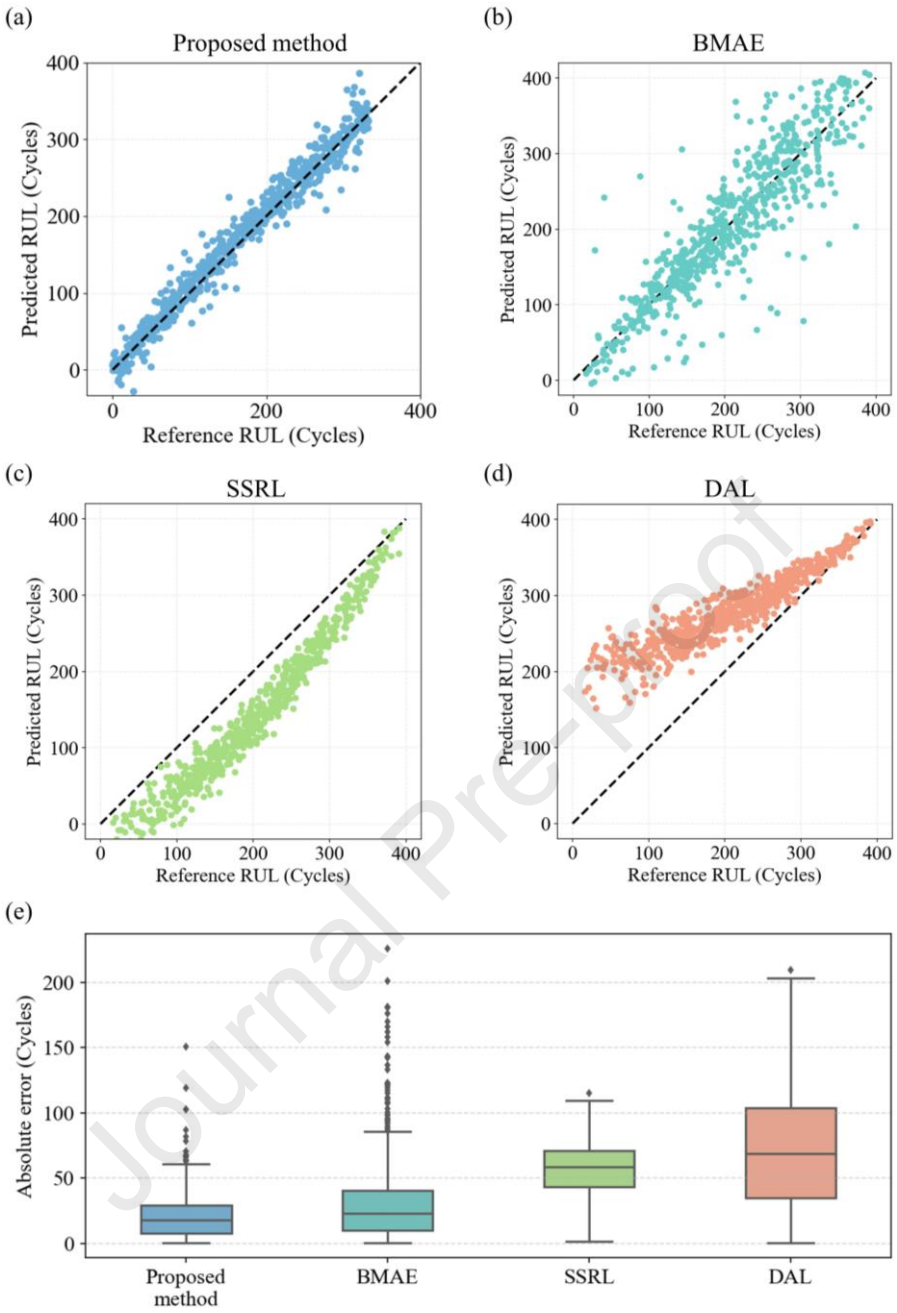


Fig. 10. Comparison of the proposed method with state-of-the-art methods.

Table 8. Quantitative comparison of the proposed method with state-of-the-art methods.

Methods	RMSE (Mean)	RMSE (Median)	RMSE (95th perc.)
Proposed method	27	15	56
BMAE [57]	45	23	92
SSRL [58]	61	58	88
DAL [59]	86	69	157

Since the temporal window determines how much historical degradation information is involved in each prediction, its influence on model performance is further investigated. **Table 9** reports the sensitivity

results of the downstream fine-tuning stage under different window sizes. As the window size increases from 5 to 20 cycles, both prediction accuracy and inference efficiency change noticeably. A window size of 5 cycles gives the lowest computational cost, with 4.042 G FLOPs, 0.477 ms per sample, and 1916 MB peak GPU memory, while the mean RMSE is 35 cycles. When the window size increases to 10 and 15 cycles, the mean RMSE decreases to 27 cycles, showing that a moderate temporal context helps the model capture degradation information more effectively. When the window size is further increased to 20 cycles, the mean RMSE rises to 42 cycles, accompanied by a substantial increase in computational burden to 14.14 G FLOPs, 1.027 ms per sample, and 7489 MB peak GPU memory. Based on these results, the window size is set to 10 cycles in this study. Although the settings of 10 and 15 cycles achieve the same mean RMSE, a window size of 10 cycles requires fewer inference FLOPs, shorter inference time, and lower peak GPU memory, at 7.409 G, 0.598 ms per sample, and 3764 MB, respectively. This indicates that a window size of 10 provides a more suitable balance between smoothing local fluctuations and preserving informative degradation dynamics.

Table 9. Prediction accuracy and inference efficiency of the proposed model under different window sizes.

Window size	RMSE (Mean)	Inference FLOPs (G)	Inference time (ms/sample)	Inference GPU peak allocated (MB)
5	35	4.042	0.477	1916
10	27	7.409	0.598	3764
15	27	10.77	0.825	5622
20	42	14.14	1.027	7489

Table 10 further presents the sensitivity of the proposed model to interpolation length during temporal normalization. As the interpolation length increases from 64 to 256, the mean RMSE gradually decreases from 42 to 27, indicating that a finer temporal resolution preserves richer aging related electrochemical information in the retained charging segments. By comparison, shorter interpolated sequences are associated with relatively large prediction errors, suggesting that coarse temporal representations are insufficient to capture subtle degradation characteristics. When the interpolation length is further increased to 384, the mean RMSE rises to 31, accompanied by a clear increase in inference cost. This trend indicates that the gain from increasing temporal resolution becomes limited beyond a certain point, whereas the computational burden continues to grow. Accordingly, the interpolation length is set to 256 in this study, since this setting yields the lowest prediction error and provides a favorable balance between electrochemical feature preservation and computational efficiency.

Table 10. Prediction accuracy and efficiency of the proposed model under different interpolation lengths.

Interpolation length	RMSE (Mean)	Inference FLOPs (G)	Inference time (ms/sample)	Inference GPU peak allocated (MB)
64	42	1.853	0.1495	999
128	38	3.705	0.387	1925
192	33	5.560	0.412	2849
256	27	7.409	0.598	3764
384	31	11.113	1.065	5303

4.6. Computational and labeling cost evaluation

High-quality RUL prediction for EV batteries relies on accurate health labels, yet acquiring such labels through controlled aging evaluation remains expensive and time-consuming in practice. Prior studies have shown that conventional battery health evaluation procedures typically involve complete discharge and recharge cycles conducted in workshop environments, often exceeding several hours per vehicle and requiring dedicated testing infrastructure and personnel resources [47]. Additional cost analyses further indicate that expenses associated with testing equipment, thermal chambers, labor, and battery handling scale almost linearly with the number of labeled vehicles, making large-scale supervised labeling economically prohibitive [60]. Motivated by these observations, a cost model for RUL label construction is formulated by integrating the reported cost structures with the practical labeling strategy adopted in this study [47,60]. The cost model represents an order-of-magnitude estimation rather than an exact accounting model. It assumes independent testing of labeled batteries under controlled aging, without parallel efficiency gains from large-scale infrastructure. Therefore, the total cost scales approximately linearly with the number of labeled batteries. The total cost for RUL labeling is presented by Eqn. (15):

$$C_{\mathrm{RUL}} = N_{\mathrm{lab}} \times \left[T_{\mathrm{test}} \cdot \left(C_{\mathrm{cyc}} + C_{\mathrm{temp}} \right) + H_{\mathrm{lab}} \cdot C_{\mathrm{lab}} + C_{\mathrm{set}} \right] \tag{15}$$

where C_{RUL} denotes the total cost of RUL label construction, N_{lab} is the number of vehicles receiving explicit RUL supervision, T_{test} represents the duration of one controlled aging evaluation, C_{cyc} and C_{temp} denote the unit-time costs of battery cyclers and temperature chambers, respectively, H_{lab} is the required engineering labor time per vehicle, C_{lab} is the labor cost per unit time, and C_{set} accounts for fixed setup, disassembly, and management costs per vehicle.

Based on Eqn. (15), the cost–performance trade-off of different labeling strategies is evaluated in **Fig. 11**. As shown in **Fig. 11**(a), prediction accuracy improves rapidly as labeling cost increases in the low-cost

regime, but the marginal performance gain diminishes beyond a certain point. First, the per-vehicle labeling cost is directly quantified by the term inside the brackets of Eqn. (15). Under the cost configuration used in this study, the RUL supervision cost per labeled vehicle is $10,000, and total labeling cost therefore scales linearly with the number of labeled vehicles. Consequently, the baseline scenario using 100% labeled data incurs approximately $100,000, while the proposed method using only 30% labeled data requires only $30,000, representing a 70% reduction in cost. This reduction is achieved without sacrificing accuracy: using labels from 30% of vehicles, the model attains an RMSE of 27 cycles, compared with 18 cycles under full supervision. **Fig. 11**(b) further corroborates this scaling behavior by decomposing total labeling costs into equipment rental, engineering labor, and setup-related expenses, highlighting the substantial savings of the proposed approach.

On the other hand, an accuracy-driven labeling requirement can be derived from the trade-off curve in **Fig. 11**(a) together with the RMSE distributions in **Fig. 11**(c). When a target error level of around 30 cycles is considered acceptable for fleet-level RUL management, the required labeled ratio is approximately 30% under the proposed framework, which corresponds to 30% labeled vehicles and a cost of about $30,000. Increasing the labeled ratio beyond 30% produces smaller accuracy gains. The RMSE improvement from 30% to 100% supervision takes about 9 cycles, while the labeling cost increases by about $ 70,000. This observation indicates that the marginal cost per additional unit of accuracy increases sharply after the 30% labeled regime. The robustness gain is also substantial in the low-label regime, but the reduction in variability becomes less pronounced after the labeled ratio reaches around 30%. Therefore, 30% labeling emerges as a practical operating point that balances three objectives simultaneously: the labeling cost remains low at $30,000, the mean accuracy remains close to the fully supervised settings, and the worst-case error is significantly reduced compared with the 5% to 20% regimes.

In addition to labeling cost, the computational overhead of the proposed framework is quantitatively analyzed in **Table 11** by comparing the pre-training, fine-tuning, and inference stages. Pre-training incurs the highest computational cost, with 43.08 GFLOPs, 2.72 million trainable parameters, and a total runtime of 2356 sec, reflecting the one-time expense of learning general degradation representations from large-scale unlabeled data. In contrast, the fine-tuning stage demonstrates substantially lower computational demand. The FLOPs decrease to 21.54 GFLOPs, and the total runtime drops to 827 sec, with a reduced GPU memory requirement of 2132 MB. This computational profile confirms that adapting the model to a new vehicle fleet

with limited labeled data introduces only moderate overhead and is feasible under typical industrial computing constraints. The inference stage exhibits the lowest computational burden among all stages. No trainable parameters are involved, and the FLOPs further decrease to 7.409 GFLOPs, with a total execution time of 42.3 sec and a peak GPU memory usage of 3764 MB. These results indicate that offloading heavy computation to an offline pre-training stage enables cost-effective deployment of RUL models across large EV fleets under limited labeled data and constrained computational resources.

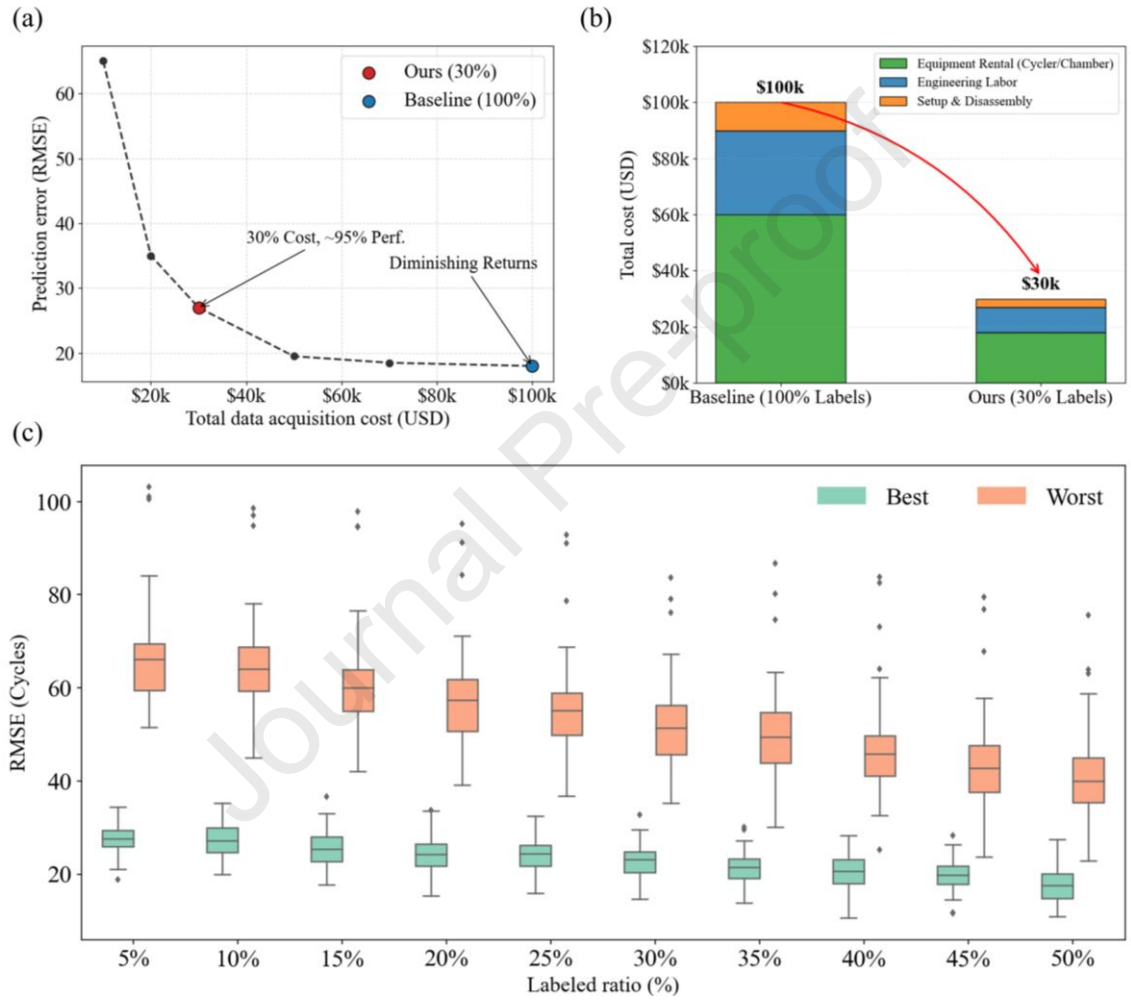


Fig. 11. Cost-benefit analysis of data labeling and model performance.

Table 11. Computational and performance metrics across pre-training, fine-tuning, and inference.

Metric	Pre-training	Fine-tuning	Inference
Epochs	150	50	0
FLOPs (G)	43.08	21.54	7.409
Trainable params (M)	2.72	1.41	0
Total time (sec)	2,356	827	42.3
GPU peak allocated (MB)	4,102	4152	3764

5. Conclusions

Reliable RUL prognostics for batteries are critical to the safe, economical, and sustainable operation of EVs. However, many existing methods are constrained by the mismatch between abundant operational data and scarce labels, a limitation amplified by complex, heterogeneous usage patterns. Bridging this gap is essential for scalable fleet-level health management and maintenance planning. This study advances a data-efficient learning paradigm that leverages large-scale unlabeled operational data to build degradation-relevant representations, thereby improving the accuracy and robustness of RUL prediction under real-world operating conditions.

Comprehensive evaluations conducted on three real-world EV datasets demonstrate the accuracy, robustness, and scalability of the proposed framework. On the heterogeneous source fleet (Dataset #1), the model achieves an RMSE of 27 cycles, outperforming supervised and semi-supervised baselines. Cross-fleet transfer experiments further show that stable generalizability performance is maintained when the pretrained model is adapted to target fleets using labeled data from only 30% of vehicles, indicating that the learned representations are transferable across variations in vehicle usage, charging behaviors, and battery configurations. In particular, approximately 95% of fully supervised prediction accuracy is retained while reducing the RUL labeling cost by about 70%, highlighting a favorable accuracy–cost trade-off for large-scale fleet applications. In addition, uncertainty-aware trajectory analyses and post-hoc interpretation consistently indicate that the learned latent representations and temporal aggregation capture degradation-relevant dynamics in real-world operation, thereby supporting transparent and reliable prognostic decision-making.

Despite these promising results, several limitations remain. The current framework is primarily validated on LIB systems, while its computational efficiency and real-time deployment capability on resource-constrained battery management systems have not yet been fully explored. Future work will extend the proposed framework to a broader range of battery chemistries and pack configurations, including emerging systems such as high-nickel and sodium-ion batteries, to further examine cross-chemistry transferability. Efforts will also focus on improving computational efficiency to support real-time inference on embedded BMS hardware and on integrating user behavior modeling and maintenance decision support, enabling more proactive and personalized battery health management in next-generation EV fleets.

Declaration of Competing Interest

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

Acknowledgement

This work is supported by the National Key Research and Development Program of China (2024YFE0115800) and the Department of Science and Technology of Guangdong Province (2023ZT10L145). All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of sponsors.

Data availability

Dataset #1 is confidential, and the authors do not have permission to share the data. Dataset #2 contains data for 20 EVs and is openly available at https://github.com/BatICM/battery-charging-data-of-on-road-electric-vehicles. Dataset #3 includes data from 300 EVs and is available at http://ivstskl.changan.com.cn/?p=2697.

Code availability

For access to the code associated with the deep learning developed in this study, please reach out to the corresponding author.

References

[1] Dunn, B., Kamath, H., & Tarascon, J. M. (2011). Electrical energy storage for the grid: a battery of choices. Science, 334(6058), 928-935.

[2] Schmuch, R., Wagner, R., Hörpel, G., Placke, T., & Winter, M. (2018). Performance and cost of materials for lithium-based rechargeable automotive batteries. Nature Energy, 3(4), 267-278.

[3] Palacín, M. R., & de Guibert, A. (2016). Why do batteries fail? Science, 351(6273), 1253292.

[4] Liu, K., Liu, Y., Lin, D., Pei, A., & Cui, Y. (2018). Materials for lithium-ion battery safety. Science Advances, 4(6), eaas9820.

[5] Zhao, J., Lv, Z., Li, D., Feng, X., Wang, Z., Wu, Y., ... & Burke, A. F. (2024). Battery engineering safety technologies (BEST): M5 framework of mechanisms, modes, metrics, modeling, and mitigation. eTransportation, 22, 100364.

[6] Qu, X., Shi, D., Zhao, J., Tran, M. K., Wang, Z., Fowler, M., ... & Burke, A. F. (2024). Insights and reviews on battery lifetime prediction from research to practice. Journal of Energy Chemistry, 94, 716-739.

[7] Jafari, M., Gauchia, A., Zhang, K., & Gauchia, L. (2015). Simulation and analysis of the effect of real-world driving styles in an EV battery performance and aging. IEEE Transactions on Transportation Electrification, 1(4), 391-401.

[8] Bamdezh, M. A., & Molaeimanesh, G. R. (2024). Aging behavior of an electric vehicle battery system considering real drive conditions. Energy Conversion and Management, 304, 118213.

[9] Qi, H., Ou, S. S., Jia, Y. H., Li, Z., & Lin, Y. (2026). Cross-temporal framework for driving behavior impact on electric vehicle battery health. Communications in Transportation Research, 6(1), 9640013.

[10] Xu, Z., Wang, J., Lund, P. D., & Zhang, Y. (2022). Co-estimating the state of charge and health of lithium batteries through combining a minimalist electrochemical model and an equivalent circuit model. Energy, 240, 122815.

[11] Xiong, R., Li, L., Li, Z., Yu, Q., & Mu, H. (2018). An electrochemical model based degradation state identification method of lithium-ion battery for all-climate electric vehicles application. Applied Energy, 219, 264-275.

[12] Liu, W., Liu, P., & Mitlin, D. (2020). Review of emerging concepts in SEI analysis and artificial SEI membranes for lithium, sodium, and potassium metal battery anodes. Advanced Energy Materials, 10(43), 2002297.

[13] Li, J., Landers, R. G., & Park, J. (2020). A comprehensive single-particle-degradation model for battery state-of-health prediction. Journal of Power Sources, 456, 227950.

[14] Liu, X., Ai, W., Marlow, M. N., Patel, Y., & Wu, B. (2019). The effect of cell-to-cell variations and thermal gradients on the performance and degradation of lithium-ion battery packs. Applied Energy, 248, 489-499.

[15] Lu, J., Xiong, R., Tian, J., Wang, C., Hsu, C. W., Tsou, N. T., ... & Li, J. (2022). Battery degradation prediction against uncertain future conditions with recurrent neural network enabled deep learning. Energy Storage Materials, 50, 139-151.

[16] Khaleghi, S., Karimi, D., Beheshti, S. H., Hosen, M. S., Behi, H., Berecibar, M., & Van Mierlo, J. (2021). Online health diagnosis of lithium-ion batteries based on nonlinear autoregressive neural network. Applied Energy, 282, 116159.

[17] Wang, F., Zhai, Z., Zhao, Z., Di, Y., & Chen, X. (2024). Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis. Nature Communications, 15(1), 4332.

[18] Aitio, A., & Howey, D. A. (2021). Predicting battery end of life from solar off-grid system field data using machine learning. Joule, 5(12), 3204-3220.

[19] . Li, J., Deng, Z., Che, Y., Xie, Y., Hu, X., & Teodorescu, R. (2023). Degradation pattern recognition and features extrapolation for battery capacity trajectory prediction. IEEE Transactions on Transportation Electrification, 10(3), 7565-7579.

[20] Zhao, J., & Wang, Z. (2024). Specialized convolutional transformer networks for estimating battery health via transfer learning. Energy Storage Materials, 71, 103668.

[21] Weng, A., Dufek, E., & Stefanopoulou, A. (2023). Battery passports for promoting electric vehicle resale and repurposing. Joule, 7(5), 837-842.

[22] Chen, S. Z., Liang, Z., Yuan, H., Yang, L., Xu, F., & Fan, Y. (2023). A novel state of health estimation method for lithium-ion batteries based on constant-voltage charging partial data and convolutional neural network. Energy, 283, 129103.

[23] Zhang, C., Luo, L., Yang, Z., Du, B., Zhou, Z., Wu, J., & Chen, L. (2024). Flexible method for estimating the state of health of lithium-ion batteries using partial charging segments. Energy, 295, 131009.

[24] Zhang, C., Tu, L., Yang, Z., Du, B., Zhou, Z., Wu, J., & Chen, L. (2025). A CMMOG-based lithium-battery SOH estimation method using multi-task learning framework. Journal of Energy Storage, 107, 114884.

[25] Zhao, J., Qu, X., Li, Y., Nan, J., & Burke, A. F. (2025). Real-time prediction of battery remaining useful life using hybrid-fusion deep neural networks. Energy, 328, 136618.

[26] Li, X., Yuan, C., Li, X., & Wang, Z. (2020). State of health estimation for Li-ion battery using incremental capacity analysis and Gaussian process regression. Energy, 190, 116467.

[27] Pan, R., Liu, T., Huang, W., Wang, Y., Yang, D., & Chen, J. (2023). State of health estimation for lithium-ion batteries based on two-stage features extraction and gradient boosting decision tree. Energy, 285, 129460.

[28] Yang, D., Wang, Y., Pan, R., Chen, R., & Chen, Z. (2018). State-of-health estimation for the lithium-ion battery based on support vector regression. Applied Energy, 227, 273-283.

[29] Severson, K. A., Attia, P. M., Jin, N., Perkins, N., Jiang, B., Yang, Z., ... & Braatz, R. D. (2019). Data-driven prediction of battery cycle life before capacity degradation. Nature Energy, 4(5), 383-391.

[30] Pozzato, G., Allam, A., Pulvirenti, L., Negoita, G. A., Paxton, W. A., & Onori, S. (2023). Analysis and key findings from real-world electric vehicle field data. Joule, 7(9), 2035-2053.

[31] Steininger, V., Rumpf, K., Hüsson, P., Li, W., & Sauer, D. U. (2023). Automated feature extraction to integrate field and laboratory data for aging diagnosis of automotive lithium-ion batteries. Cell Reports Physical Science, 4(10).

[32] Ma, G., Xu, S., Jiang, B., Cheng, C., Yang, X., Shen, Y., ... & Yuan, Y. (2022). Real-time personalized health status prediction of lithium-ion batteries using deep transfer learning. Energy & Environmental Science, 15(10), 4083-4094.

[33] Ma, J., Shang, P., Zou, X., Ma, N., Ding, Y., Sun, J., ... & Lin, Y. (2021). A hybrid transfer learning scheme for remaining useful life prediction and cycle life test optimization of different formulation Li-ion power batteries. Applied Energy, 282, 116167.

[34] Chen, J., Han, X., Sun, T., & Zheng, Y. (2024). Analysis and prediction of battery aging modes based on transfer learning. Applied Energy, 356, 122330.

[35] Zhao, J., Qu, X., Li, Y., Nan, J., & Burke, A. F. (2025). Real-time prediction of battery remaining useful life using hybrid-fusion deep neural networks. Energy, 136618.

[36] Guo, N., Chen, S., Tao, J., Liu, Y., Wan, J., & Li, X. (2024). Semi-supervised learning for explainable few-shot battery lifetime prediction. Joule, 8(6), 1820-1836.

[37] Li, W., Samsukha, H., van Vlijmen, B., Yan, L., Greenbank, S., Onori, S., & Viswanathan, V. (2025). Fast data augmentation for battery degradation prediction. Energy and AI, 100542.

[38] Song, J., Wang, H., Liu, Y., Wang, R., & Wang, K. (2025). Enhancing variational autoencoder for estimation of lithium-ion batteries state-of-health using impedance data. Energy, 138739.

[39] Sun, J., Gu, A., & Kainz, J. (2025). A solution framework for the experimental data shortage problem of lithium-ion batteries: Generative adversarial network-based data augmentation for battery state estimation. Journal of Energy Chemistry, 103, 476-497.

[40] Zhang, S., Liu, M., Guo, R., Tian, J., Man, Z., & Shen, W. (2026). Battery life prediction with scarce data using physics-informed data generation and adaptive autoencoder. IEEE Transactions on Transportation Electrification, 12(1), 1223-1234.

[41] Ma, L., Tian, J., Zhang, T., Guo, Q., & Chung, C. Y. (2025). Enhanced battery life prediction with reduced data demand via semi-supervised representation learning. Journal of Energy Chemistry, 101, 524-534.

[42] Zhang, S., Li, Y., Tian, J., Man, Z., Chung, C. Y., & Shen, W. (2024). Improving battery life prediction with unlabeled data: Confidence-weighted semi-supervised learning with label propagation. IEEE Transactions on Transportation Electrification, 11(2), 5938-5949.

[43] Hu, J., Weng, L., Gao, Z., & Yang, B. (2022). State of health estimation and remaining useful life prediction of electric vehicles based on real-world driving and charging data. IEEE Transactions on Vehicular Technology, 72(1), 382-394.

[44] Li, F., Feng, H., Min, Y., Zhang, Y., Zuo, H., Bai, F., & Zhang, Y. (2025). Prediction of lithium-ion battery degradation trajectory in electric vehicles under real-world scenarios. Energy, 317, 134663.

[45] Zhang, D., Wang, Z., Li, X., Liu, P., Sun, H., Wang, Q., ... & She, C. (2024). A battery degradation prediction framework considering differences in electric vehicle operating characteristics. IEEE Transactions on Transportation Electrification, 11(2), 5223-5236.

[46] Deng, Z., Xu, L., Liu, H., Hu, X., Duan, Z., & Xu, Y. (2023). Prognostics of battery capacity based on charging data and data-driven methods for on-road vehicles. Applied Energy, 339, 120954.

[47] Liu, H., Li, C., Hu, X., Li, J., Zhang, K., Xie, Y., ... & Song, Z. (2025). Multi-modal framework for battery state of health evaluation using open-source electric vehicle data. Nature Communications, 16(1), 1137.

[48] Wen, S., Lin, N., Huang, S., Li, X., Wang, Z., & Zhang, Z. (2024). Lithium battery state of health estimation using real-world vehicle data and an interpretable hybrid framework. Journal of Energy Storage, 96, 112623.

[49] Keil, P., & Jossen, A. (2016). Charging protocols for lithium-ion batteries and their impact on cycle life—An experimental study with different 18650 high-power cells. Journal of Energy Storage, 6, 125-141.

[50] Rumpf, K., Rheinfeld, A., Schindler, M., Keil, J., Schua, T., & Jossen, A. (2018). Influence of cell-to-cell variations on the inhomogeneity of lithium-ion battery modules. Journal of The Electrochemical Society, 165(11), A2587-A2607.

[51] Barré, A., Deguilhem, B., Grolleau, S., Gérard, M., Suard, F., & Riu, D. (2013). A review on lithium-ion battery ageing mechanisms and estimations for automotive applications. Journal of Power Sources, 241, 680-689.

[52] Jing, H., Hu, J., Ou, S. S., Lv, Z., Lyu, R., & Zhao, J. (2025). Scalable and generalizable deep learning for battery state of health estimation in on-road electric vehicles. Journal of Energy Chemistry.

[53] Xia, F., Yu, Y., & Chen, J. (2024). SOH and RUL prediction of lithium batteries based on fusions of RLOESS filtered electrochemical and thermal features by bidirectional gated recurrent unit network. Journal of Energy Storage, 102, 114134.

[54] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.

[55] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[56] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11).

[57] Chen, J., Wang, Y., Guo, D., Shen, Y., Sun, T., Han, X., ... & Ouyang, M. (2025). Deep learning model for remaining useful life prediction with reduced labeling data dependency. Applied Energy, 402, 126885.

[58] Ma, L., Tian, J., Zhang, T., Guo, Q., & Chung, C. Y. (2025). Enhanced battery life prediction with reduced data demand via semi-supervised representation learning. Journal of Energy Chemistry, 101, 524-534.

[59] Zhang, Z., Wang, Y., Ruan, X., & Zhang, X. (2025). Lithium-ion batteries lifetime early prediction using domain adversarial learning. Renewable and Sustainable Energy Reviews, 208, 115035.

[60] Rochester Institute of Technology. Testing pricing. https://www.rit.edu/batterydevelopment/testing-pricing. Accessed on 20 December 2025.

Highlights

- Self-supervised VAE-LSTM predicts EV battery RUL using unlabeled field data.
- Proposed framework achieves a low RMSE of 27 cycles on real-world EV fleets.
- Cross-fleet generalization on 340 EVs reduces RMSE by 42% compared to baselines.
- Interpretability analysis reveals principal degradation patterns in latent space.
- Using 30% labeled data retains 95% accuracy and reduces labeling cost by 70%.

Declaration of interests

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

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: