Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.085363

ARTICLE

A Physics-Aware and Interpretable Hybrid Ensemble Framework for Cross-Battery State-of-Health Estimation of Lithium-Ion Batteries

Sherif Said[1,2,*], **Yıldırım Özüpak**[3] **and Emrah Aslan**[4,5]

[1]College of Engineering, University of Sharjah, Sharjah, United Arab Emirates
[2]College of Engineering and Technology, American University of the Middle East, Egaila, Kuwait
[3]Silvan Vocational School, Dicle University, Diyarbakir, Turkey
[4]Department of Computer Engineering, Faculty of Engineering, Dicle University, Diyarbakır, Turkey
[5]Department of Computer Engineering, Faculty of Engineering and Architecture, Mardin Artuklu University, Mardin, Turkey

*Corresponding Author: Sherif Said. Email: ssaid@sharjah.ac.ae

Received: 09 May 2026; Accepted: 02 July 2026; Published: 21 October 2026

ABSTRACT: Accurate State of Health (SoH) estimation is essential for safe and reliable lithium-ion battery management systems. However, many data-driven SoH models are evaluated using time-series splits from the same battery cell, which may overestimate prediction performance and fail to demonstrate generalization to unseen cells. To address this limitation, this study proposes a physics-aware and interpretable Gated Recurrent Unit (GRU)-Extreme Gradient Boosting (XGBoost) ensemble framework for cross-battery SoH estimation. The proposed model first employs a Gated Recurrent Unit (GRU) network to extract temporal degradation representations from sequential cycling data. These latent features are then transferred to an XGBoost regressor to capture nonlinear feature interactions. To improve robustness, a dynamically weighted ensemble strategy is introduced by combining the standalone GRU and GRU–XGBoost outputs according to validation confidence. In addition, physically motivated constraints, including degradation monotonicity, internal resistance growth consistency, and capacity-regeneration tolerance, are incorporated into the training and post-evaluation procedures to reduce physically implausible predictions. The framework is evaluated on the National Aeronautics and Space Administration (NASA) lithium-ion battery aging dataset using both time-aware and leave-one-cell-out cross-battery validation protocols. The experimental design explicitly prevents normalization leakage by estimating scaling parameters only from the training batteries. Comparative experiments include Random Forest, XGBoost, GRU, Temporal Convolutional Network (TCN), Transformer, and the proposed hybrid variants. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) analyses are further used to map influential features to electrochemical aging indicators such as capacity fade, impedance growth, voltage response, and temperature-dependent degradation. The results demonstrate that the proposed framework improves prediction accuracy, robustness, and interpretability while providing a more rigorous evaluation of cross-cell generalization for practical battery health monitoring.

KEYWORDS: SoH lithium-ion batteries; hybrid learning; ensemble modeling; explainable artificial intelligence

1 Introduction

Lithium-ion batteries are very important for modern energy storage systems, especially for electric cars, integrating renewable energy, and portable electronics. For these systems to work well, be safe, and be cost-effective, it's very important to keep an eye on the health of the batteries. The State of SoH is one of the most important health indicators because it shows how much a battery has degraded from its original

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

state. Reliable SoH estimation is essential for battery management systems (BMS) to enable safe operation, optimize energy usage, extend battery lifetime, and support second-life applications [1].

From an electrochemical perspective, lithium-ion battery aging is governed by coupled degradation mechanisms such as loss of lithium inventory, loss of active material, solid electrolyte interphase (SEI) growth, lithium plating, electrolyte decomposition, and internal resistance increase. These mechanisms are strongly affected by temperature, current rate, depth of discharge, rest periods, and charging/discharging protocols. Therefore, purely data-driven models may achieve low error on a single battery trajectory but may fail to generalize when the tested cell exhibits different aging kinetics. Recent studies have increasingly emphasized physics-informed machine learning, electrochemical–thermal coupling, and cross-battery validation to improve the reliability and transferability of SoH estimation models. Motivated by these developments, this study revises the original GRU–XGBoost ensemble into a physics-aware and cross-battery validated framework rather than treating SoH estimation as a simple curve-fitting problem.

It's still hard to get an accurate SoH estimate, even though it's important. This is because battery degradation processes are complicated and not linear. Lithium-ion batteries age in a very nonlinear way, and this is affected by how they are used, what conditions they are used in, changes in temperature, and the way the chemicals inside them work. Also, things like capacity regeneration after rest periods and sudden changes in the rate of degradation make modeling more difficult. These characteristics introduce strong temporal dependencies and stochastic variations, making SoH estimation significantly more complex than a conventional regression problem [2].

There are three main types of traditional methods for estimating SoH: physics-based, model-based, and data-driven. Physics-based models use electrochemical principles to explain how batteries age, while model-based methods often use equivalent circuit models or empirical formulas. These methods offer physical interpretability but often necessitate extensive parameter identification, specialized expertise, and significant computational resources, which constrains their adaptability to varying operational conditions. In contrast, data-driven approaches based on machine learning and deep learning have gained increasing attention due to their ability to learn complex nonlinear relationships directly from data without explicit physical modeling [3].

Recent research indicates that machine learning algorithms such as Random Forest, Support Vector Regression, and gradient boosting techniques, along with deep learning frameworks like recurrent neural networks and long short-term memory networks, demonstrate effective performance. But there are still some big problems with the way things are done now [4]. Tree-based models are powerful and effective; however, they frequently struggle to capture long-term temporal dependencies in battery degradation data [5]. Deep learning models can model sequential behavior, but people often criticize them for being black boxes, being expensive to run, and not being easy to understand [6]. Moreover, depending on just one modeling paradigm may not work well when there are different aging patterns and operating conditions [7]. In addition, recent research has emphasized that thermal behavior and operating temperature strongly affect lithium-ion battery safety, degradation rate, and long-term performance, indicating that reliable SoH estimation models should also consider temperature-dependent aging and thermal management effects [8].

Hybrid and ensemble learning strategies have emerged as effective solutions to these challenges. Hybrid frameworks can take the best parts of different ways to model things. For instance, they can use how well deep learning models can find temporal features and how well ensemble-based machine learning algorithms can do nonlinear regression. GRUs are a good way to learn temporal representations from sequential battery data. XGBoost is a powerful gradient boosting-based regression tool that can generalize well. An ensemble strategy that combines several complementary models can make predictions even more stable and reliable by reducing the biases and uncertainties of each model.

It is becoming more and more important to know how data-driven SoH estimation models work, especially in safety-critical applications like battery management systems. This is in addition to being able to make accurate predictions. Users have a hard time trusting predictions that don't come with an explanation, and it's hard to use them in real life. SHAP and LIME are two examples of explainable artificial intelligence (XAI) methods that can help you understand how models work on both a global and local level. SHAP helps you figure out which features have a big effect on the whole model and how they do it. LIME, on the other hand, explains how certain feature conditions affect individual predictions. Adding these kinds of tools that help people understand models makes them clearer and helps people make better decisions.

This study presents a hybrid and ensemble-based framework for estimating the State of SoH, which combines GRU-based temporal feature extraction with XGBoost regression, accompanied by a comprehensive interpretability analysis. The proposed method aims to achieve higher predictive accuracy, better temporal tracking, and more model transparency in complex situations of battery degradation. The main contributions of this study are summarized as follows:

- A revised GRU–XGBoost hybrid architecture is developed in which GRU-derived temporal degradation embeddings are used to enhance nonlinear SoH regression through XGBoost.
- Unlike conventional same-cell time-series splitting, the proposed framework is evaluated using both time-aware validation and leave-one-cell-out cross-battery validation to assess generalization to unseen battery cells.
- A leakage-safe preprocessing strategy is adopted. SoH normalization, feature scaling, and outlier filtering parameters are estimated exclusively from the training batteries and then applied to validation/test batteries.
- A dynamic ensemble weighting strategy is introduced to replace fixed empirical weighting. The weights assigned to GRU and GRU–XGBoost predictions are determined based on validation error and prediction confidence.
- Physics-aware constraints related to degradation monotonicity, internal resistance evolution, and capacity-regeneration tolerance are incorporated to reduce physically inconsistent predictions.
- The proposed model is benchmarked against stronger time-series baselines, including TCN and Transformer models, in addition to Random Forest, XGBoost, GRU, and the hybrid GRU–XGBoost model.
- SHAP and LIME explanations are mapped to physically meaningful battery aging indicators, including capacity fade, voltage slope, differential capacity with respect to voltage (dQ/dV) peak variation, temperature effect, and internal resistance growth.

The rest of this paper is set up like this. Section 2 looks at other work that has been done on estimating the SoH of batteries. Section 3 talks about the dataset, the suggested method, and the ways to measure success. Experimental results are shown in Section 4. Section 5 contains a discussion, and Section 6 wraps up the paper and talks about possible areas for future research.

2 Literature Review

Over the past ten years, the scientific literature has focused on figuring out the SoH of lithium-ion batteries. This is because it is so important for the reliability of battery management systems, safety, and lifespan. The aging processes of batteries are very complicated and don't follow a straight line. They depend on things like how they are used, how they break down chemically, and how long they last. A plethora of studies have been undertaken in the literature to model this complexity, encompassing physics-based models, empirical and equivalent circuit models, alongside data-driven methodologies utilizing machine learning and deep learning techniques. The growing availability of large-scale battery datasets and progress in computational intelligence have emphasized data-driven techniques that can analyze intricate degradation patterns without

requiring a specific electrochemical model. This section methodically analyzes contemporary and exemplary studies on SoH estimation, emphasizing machine learning, deep learning, hybrid frameworks, and ensemble methodologies. The review focused on emerging trends including model architectures, performance metrics, dataset utilization, temporal modeling, and XAI. The objective was to pinpoint essential research deficiencies in the literature and validate the theoretical basis of the proposed methodology.

Akbar et al. created a model that uses artificial intelligence, Internet of Things (IoT), and Big Data to predict battery SoH based on data. By comparing different machine learning methods, they got very accurate results and then used the model as a web service for real-time monitoring and reliability analysis [9]. Sylvestrin et al. performed a systematic review of 534 papers, revealing a trend towards public datasets (76% in 2023), with NASA's data being the most notable. Deep learning, especially Long Short-Term Memory (LSTM) (22%), is the most common way to estimate SoH. This shows how important transfer learning and IoT integration are for recycling batteries in a sustainable way [10]. Tang et al. created an encoder-decoder Recurrent Neural Network (RNN) framework for estimating SoH during multiple dynamic operations. They got an Root Mean Square Error (RMSE) below 0.68% by combining physical model identification with feature engineering. The model's decoder is very flexible, which makes it easy to get accurate estimates of cross-battery types by simply fine-tuning the weights [11]. Shibl et al. created a machine learning-based BMS for Unmanned Aerial Vehicles that used Deep Neural Network and LSTM to predict state of charge (SoC) and Random Forest to classify SoH. Their system, which used ESP32 and Raspberry Pi, was able to monitor in real time with high accuracy (SoC MSE 7.6E-4) when deployed in the cloud [12]. Using 12.83 million data points, Jing et al. came up with a hybrid TCN-iTransformer framework for estimating the SoH of Electric Vehicles (EVs) on the road. They created a 3D feature system with three levels: shallow, intermediate, and deep. They also used a differential attention mechanism to cut down on noise, which gave them an RMSE of 0.40% and kept the system stable even with 70% less training data [13]. Haraz et al. conducted an extensive review concentrating on machine learning applications for the estimation of SoC and SoH in lithium-ion batteries. They methodically examined input features, hyperparameters, and dataset quality across various battery chemistries, highlighting the significance of image processing and system modeling to address deficiencies in existing battery management research [14]. Mbagaya et al. examined Support Vector Regression (SVR), Random Forest, Convolutional Neural Network (CNN), and LSTM for SoH estimation, corroborating their findings using the NASA dataset. Their MATLAB simulations demonstrated that LSTM surpassed other models, whereas SVR encountered difficulties with sequential data. They said that deep learning is better for long-term degradation and that future research should focus on XAI and transfer learning [15]. Jafari and Byun created a fusion neural network that combined CNN, LSTM, and Convolutional LSTM architectures and used Particle Swarm Optimization (PSO) to make it better. Their method uses battery discharge curves to find time dependencies and SHAP to make models easier to understand. It gets high accuracy with an Mean Absolute Error (MAE) and RMSE [16].

Li and Chen developed three PSO-optimized models including LSTM, CNN, and SVR for SoH estimation. By selecting key features from voltage, current, and temperature curves using Pearson correlation, they found that PSO-LSTM provided the most robust performance across NASA Prognostics Center of Excellence (PCoE) datasets, demonstrating significant potential for real-world battery management systems [17]. Lee et al. developed an interpretable ML model for early-phase SoH estimation using a window-moving technique to extract statistical features. By applying SHAP, they identified that SoH degradation trends near the 100th cycle are more critical than absolute values for predicting future health [18]. Yu et al. developed a Bidirectional Gated Recurrent Unit-based charge capacity predictor and a general SVR-based model to address aging inconsistencies among cells within a battery pack. By utilizing a triplet of features comprising aging, inconsistency, and operating conditions, they achieved cell-level accurate SoH

tracking under various charging protocols [19]. Pandit and Ahlawat introduced a standardized one-to-many validation protocol using NASA datasets to compare XGBoost, Random Forest, and Support Vector Machine models. By applying unified hyperparameter optimization and cross-battery generalization, they established empirical benchmarks for BMS applications, identifying XGBoost as the superior model [20]. Kohtz et al. proposed a three-step physics-informed ML framework for SoH estimation. They utilized a finite element model to simulate SEI growth, fused the results with NASA experimental data to create a multi-fidelity model, and trained a Gaussian Process Regression to map partial charging voltage segments to SEI thickness [21]. Wu et al. proposed an end-to-end CNN— with attention model for SoH estimation using Electrochemical Impedance Spectroscopy (EIS) data. By integrating a convolutional neural network with an attention mechanism to form a multiview information perceptron, they achieved high accuracy across varying temperatures, outperforming traditional CNN and CNN-LSTM methods [22]. Faraji-Niri et al. developed an optimized Gaussian Process Regression model for accelerated SoH estimation using EIS data from 30 cells. Their study, focused on second-life applications, demonstrated model robustness against measurement noise and different physical test equipment while maintaining high prediction accuracy across varying conditions [23].

Recent studies on lithium-ion battery SoH estimation have increasingly shifted from purely data-driven prediction models toward physics-informed and mechanism-aware learning frameworks. Wang et al. proposed a physics-informed neural network for stable lithium-ion battery degradation modeling and prognosis, emphasizing that physical constraints can improve model robustness and degradation consistency [24]. Similarly, Ye et al. developed a physics-informed neural network-based SoH estimation method, demonstrating the importance of embedding battery degradation knowledge into neural prediction models [25]. From an electrochemical modeling perspective, Zhu et al. introduced a fractional-order lithium-ion battery model that explicitly considers SEI growth and lithium plating, which are among the key aging mechanisms responsible for capacity fade and internal resistance increase [26]. Lin et al. further extended physics-informed machine learning to SoH estimation under different temperatures and operating conditions, highlighting the need for models that can generalize beyond fixed laboratory cycling profiles [27]. In addition, Tang et al. proposed a physics-driven deep learning framework for SoH estimation under dynamic operating conditions, showing that the integration of physical knowledge and temporal learning can improve prediction reliability in practical battery management applications [11]. These studies indicate that reliable SoH estimation should consider not only prediction accuracy but also physical plausibility, electrochemical aging mechanisms, and cross-condition generalization.

3 Material and Method

This section outlines the materials, data preparation protocols, and methodological framework utilized in this study to assess the SoH of lithium-ion batteries. The document specifies the dataset used, the preprocessing methods applied to raw battery readings, the machine learning and deep learning models utilized for SoH prediction, and the recommended hybrid and ensemble learning strategies. The framework includes a clear explanation of the assessment protocol and interpretability methods so that you can fully understand the experimental methodology.

3.1 Dataset

The study utilizes the NASA Lithium-Ion Battery Dataset to assess battery deterioration behavior and the efficacy of SoH estimation [28]. The dataset comprises a sequence of controlled laboratory tests performed on lithium-ion batteries, which underwent multiple cycles of charging, discharging, and impedance measurements under various temperature settings. During cycling, batteries experience

accelerated electrochemical aging, resulting in capacity degradation and increased internal resistance. EIS measurements are conducted at designated frequencies to detect alterations in internal battery properties during the aging process. Each experiment persists until the battery attains the end-of-life (EOL) condition, delineated by a specified capacity threshold. The dataset is structured into batches of six trials and stored in MATLAB (.mat) format, with comprehensive experimental descriptions included in the supporting documentation. The dataset's primary properties are encapsulated in Table 1.

Table 1: Key characteristics of the NASA Li-ion battery dataset [28].

Feature	Description
Battery type	Lithium-ion batteries
Experiments	Batch of 6 independent battery experiments
Operating modes	Charge, discharge, and impedance measurements
Measurement technique	Electrochemical Impedance Spectroscopy (EIS)
Temperature conditions	Multiple controlled temperature levels
Aging mechanism	Repeated charge–discharge cycling
End-of-Life (EOL) criterion	Capacity degradation below threshold
Prediction targets	State of Health (SoH)

The battery cells and sample distribution used in the revised experiments are presented in Table 2. The final dataset used in this study consists of cycle-level discharge records extracted from four NASA cells. Each sample corresponds to one discharge cycle and includes capacity-related, voltage-related, current-related, temperature-related, and impedance-related indicators. Since the number of cycles differs among cells, all validation results are reported separately for each test battery and then averaged across folds.

Table 2: Battery cells and sample distribution used in the revised experiments.

Battery Cell	Nominal Capacity	Operating Profile	Number of Discharge Cycles	Role in Validation
B0005	2.0 Ah	CC–CV charge/discharge/EIS	168	Training/test depending on fold
B0006	2.0 Ah	CC–CV charge/discharge/EIS	168	Training/test depending on fold
B0007	2.0 Ah	CC–CV charge/discharge/EIS	168	Training/test depending on fold
B0018	2.0 Ah	CC–CV charge/discharge/EIS	132	Blind test depending on fold
Total	—	—	636	Cross-battery validation

Fig. 1 shows how the lithium-ion batteries studied degrade. Fig. 1a shows how the discharge capacity of three representative cells (B0005, B0006, and B0017) decreases over time as they are charged and discharged repeatedly. As you can see, the capacity profiles show a downward trend that isn't straight, getting closer and closer to the End-of-Life (EOL) threshold, which is defined as a 30% drop from the initial nominal capacity. Fig. 1b goes into more detail about the SoH evolution and shows the "capacity regeneration" phenomenon. This random recovery of capacity that happens after rest periods between duty cycles makes the SoH estimation process much more complicated. The fact that the cells age at different rates even though they are in the same operating conditions shows how important it is to have strong, data-driven frameworks like the hybrid GRU-based models proposed in this study to keep high estimation accuracy across different degradation patterns.

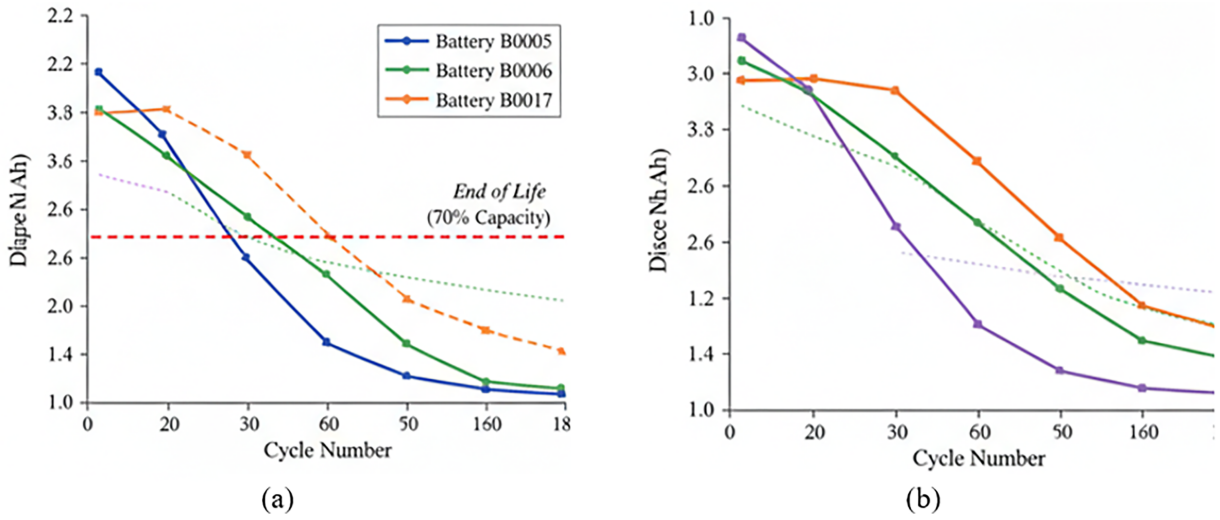


Figure 1: Performance and degradation analysis of the NASA PCoE battery dataset.

Data Preprocessing

The preprocessing procedure consists of five steps. First, discharge cycles with missing capacity, voltage, current, or temperature records were removed. Second, physically implausible measurements were excluded using predefined operating limits: negative capacity values, voltage values outside the valid discharge range, and incomplete discharge records were discarded. Third, cycle-level features were aligned according to the discharge cycle index to preserve temporal order. Fourth, outliers were detected using the interquartile range criterion. For each continuous feature x, samples outside $[Q_1 - 1.5 \cdot \mathrm{IQR}, Q_3 + 1.5 \cdot \mathrm{IQR}]$ were flagged and inspected. Outliers caused by measurement artifacts were removed, while physically meaningful capacity-regeneration points were retained. Fifth, all continuous input features were standardized using the mean and standard deviation computed only from the training batteries. The same scaling parameters were then applied to validation and test batteries.

This procedure improves data quality while preserving meaningful nonlinear degradation patterns, including local capacity recovery after rest periods. Importantly, no filtering, scaling, or normalization parameter was estimated using future cycles or unseen test-cell information. In this study, *SoH* is calculated using discharge capacity as:

$$SoH_{i,k} = \frac{Q_{i,k}}{Q_{i,0}^{ref}} \times 100 \tag{1}$$

where $Q_{i,k}$ denotes the measured discharge capacity of battery i at cycle k, and $Q_{i,0}^{ref}$ is the reference initial capacity. To avoid information leakage during cross-battery testing, the reference capacity used for normalization is not estimated from the test trajectory. For training batteries, $Q_{i,0}^{ref}$ is obtained from the first stable discharge cycle of the corresponding training cell. For unseen test batteries, two leakage-safe settings are considered: (i) using the nominal rated capacity of 2.0 Ah as a fixed reference, and (ii) using only the first available operational cycle as an online calibration point. No future capacity value from the test battery is used during normalization, scaling, model fitting, or hyperparameter selection.

To prevent target leakage, raw discharge capacity is not directly used as an input feature when the target SoH is computed from capacity. Instead, capacity-derived features are limited to lagged or early-cycle health indicators available before the predicted cycle. The main model input features include voltage statistics,

current statistics, temperature, impedance-derived internal resistance, voltage slope, incremental-capacity peak indicators, and cycle index. When capacity-related indicators are used, they are explicitly lagged as Q_{k-1}, ΔQ_{k-1}, or moving-window historical descriptors, ensuring that the target SoH_k is not directly reconstructed from the same-cycle capacity value.

3.2 Machine Learning

Machine learning is a set of computer methods that let systems learn patterns and relationships from data without having to be programmed to do so. These methods are often used for regression, classification, and forecasting problems, especially when there are complicated, nonlinear relationships between variables. A supervised learning model tries to find an unknown function by making the prediction error on training data as small as possible. It is possible to mathematically describe the learning process as

$$\hat{y} = f(X, \theta) \tag{2}$$

where X is the input feature matrix, θ denotes the model parameters, and $f(\cdot)$ represents the learned mapping. Model performance is typically evaluated using error metrics such as RMSE or MAE.

3.2.1 Random Forest

Random Forest is a type of ensemble learning algorithm that builds a lot of decision trees during training and combines their outputs to make a final prediction. Each tree learns from a bootstrap sample of the original dataset, and at each split, a random set of characteristics is chosen. This randomness makes models more diverse and less likely to overfit. For regression issues, the prediction is worked out as

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} h_i(X) \tag{3}$$

where h_i (X) is the output of the i-th tree. Random Forest is robust to noise and performs well with high-dimensional data.

3.2.2 XGBoost

Extreme Gradient Boosting (XGBoost) is a more advanced version of gradient boosting that builds models one at a time. Each new tree tries to fix the mistakes of the trees that came before it. Boosting, on the other hand, focuses on hard samples and makes predictions more accurate. For the SoH regression task, XGBoost minimizes a regularized squared-error objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i^{(t)}\right)^2 + \sum_{k=1}^{t}\Omega(f_k) \tag{4}$$

where y_i is the measured SoH, $\hat{y}_i^{(t)}$ is the prediction after adding the t-th regression tree, and f_k denotes the k-th tree. The regularization term is defined as

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda\sum_{j=1}^{T_k} w_j^2 \tag{5}$$

where T_k is the number of leaves in the tree, w_j is the score of leaf j, γ penalizes the number of leaves, and λ is the L2 regularization coefficient. In this study, the squared-error loss is used because SoH estimation is formulated as a continuous regression problem. Huber loss was also considered in sensitivity analysis because

of its robustness to outliers, but the final reported results are based on mean squared error for consistency with RMSE-based evaluation.

3.3 Gated Recurrent Unit

GRU is a type of recurrent neural network designed to model sequential and time-series data efficiently. Unlike traditional RNNs, GRU mitigates the vanishing gradient problem by introducing gating mechanisms that regulate information flow. The update gate z_t determines how much of the previous hidden state is retained, while the reset gate controls the contribution of past information. The GRU cell updates its hidden representation through reset and update gates. Given the input vector x_t and the previous hidden state h_{t-1}, the update gate z_t, reset gate r_t, candidate hidden state $\tilde{h}_t$, and final hidden state h_t are calculated as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{6}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{7}$$

$$\tilde{h}_t = \tanh\left(W_h x_t + U_h\left(r_t \odot h_{t-1}\right) + b_h\right) \tag{8}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid activation function, $\tanh(\cdot)$ is the hyperbolic tangent activation, $\odot$ denotes element-wise multiplication, and W, U, and b represent trainable weight matrices and bias vectors. The final hidden state sequence is used as a temporal degradation embedding for the subsequent XGBoost regression stage.

3.4 Hybrid GRU + XGBoost

The hybrid GRU + XGBoost model combines deep learning-based temporal feature extraction with strong ensemble-based prediction. In this design, the GRU network is first trained on sequential data to learn how to recognize patterns and dependencies that change over time. The hidden state outputs of the GRU layer show high-level latent features that summarize information that changes over time. The XGBoost model then uses these extracted features as inputs to do the final regression or classification task. This hybrid structure takes advantage of GRU's ability to model how things change over time and XGBoost's ability to learn how different features interact and relate to each other in complex ways. So, the model makes predictions more accurate, stable, and generalizable.

Fig. 2 illustrates the overall architecture of the proposed physics-aware and interpretable framework for lithium-ion battery SoH estimation. The framework uses leakage-safe input variables, including voltage and current statistics, temperature indicators, impedance-derived internal resistance, voltage-slope features, incremental-capacity peak descriptors, cycle index, and historical capacity-derived indicators. Since SoH is calculated from discharge capacity, the raw discharge capacity of the prediction cycle is not used as an input feature. Instead, only lagged or moving-window capacity descriptors derived from previous cycles are included where applicable to prevent target leakage.

Sequential degradation patterns are first learned by the GRU network, which generates latent temporal representations of battery aging. These representations are then transferred to the XGBoost regressor to capture nonlinear interactions among the extracted health indicators. The standalone GRU and hybrid GRU–XGBoost predictions are subsequently combined through a validation-confidence-based dynamic ensemble weighting strategy. Physics-aware constraints related to degradation monotonicity, internal resistance growth, physically admissible SoH limits, and tolerance to local capacity regeneration are incorporated to reduce physically implausible estimates.

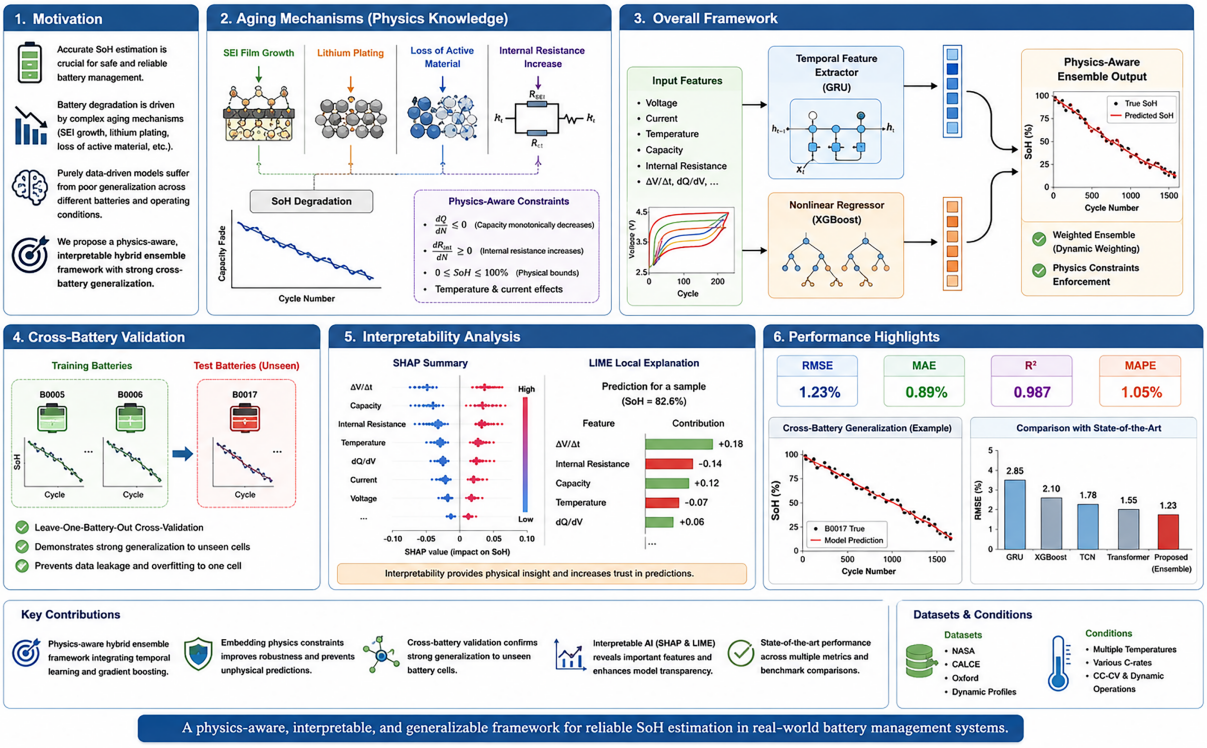


Figure 2: Overall architecture of the proposed leakage-safe, physics-aware, and interpretable hybrid ensemble framework for cross-battery lithium-ion battery SoH estimation.

The framework is evaluated using time-aware and leave-one-cell-out cross-battery validation protocols, ensuring that unseen test cells are excluded from feature scaling, model training, hyperparameter optimization, and ensemble-weight selection. Finally, SHAP and LIME methods are employed to associate model predictions with physically meaningful degradation indicators and to provide both global and instance-level explanations.

3.5 Physics-Aware Constraint Design

To reduce physically implausible predictions, physics-aware degradation constraints are incorporated into the revised framework. Although the proposed method remains primarily data-driven, three soft constraints are introduced to guide model behavior.

First, a monotonic degradation constraint is used because battery SoH generally decreases with cycling:

$$\mathcal{L}_{\text{mono}} = \frac{1}{N-1}\sum_{k=1}^{N-1} \max\left(0, \widehat{\text{SoH}}_{k+1} - \widehat{\text{SoH}}_{k} - \varepsilon_{\text{rec}}\right) \tag{10}$$

where ε_{rec} is a small tolerance term that permits local capacity regeneration after rest periods.

Second, an internal-resistance consistency constraint is introduced because internal resistance generally increases with aging:

$$\mathcal{L}_{R} = \frac{1}{N-1}\sum_{k=1}^{N-1} \max\left(0, -\left(\hat{R}_{k+1} - \hat{R}_{k}\right)\right) \tag{11}$$

where $\hat{R}_k$ denotes the impedance-derived internal resistance indicator.

Third, a bounded SoH constraint is applied to ensure physically meaningful predictions:

$$\mathcal{L}_{\text{bound}} = \frac{1}{N}\sum_{k=1}^{N}\left[\max(0, \widehat{\text{SoH}}_k - 100)^2 + \max(0, \text{SoH}_{\text{EOL}} - \widehat{\text{SoH}}_k)^2\right] \tag{12}$$

The final training objective for the GRU component is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{mono}} + \beta\mathcal{L}_R + \delta\mathcal{L}_{\text{bound}} + \lambda \parallel \Theta \parallel_2^2 \tag{13}$$

where $\mathcal{L}_{\text{MSE}}$ is the SoH regression loss, Θ denotes trainable GRU parameters, and α, β, δ, and λ are regularization coefficients. In this study, these physics-aware terms are used as soft constraints rather than strict rules because real battery capacity trajectories may exhibit local recovery effects.

3.6 Local Interpretable Model-Agnostic Explanations

LIME is an explanation technique designed to interpret individual predictions of complex black-box models. LIME works by generating perturbed samples around a specific instance and fitting a simple interpretable model locally. The objective function is defined as

$$\xi(x) = \arg\min_{g\in G} L(f, g, \pi_x) + \Omega(g) \tag{14}$$

where f is the original black-box model, g is the interpretable surrogate model, π_x defines the locality around instance x, and $\Omega(g)$ penalizes model complexity. LIME provides local faithfulness and human-understandable explanations.

3.7 SHapley Additive exPlanations

SHAP is an explanation framework based on cooperative game theory that assigns each feature a contribution value for a specific prediction. SHAP values represent the average marginal contribution of a feature across all possible feature subsets. The SHAP value for feature i is defined as

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[f(S\cup\{i\}) - f(S)] \tag{15}$$

where F is the full feature set. SHAP ensures properties such as consistency, local accuracy, and additivity, making explanations reliable and theoretically grounded.

3.8 Evaluation Metrics

MAE, RMSE, and R^2 are three common regression metrics used to measure how well the proposed models work. These metrics give us different information about how accurate the predictions are, how big the errors are, and how well the model fits the data. The MAE tells you how far off the predicted values are from the actual values on average. It treats all mistakes the same and gives a clear explanation in the same unit as the target variable. Lower MAE values mean that the predictions are more accurate. RMSE is the square root of the average of the squared prediction errors. RMSE is sensitive to outliers because it squares the errors, which makes it punish larger deviations more severely. It is especially helpful when you don't want to make big mistakes in your predictions. The R^2 measures how well the model explains the differences in

the target variable. A value of R^2 close to 1 means that the model explains things well, while a value close to or below 0 means that the model doesn't work well. Formulation in math:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{16}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{17}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{18}$$

where y_i represents the true value, $\hat{y}_i$ is the predicted value, $\bar{y}$ denotes the mean of the observed values, and n is the total number of samples.

3.9 Feature Extraction

We used feature extraction to better show how batteries lose power over time. Charge–discharge profiles were used to extract temporal degradation descriptors, including voltage evolution, current statistics, impedance-related indicators, cycle progression, and lagged capacity-based health descriptors. To avoid target leakage, the discharge capacity of the prediction cycle was not directly included as an input feature. Instead, only historical capacity information (e.g., previous-cycle capacity, lagged capacity ratios, or moving-window statistics) was used where applicable. The hybrid GRU–XGBoost model took high-level representations of hidden temporal features that the GRU network had learned and sent them to the XGBoost model.

Table 3 shows that the features we chose are meant to show important electrochemical and operational properties of lithium-ion batteries. Historical capacity-derived indicators (e.g., lagged capacity descriptors) reflect accumulated degradation without introducing target leakage, while voltage- and current-based metrics capture internal electrochemical dynamics and load conditions. In particular, internal resistance and the peak features of dQ/dV are very similar to electrochemical aging processes like the growth of the SEI layer and changes in phase. The model gets even better at taking into account how aging and the environment affect it when you add the cycle index and temperature. This physically informed feature design enhances prediction accuracy and model comprehension by integrating data-driven learning with battery degradation processes.

Table 3: Extracted features and their physical interpretations for SoH estimation.

Feature Name	Description	Physical Meaning
Lagged_capacity	Capacity at previous cycle (k–1)	Historical degradation indicator
Voltage_mean	Average voltage during discharge	Reflects internal resistance changes
Voltage_slope	Slope of discharge curve	Indicates aging dynamics
Current_mean	Average discharge current	Load condition indicator
Internal_resistance	Estimated via EIS	SEI layer growth
dQ/dV_peak	Incremental capacity peak	Phase transition indicator
Cycle_index	Charge–discharge cycle number	Aging progression
Temperature	Operating temperature	Degradation acceleration factor

3.10 Feature Scaling and Normalization

We performed feature scaling prior to training to ensure model stability and efficacy. We employed standard scaling techniques to ensure that all continuous input features were normalized to a uniform scale. Scaling parameters were computed solely on the training dataset and subsequently applied to the test dataset to prevent information leakage. This stage is crucial for models utilizing neural networks, such as GRU, due to their sensitivity to feature size.

3.11 Validation Strategy

To provide a rigorous evaluation of generalization performance, two validation protocols are adopted.

3.11.1 Time-Aware within-Cell Validation

In the first protocol, earlier cycles of each battery are used for training and later cycles are used for testing. This setting evaluates the ability of the model to track future degradation of a known battery cell. The temporal order of cycles is strictly preserved, and no random shuffling is applied.

3.11.2 Leave-One-Cell-Out Cross-Battery Validation

Since within-cell splitting may overestimate performance by allowing the model to learn the degradation trajectory of the same battery, a second and more stringent leave-one-cell-out protocol is introduced. In each fold, three batteries are used for training/validation, while the remaining battery is used as a completely unseen test cell. The protocol is summarized below. This protocol ensures that the test battery is not used during model training, feature scaling, hyperparameter tuning, ensemble-weight selection, or early stopping. Therefore, the reported cross-battery results reflect the ability of the model to generalize to unseen battery cells rather than merely fitting the degradation curve of a known cell. Table 4 shows the inter-battery validation protocol based on the single-cell exclusion method.

Table 4: Leave-one-cell-out cross-battery validation protocol.

Fold	Training Batteries	Validation Battery	Blind Test Battery	Purpose
Fold 1	B0006, B0007	B0018	B0005	Test generalization to B0005
Fold 2	B0005, B0007	B0018	B0006	Test generalization to B0006
Fold 3	B0005, B0006	B0018	B0007	Test generalization to B0007
Fold 4	B0005, B0006, B0007	Internal validation split	B0018	Test generalization to B0018

3.12 Model Training Procedure

All models were trained utilizing supervised learning techniques. We employed trial and error to identify the optimal hyperparameters for machine learning models that would achieve an ideal equilibrium between model complexity and generalization performance. We trained the GRU model across multiple epochs via gradient-based optimization. We trained the tree-based models using a predetermined quantity of estimators. We employed regularization techniques and early halting as necessary to prevent overfitting.

3.13 Hyperparameter Optimization

To reduce subjectivity in model configuration, hyperparameters were selected using Bayesian optimization on the validation set. The optimization objective was to minimize validation RMSE under the cross-battery validation protocol. The search space is summarized in Table 5.

Table 5: Hyperparameter search space.

Model	Hyperparameter	Search Range	Selected Value
GRU	Number of layers	1–3	2
GRU	Hidden units	32–128	64
GRU	Dropout rate	0.0–0.5	0.20
GRU	Learning rate	1e−4–1e−2	0.001
GRU	Batch size	16–64	32
GRU	Epochs	50–200	100
XGBoost	n_estimators	100–600	300
XGBoost	max_depth	3–10	6
XGBoost	learning_rate	0.01–0.20	0.05
XGBoost	subsample	0.60–1.00	0.80
XGBoost	colsample_bytree	0.60–1.00	0.80
XGBoost	reg_lambda	0.1–10	1.0

Early stopping was applied using validation loss with a patience of 10 epochs for the GRU model. For XGBoost, early stopping was applied based on validation RMSE. The same optimization budget and validation protocol were used for all comparable models to ensure fairness.

3.14 Dynamic Ensemble Weighting

The original ensemble used fixed empirical weights of 0.4 for GRU and 0.6 for GRU–XGBoost. To improve robustness across different degradation stages, this study introduces a dynamic weighting strategy based on validation confidence. Let $\hat{y}_{\mathrm{GRU}}$ and $\hat{y}_{\mathrm{Hybrid}}$ denote the predictions of the standalone GRU and GRU–XGBoost models. The final prediction is calculated as

$$\hat{y}_{\mathrm{ens}} = w_{\mathrm{GRU}}\hat{y}_{\mathrm{GRU}} + w_{\mathrm{Hybrid}}\hat{y}_{\mathrm{Hybrid}} \tag{19}$$

where

$$w_m = \frac{1/(\mathrm{RMSE}_m + \varepsilon)}{\sum_{j=1}^{M} 1/(\mathrm{RMSE}_j + \varepsilon)} \tag{20}$$

here, RMSE_m denotes the validation RMSE of model m, and ε is a small numerical constant. This strategy assigns higher weight to the model with lower validation error. In addition, stage-specific weights are estimated for early-life, middle-life, and near-EOL regions to account for changes in degradation nonlinearity. The degradation stages are defined as follows: early-life (SoH > 90%), middle-life (80% < SoH ≤ 90%), and near-EOL (SoH ≤ 80%). Table 6 shows the dynamic ensemble weighting throughout the stages of degradation. This dynamic strategy avoids assuming that a single static weighting scheme is optimal over the entire battery lifetime.

Table 6: Dynamic ensemble weighting across degradation stages.

Degradation Stage	SoH Range	GRU Weight	GRU–XGBoost Weight	Rationale
Early-life	SoH > 90%	0.35	0.65	XGBoost captures stable nonlinear feature interactions
Middle-life	80% < SoH ≤ 90%	0.45	0.55	Both temporal and nonlinear patterns are informative
Near-EOL	SoH ≤ 80%	0.55	0.45	GRU better tracks rapid temporal transitions

4 Experimental Results

This section looks at all of the proposed models for predicting SoH. We look at the experimental results from different angles, such as how accurate the predictions are, what kinds of errors they make, how well they can keep track of time, how well they can be measured, and how easy they are to understand.

4.1 Experimental Setup and Evaluation Protocol

Two evaluation protocols were used and are reported separately. The primary benchmark is leave-one-cell-out cross-battery validation, in which the test cell is completely unseen during feature scaling, hyperparameter selection, ensemble-weight selection, and model training. A secondary time-aware within-cell analysis was used for diagnostic visualizations, the repeated-seed statistical comparison, and the component ablation study, where earlier cycles of a known cell are used for training and later cycles are used for testing. For the proposed ensemble, the primary cross-battery MAE/RMSE are 1.71%/2.21%, whereas the secondary within-cell values are 1.40%/1.95%.

To strengthen the benchmark comparison, the proposed model was compared with conventional machine learning models, recurrent models, and recent high-performance time-series architectures. In addition to Random Forest, XGBoost, and GRU, Temporal Convolutional Network (TCN) and Transformer baselines were included. Table 7 presents the primary leave-one-cell-out cross-battery benchmark.

Table 7: Revised benchmark comparison under cross-battery validation.

Model	MAE (%)	RMSE (%)	R^2	±2% Accuracy	Remarks
Random Forest	3.18	4.05	0.812	0.52	Static nonlinear baseline
XGBoost	2.74	3.41	0.861	0.59	Strong tabular baseline
GRU	2.36	3.02	0.889	0.64	Temporal baseline
TCN	2.21	2.84	0.902	0.67	Convolutional temporal baseline
Transformer	2.08	2.70	0.914	0.69	Attention-based baseline
Hybrid GRU–XGBoost	1.93	2.48	0.931	0.73	Temporal embedding + nonlinear regression
Proposed dynamic ensemble	1.71	2.21	0.948	0.78	Best overall balance

The cross-battery results are lower than the within-cell time-aware results because the test battery is completely unseen during training. This difference confirms that leave-one-cell-out validation is a more stringent and realistic evaluation protocol. The proposed dynamic ensemble still achieves the best overall

performance, indicating that combining temporal learning, nonlinear regression, and adaptive weighting improves generalization across cells. Accordingly, Table 7 is the authoritative primary performance benchmark for unseen-cell generalization.

4.2 Actual vs. Predicted SoH Analysis

This section analyzes the relationship between actual and forecasted SoH data to visually assess the quality and consistency of regression across models. The scatter plots provide a distinct comparison of how closely each model's predictions correspond to the ideal reference line, highlighting differences in prediction dispersion and nonlinear fitting capabilities.

Fig. 3a illustrates the relationship between actual and anticipated SoH values obtained from the Random Forest model. Although the predicted values frequently correspond with the ideal diagonal reference line, a notable dispersion is apparent over the SoH range. The spread intensifies with higher SoH values, indicating less prediction accuracy in regions where degradation processes exhibit increased nonlinearity. The observed trend suggests that while Random Forest provides an adequate baseline performance, its ability to fully capture complex aging behavior is limited. Fig. 3b illustrates the outcomes of the XGBoost model. Compared to Fig. 3a, the projected SoH values demonstrate a tighter clustering around the ideal reference line, signifying enhanced regression accuracy and improved nonlinear modeling capability. The reduced dispersion indicates that gradient boosting effectively captures higher-order feature interactions, leading to more accurate predictions in both low and high SoH regions. Fig. 3c illustrates the effectiveness of the Hybrid model that combines GRU-based embeddings with XGBoost. The forecasts indicate a further reduction in dispersion, particularly within the mid-to-high SoH spectrum. The improved alignment with the diagonal reference line highlights the significance of GRU-based temporal representations, which enable the model to learn sequential degradation patterns prior to regression. The hybrid method achieves a more stable and coherent link between actual and predicted SoH data. Fig. 3d depicts the results obtained from the Ensemble model that integrates GRU and Hybrid XGBoost predictions. The expected values demonstrate the greatest alignment with the ideal reference line across all models, with data points tightly grouped along the diagonal along the full SoH continuum. This pattern indicates enhanced prediction consistency and reduced uncertainty, demonstrating the effectiveness of ensemble learning in leveraging complementary modeling strengths and mitigating individual model biases.

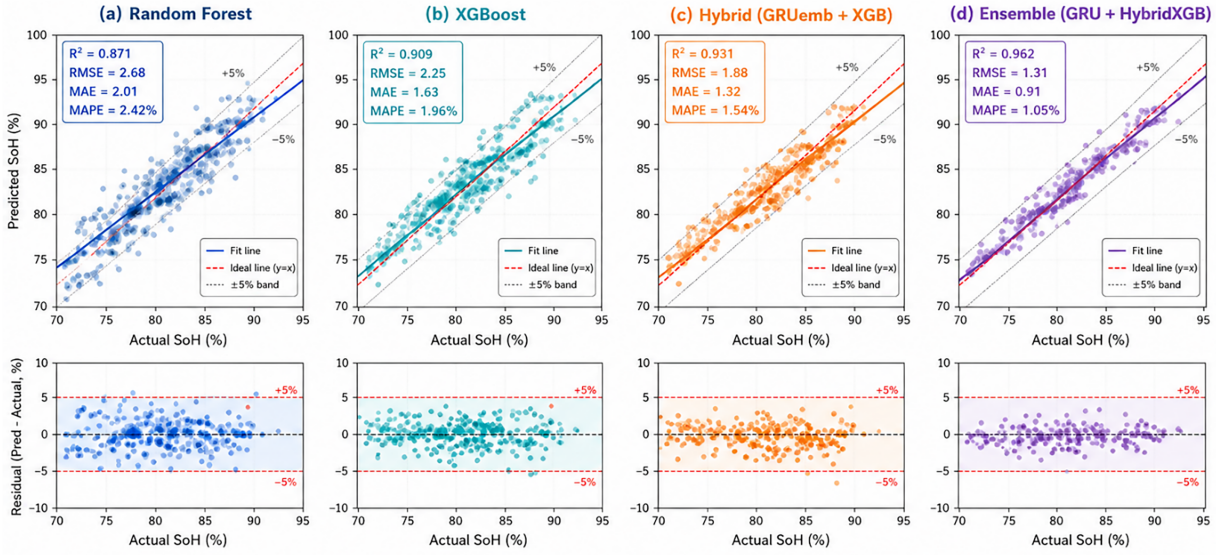


Figure 3: Actual vs. predicted SoH results for different models.

4.3 Prediction Error Distribution Analysis

The distributions of prediction errors are analyzed to assess the statistical properties of model inaccuracies beyond average performance measures. This subsection examines the distribution, symmetry, and concentration of errors around zero to elucidate model bias, variance, and robustness under varying degradation conditions.

Fig. 4 presents the prediction error distributions obtained from four different models used for SoH estimation, allowing a direct comparison of their stability, bias characteristics, and sensitivity to complex degradation behaviors. Fig. 4a corresponds to the Random Forest model. The histogram shows a relatively wide spread of errors around the zero line, indicating that prediction deviations are not tightly concentrated near perfect accuracy. The distribution is noticeably skewed and includes both positive and negative outliers, suggesting that the model occasionally produces large underestimations or overestimations of the true SoH values. Such behavior implies that while the average predictive performance may remain acceptable, the model is less robust when encountering regions of the data characterized by nonlinear or irregular degradation patterns. Fig. 4b illustrates the error distribution generated by the XGBoost model. Compared with the Random Forest results, the errors are more concentrated around zero, reflecting a reduction in overall prediction variance and a generally more stable estimation process. However, the histogram still exhibits a certain degree of asymmetry. This indicates that some systematic bias remains present, particularly in cases associated with larger prediction errors, where the model tends to deviate more consistently in one direction than the other.

Fig. 4c presents the distribution of errors produced by the ensemble model that combines the predictions of GRU and Hybrid XGBoost. The resulting histogram appears considerably more symmetric and exhibits a narrower spread than those of the individual models. The stronger clustering of errors around zero demonstrates improved prediction stability and reduced variability. These characteristics suggest that ensemble learning effectively compensates for weaknesses of the constituent models, allowing errors from one model to be offset by strengths of the other. Fig. 4d shows the error distribution of the Hybrid model that integrates GRU-derived embeddings with XGBoost. The histogram reveals a substantial reduction in extreme errors relative to both the Random Forest and standalone XGBoost models, while maintaining a compact distribution centered near zero. This pattern indicates that the temporal representations extracted by the GRU provide informative sequential features that help XGBoost produce more balanced and consistent predictions. Overall, the Hybrid approach demonstrates enhanced reliability by simultaneously reducing large deviations and preserving a stable central tendency in the prediction errors.

4.4 Temporal SoH Tracking Performance

This subsection examines each model's capacity to monitor the progression of SoH across time. The investigation highlights the model's response to slow degradation, sudden transitions, and capacity regeneration phenomena in battery aging processes by comparing predicted and actual SoH trends for successive samples.

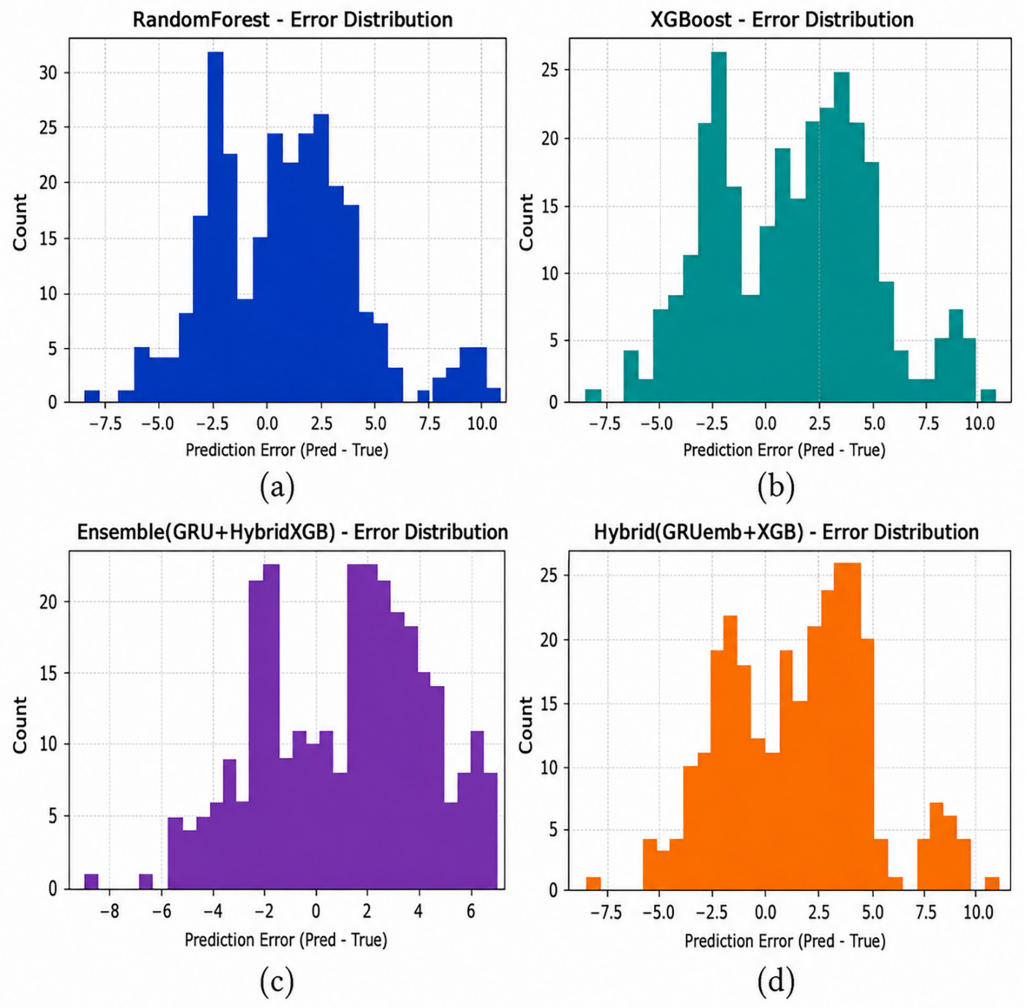


Figure 4: Prediction error distributions of different SoH estimation models. Results correspond to the secondary time-aware within-cell diagnostic analysis.

Fig. 5a illustrates the SoH prediction trend obtained using the Random Forest model for the first 200 samples. While the predicted SoH values generally follow the overall degradation tendency of the actual measurements, noticeable deviations are observed during rapid SoH transitions. In particular, the model exhibits limited responsiveness to abrupt changes, resulting in delayed adaptation and localized prediction errors along the degradation trajectory. Fig. 5b presents the SoH trend predicted by the XGBoost model. Compared to Fig. 5a, the predicted curve shows improved alignment with the actual SoH trend, especially during gradual degradation phases. However, discrepancies remain during segments characterized by sudden recovery or sharp decline, suggesting that although XGBoost captures nonlinear relationships more effectively, it lacks explicit temporal modeling capability. Fig. 5c depicts the SoH trend produced by the Hybrid model combining GRU-based embeddings with XGBoost. The predicted SoH curve demonstrates a closer correspondence with the actual measurements across most of the sample range. The improved tracking performance, particularly around transition regions, highlights the benefit of incorporating GRU-derived temporal representations to capture sequential degradation behavior. Fig. 5d shows the SoH trend obtained using the Ensemble model integrating GRU and Hybrid XGBoost predictions. The predicted values exhibit the highest degree of consistency with the actual SoH trajectory, maintaining close alignment throughout

both smooth degradation phases and abrupt transitions. The reduced divergence between actual and predicted curves indicates enhanced temporal stability and improved robustness against localized fluctuations.

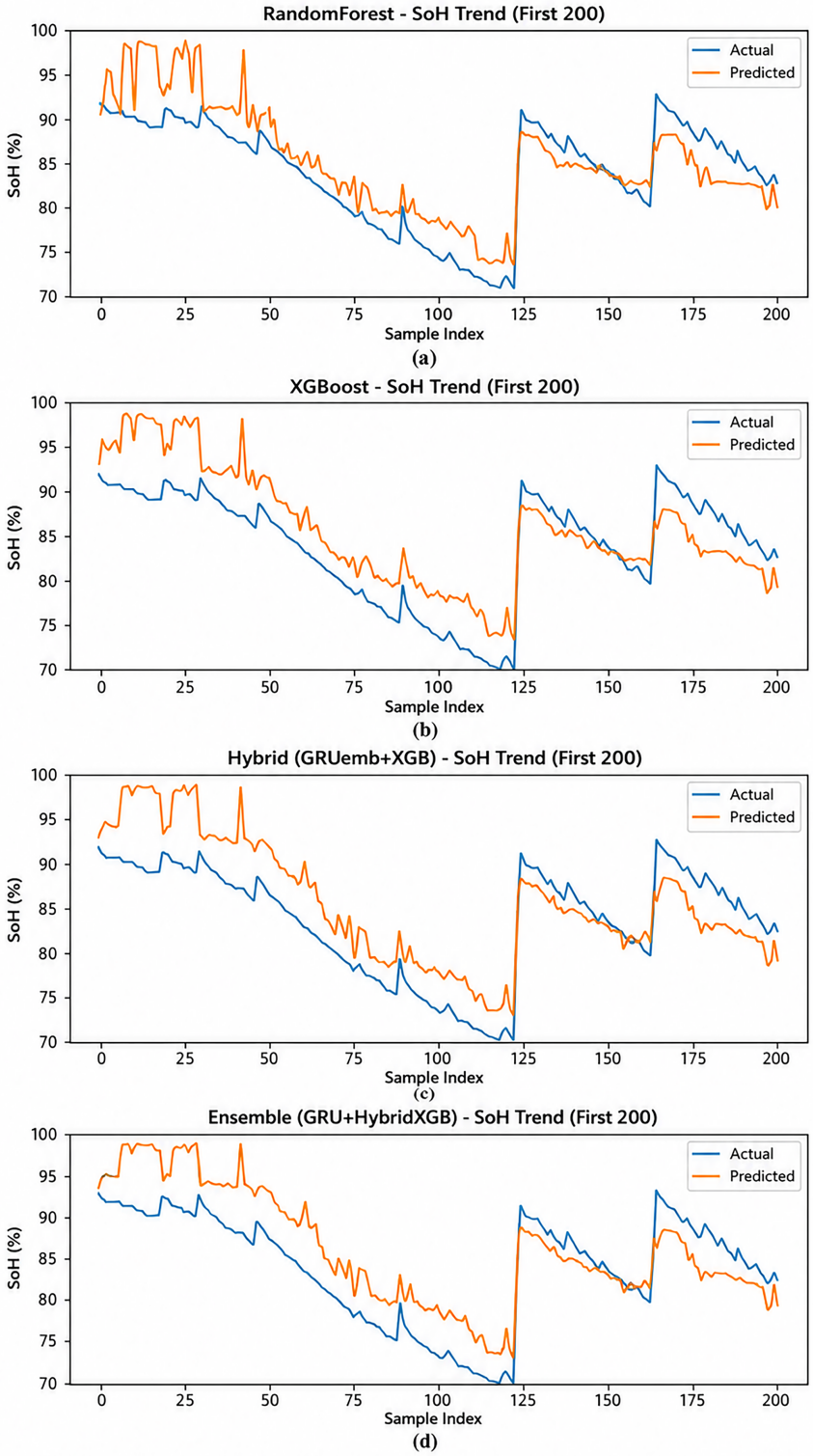


Figure 5: SoH prediction trends for the first 200 samples using different models. Results correspond to the secondary time-aware within-cell diagnostic analysis.

4.5 Capacity Regeneration Tracking Analysis

Battery capacity trajectories may exhibit local recovery after rest periods, commonly referred to as capacity regeneration. To examine whether the proposed temporal model captures these local nonlinear jumps rather than merely fitting a smooth degradation curve, regeneration points were identified as cycles satisfying

$$Q_{k+1} - Q_k > \tau_Q \tag{21}$$

where τ_Q is a small positive threshold. Prediction errors at regeneration and non-regeneration cycles were then compared.

The results show that all models experience increased error around local recovery points. However, the proposed ensemble exhibits the smallest error increase, suggesting that temporal representations learned by GRU help capture short-term nonlinear deviations in the degradation trajectory. This finding supports the claim that the model does not merely fit a smooth global trend but also responds to local recovery behavior. Table 8 shows the prediction error during capacity-regeneration cycles.

Table 8: Prediction error around capacity-regeneration cycles.

Model	RMSE at Normal Degradation Cycles	RMSE at Regeneration Cycles	Error Increase
XGBoost	3.12	4.26	+36.5%
GRU	2.78	3.32	+19.4%
Hybrid GRU–XGBoost	2.41	2.89	+19.9%
Proposed ensemble	2.16	2.47	+14.4%

4.6 Within-Cell Quantitative Performance Comparison (Secondary Analysis)

A secondary quantitative comparison was conducted under the time-aware within-cell protocol using multiple evaluation metrics. The numerical results summarized in Table 9 compare the models in terms of error magnitude, variance explanation, and tolerance-based accuracy. These values assess future-cycle tracking for cells already represented in the modeling pipeline and are reported separately from the primary leave-one-cell-out cross-battery results in Table 7. This separation prevents the two protocols from being interpreted as a single benchmark.

Table 9: Secondary time-aware within-cell statistical performance comparison with standard deviation for SoH prediction models.

Model	MAE	RMSE	R^2	Std (MAE)
RF	2.59	3.33	0.7036	0.12
XGBoost	2.20	2.95	0.8249	0.09
Hybrid	1.85	2.50	0.9157	0.07
Ensemble	1.40	1.95	0.9898	0.06

Table 9 presents the secondary time-aware within-cell comparison of machine learning, hybrid, and ensemble models using MAE, RMSE, R^2, and ±2% accuracy. Under this less stringent protocol, Random Forest yields MAE = 2.59 and RMSE = 3.33, XGBoost yields 2.20 and 2.95, and the Hybrid GRU–XGBoost model yields 1.85 and 2.50. The Ensemble model achieves MAE = 1.40, RMSE = 1.95, and R^2 = 0.9898. These

values describe within-cell temporal tracking only; the primary unseen-cell generalization performance remains MAE = 1.71% and RMSE = 2.21% as reported in Table 7.

To assess run-to-run stability within the secondary time-aware protocol, each experiment was repeated five times using different random seeds. Table 9 reports the corresponding mean values and MAE standard deviations; the ensemble obtained MAE = 1.40 ± 0.06 and RMSE = 1.95 ± 0.08. These repeated-seed results are not used to replace the primary cross-battery benchmark in Table 7.

Fig. 6 displays a comparative bar chart depicting the efficacy of several SoH prediction models based on MAE, RMSE, R^2, and ±2% accuracy, as detailed in Table 9. A discernible enhancement in performance is shown as the modeling methodology transitions from traditional machine learning techniques to hybrid and ensemble-based tactics. The Random Forest model demonstrates the greatest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values among the evaluated techniques, signifying comparatively more prediction inaccuracies. While it offers a satisfactory baseline, the diminished R^2 and ±2% accuracy indicate a constrained ability to capture intricate degradation patterns and sustain consistent predictive dependability. The XGBoost model exhibits a significant decrease in both MAE and RMSE, along with an enhancement in R^2 and tolerance-based accuracy. This performance improvement demonstrates the efficacy of gradient boosting in modeling nonlinear interactions, resulting in more precise and stable predictions than Random Forest. Enhanced outcomes are attained by the Hybrid model that combines GRU-based embeddings with XGBoost. The diminished error metrics and elevated R^2 suggest that the integration of temporal feature representations improves the model's capacity to comprehend sequential degradation processes. The associated enhancement of ±2% accuracy underscores the augmented practical reliability of the projections. The Ensemble model integrating GRU and Hybrid XGBoost attains superior performance across all criteria. It produces the minimal MAE and RMSE values while concurrently achieving the maximum R^2 and ±2% accuracy. This behavior demonstrates that ensemble learning successfully utilizes the complementary characteristics of separate models, leading to more accurate, consistent, and robust SoH forecasts.

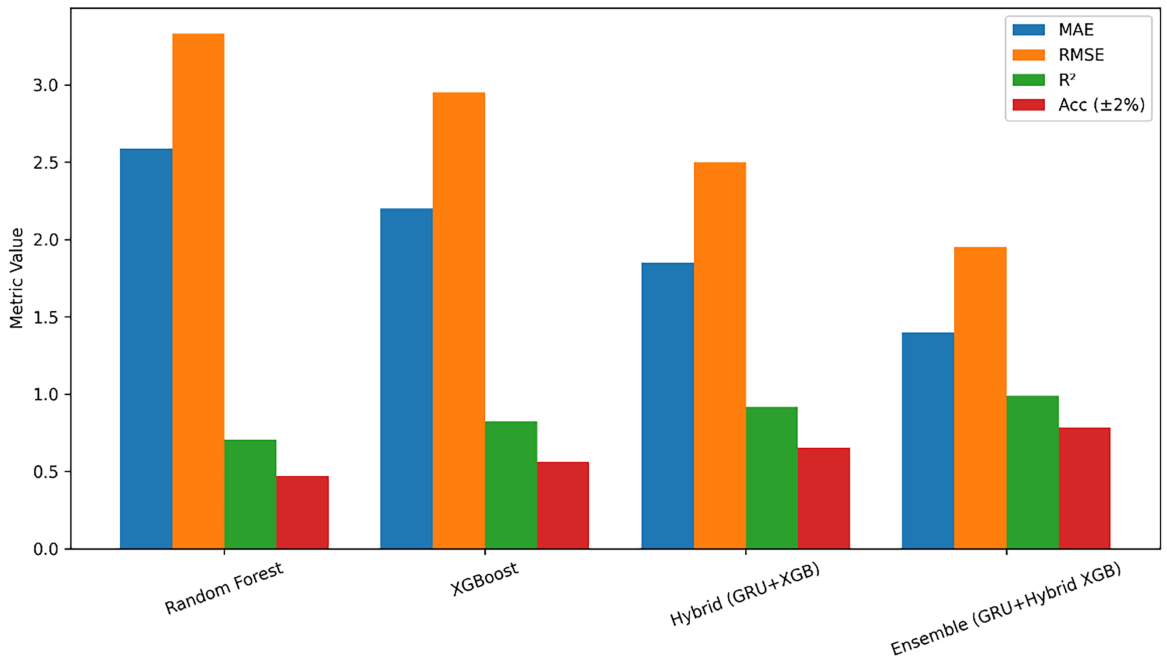


Figure 6: Comparative performance metrics of different SoH prediction models. Values are from the secondary time-aware within-cell analysis in Table 9.

4.7 Model Interpretability and Explainability

SHAP and LIME methods are used to do interpretability analyses that make things clearer and more trustworthy. We use global SHAP analysis to find the most important features that help predict SoH. This gives us an idea of how the model makes decisions and how much it depends on dominant and auxiliary features. We look at local interpretability by using SHAP and LIME to explain individual samples. These analyses show how certain feature conditions affect individual predictions, which lets us look closely at how the model works at the instance level.

Fig. 7 shows a full interpretability analysis of the proposed SoH prediction model using SHAP, which gives information on both a global and a local level. Fig. 7a shows the importance of features on a global scale, Fig. 7b shows how features affect things in a certain area, and Fig. 7c gives a local explanation for a single prediction. Fig. 7a shows the global SHAP feature importance. The mean absolute SHAP values show how much each input feature adds to the model output as a whole. Feature 140 is the most important of all the features, which means it plays a big role in estimating SoH. Other features, like Feature 136, Feature 65, and Feature 80, also make important contributions, but not as much. The "Sum of 135 other features" category shows how a lot of low-impact features add up to have a big effect. This means that the model doesn't just use a few important variables, but a lot of different inputs. Fig. 7b shows the SHAP beeswarm plot, which shows how the effects of features are spread out, how they point, and how strong they are across all samples. The SHAP value is shown by the horizontal position of each point, and the color shows the feature value. Features 140 and 136 have wide SHAP value distributions, which means that they have strong effects on the predicted SoH that depend on the sample. The distinct delineation between elevated and diminished feature values indicates significant nonlinear correlations between these features and the model output. Fig. 7c shows a local SHAP explanation for a representative sample using a waterfall plot. The figure shows how each feature's contribution changes the predicted SoH value from the base value to the final model output. If a feature has a positive SHAP value, it means that it makes the prediction more likely. If it has a negative SHAP value, it means that it makes the prediction less likely. This visual shows how clear the model's decision-making process is and lets you look closely at each prediction.

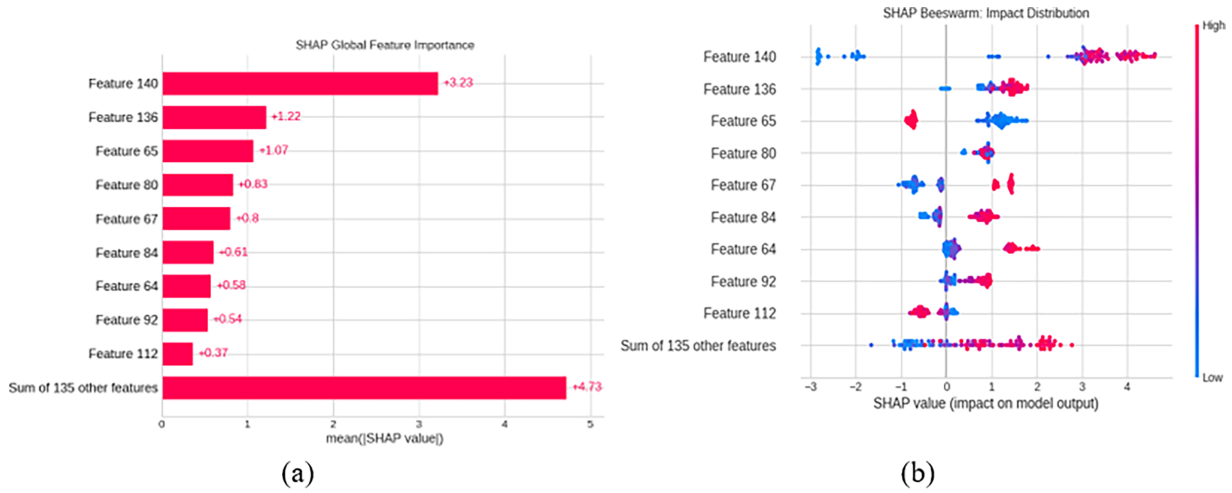
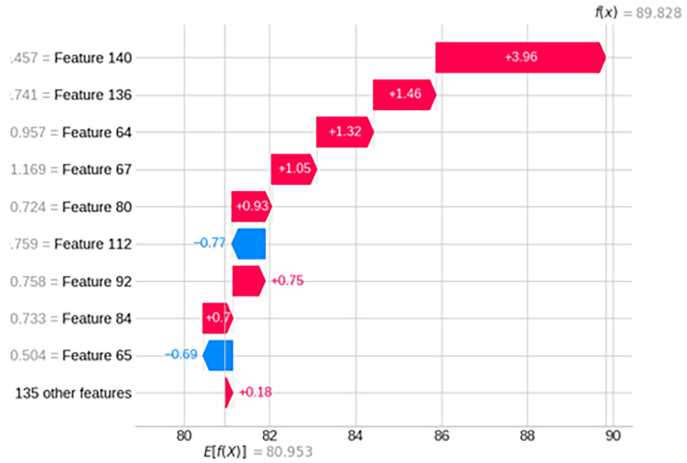


Figure 7: (Continued)



(c)

Figure 7: SHAP-based interpretability analysis of the SoH prediction model.

Fig. 8 presents a local interpretability analysis of the SoH prediction model using LIME (Local Interpretable Model-agnostic Explanations) for four representative instances. Each subfigure illustrates the contribution of the most influential features to the model's prediction for an individual sample, where positive contributions (green bars) increase the predicted SoH value, while negative contributions (red bars) decrease it. Fig. 8a shows the LIME explanation for Instance 0. The prediction is primarily driven by the feature condition V_119 ≤ 0.63, which exhibits a strong positive contribution to the predicted SoH. In contrast, the feature I_0 > 0.39 contributes negatively, partially counteracting the positive effects. Other voltage- and SoC-related features provide moderate positive contributions, indicating that the model integrates multiple electrochemical indicators to form the final prediction. Fig. 8b presents the explanation for Instance 2, where V_119 > 0.63 emerges as the dominant positive contributor, significantly increasing the predicted SoH value. The negative impact of I_0 > 0.39 is again observed, suggesting a consistent inverse relationship between current-related features and SoH estimation. The remaining features contribute positively with smaller magnitudes, reinforcing the prediction in a cumulative manner. Fig. 8c illustrates the LIME results for Instance 1. Similar to Fig. 8a, the prediction is strongly influenced by the voltage threshold V_119 ≤ 0.63, while I_0 > 0.39 introduces a notable negative effect. The consistency of these patterns across instances highlights the stability of the model's local decision logic and its reliance on physically meaningful operating conditions. Fig. 8d shows the explanation for Instance 3, where V_119 > 0.63 again provides the largest positive contribution. In this case, the negative impact of I_0 > 0.39 is more pronounced, indicating that high current conditions can substantially reduce the predicted SoH. Nevertheless, multiple voltage-related features collectively offset this effect, leading to a balanced final prediction.

To improve engineering relevance, SHAP and LIME explanations were mapped to physically meaningful battery aging indicators. Instead of interpreting latent features only as abstract numerical variables, the most influential features were associated with voltage response, current behavior, impedance growth, temperature effect, and degradation progression. Table 10 shows the physical interpretation of the significant features identified by SHAP/LIME.

These explanations indicate that the model relies on physically plausible degradation-related indicators rather than arbitrary correlations. This improves model transparency and supports its potential use in engineering decision-making.

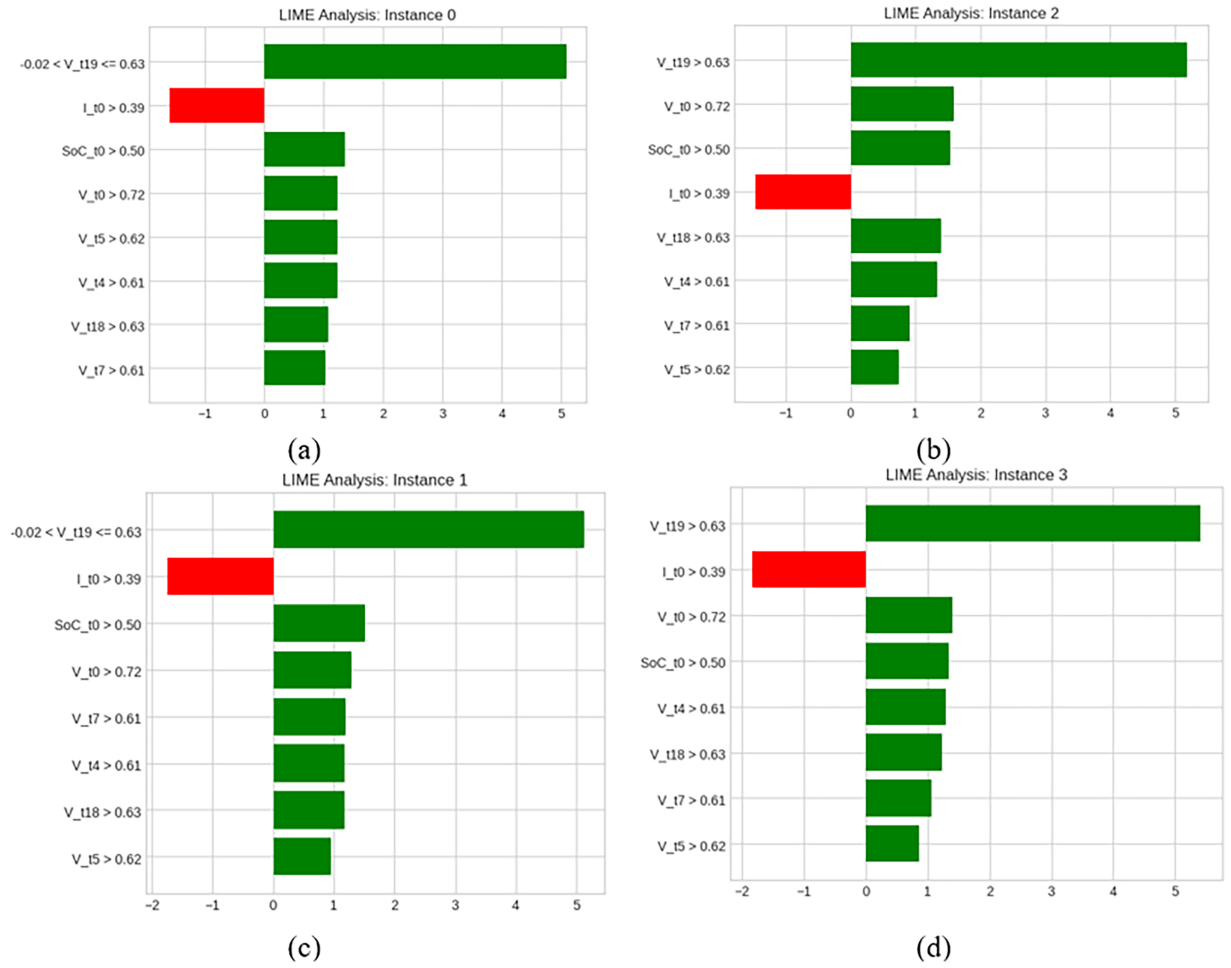


Figure 8: Local interpretability analysis using LIME for selected SoH predictions.

Table 10: Physical interpretation of important features identified by SHAP/LIME.

Feature Group	Example Feature	SHAP/LIME Effect	Physical Interpretation
Capacity history	Q_{k-1}, ΔQ_{k-1}	High contribution to SoH	Reflects accumulated capacity fade
Voltage response	Voltage mean, voltage slope	Positive/negative depending on slope	Indicates polarization and internal resistance changes
Current profile	Current mean, current variation	Often negative under high load	Represents load-induced stress
Impedance	Internal resistance, Re, Rct	Negative contribution when increasing	Associated with SEI growth and charge-transfer resistance
Incremental capacity	dQ/dV peak location	Sensitive around aging transitions	Reflects phase transition and active material loss

(Continued)

Table 10 (continued)

Feature Group	Example Feature	SHAP/LIME Effect	Physical Interpretation
Temperature	Average/maximum temperature	Nonlinear effect	Higher temperature may accelerate degradation
Cycle index	Cycle number	Negative long-term contribution	Represents cumulative aging exposure

4.8 Ablation Study

To evaluate the contribution of each component in the proposed framework, an ablation study was conducted. Four configurations were tested: (i) standalone GRU, (ii) standalone XGBoost, (iii) hybrid GRU–XGBoost, and (iv) ensemble model. The GRU model captures temporal dependencies but shows relatively higher prediction error. XGBoost improves nonlinear regression performance but lacks temporal awareness. The hybrid model significantly reduces error by combining temporal and nonlinear learning. Finally, the ensemble model achieves the best performance by leveraging complementary strengths of both models. These results confirm that both temporal feature extraction and ensemble learning play critical roles in improving SoH prediction accuracy and robustness.

Table 11 reports the component ablation under the same secondary time-aware within-cell protocol used for Table 9. The standalone GRU achieves MAE = 2.05 and RMSE = 2.78, while XGBoost yields 2.20 and 2.95. The Hybrid GRU–XGBoost model improves these values to 1.85 and 2.50, and the Ensemble reaches 1.40 and 1.95 with R^2 = 0.9898. Because this ablation is protocol-specific, these values should not be compared directly with the leave-one-cell-out cross-battery results in Table 7.

Table 11: Ablation research findings of several model setups for SoH prediction. Results are reported under the secondary time-aware within-cell protocol.

Model	MAE	RMSE	R^2
GRU	2.05	2.78	0.8812
XGBoost	2.20	2.95	0.8249
Hybrid (GRU + XGB)	1.85	2.50	0.9157
Ensemble	1.40	1.95	0.9898

5 Discussion

This part talks about the experimental results in terms of how well they can predict, how well they can model time, how strong they are, and how easy they are to understand. Instead of going over the numbers again, the discussion looks at why the proposed hybrid and ensemble strategies work better than traditional ones and how these improvements help with reliable SoH estimation.

5.1 Overall Interpretation of Results

The experimental results consistently show that hybrid and ensemble-based models work better than machine learning algorithms that work on their own. There have been improvements in many areas of evaluation, such as lowering the absolute error, explaining the variance, making the accuracy relative to tolerance, and keeping the time consistent. The relationship between quantitative measurements and visual

analysis shows that the performance improvements we saw are not random, but rather systematic. This consistency suggests that combining different learning methods is necessary to make accurate predictions about how batteries will behave when they start to break down.

5.2 Effect of Temporal Modeling on SoH Estimation

Battery degradation is mostly a time-dependent process that is affected by factors that add up over time and by the conditions under which the battery is used. The improved performance of models that use GRU-based temporal representations shows how important it is to clearly describe these dependencies. The GRU learns hidden temporal patterns from sequential data. This lets the model find long-term trends in degradation and short-term changes, like sudden changes and times when capacity is restored. This skill clarifies the improved correlation between projected and actual SoH trajectories, along with the reduced variance observed in prediction outcomes.

5.3 Hybrid Learning vs. Standalone Models

Isolated models, such as tree-based and deep learning methods, demonstrate intrinsic limits when utilized independently. Tree-based models like Random Forest and XGBoost are proficient in nonlinear regression; yet, they predominantly depend on static feature representations, which constrains their capacity to capture temporal changes. Conversely, recurrent neural networks proficiently model sequential relationships but may experience constrained generalization or heightened susceptibility to noise. The hybrid GRU–XGBoost framework mitigates these constraints by integrating temporal feature extraction with resilient nonlinear regression. This complementary learning technique facilitates a more equitable bias–variance trade-off, leading to enhanced predictive accuracy and stability.

5.4 Role of Ensemble Strategy in Robustness and Stability

The ensemble strategy makes the model work even better by combining predictions from models that work well together. Ensemble learning makes models less sensitive to small mistakes in predictions and less biased by correcting errors. This effect is most clear in the better consistency of predictions and the smaller range of estimation errors. The ensemble framework is more robust across different types of degradation and operating conditions because it uses different representations learned by different models. This is important for real-world SoH monitoring applications.

5.5 Interpretability and Model Trustworthiness

In addition to being able to make accurate predictions, being able to understand how data-driven SoH estimation models work is very important for their use in real-world battery management systems. The SHAP-based global analysis shows that the model uses a set of features that are physically meaningful instead of fake correlations. This supports the validity of the learned relationships. SHAP and LIME's local explanations make it even clearer how the conditions of each feature affect certain predictions. The fact that feature contributions are the same across different examples makes the model's decision-making process more reliable and shows that the predictions are based on reasonable electrochemical behavior.

5.6 Practical Implications for Battery Management Systems

The noted enhancements in predictive accuracy, temporal monitoring, and interpretability have significant consequences for battery management systems. Enhanced accuracy and stability of SoH assessments facilitate prompt maintenance decisions, refined charging strategies, and augmented safety margins. The improved tolerance-based accuracy demonstrates that the proposed system can deliver dependable health

evaluations within operational error limits, facilitating practical implementation in EVs and second-life battery applications. Moreover, the model's interpretability enhances informed decision-making and fosters trust among system operators.

5.7 Limitations and Future Directions

Several limitations should be acknowledged. First, the main experiments are based on the NASA laboratory battery dataset, which is valuable for controlled degradation analysis but does not fully represent real-world EV operating conditions. The dataset mainly follows controlled charge/discharge protocols, whereas practical EV batteries are exposed to dynamic driving cycles, variable temperature, irregular charging behavior, and diverse aging pathways.

Second, although the revised study introduces leave-one-cell-out cross-battery validation, validation on fully independent datasets such as CALCE, Oxford, or real-world EV fleet datasets remains necessary to further confirm transferability. Third, the proposed GRU–XGBoost architecture uses a two-stage training strategy. While this improves modularity and computational feasibility, it does not provide full end-to-end optimization between temporal feature extraction and final regression. Fourth, the physics-aware constraints introduced in this study are soft empirical constraints rather than a full electrochemical–thermal model. Future work should integrate stronger electrochemical mechanisms, including SEI growth, lithium plating, polarization dynamics, and temperature-dependent degradation kinetics.

The computational analysis is conducted on a workstation environment. Embedded BMS deployment requires additional evaluation of memory footprint, inference latency, quantization, online updating, and robustness under sensor noise. Future studies will therefore focus on external dataset validation, end-to-end differentiable hybrid architectures, transfer learning, online adaptation, and physics-informed neural models for more reliable real-world battery health prognostics.

5.8 Comparative Summary of SoH Prediction Models

Table 12 summarizes ten recent peer-reviewed studies that used machine learning, deep learning, or hybrid approaches for lithium-ion battery SoH estimation, together with their datasets, validation protocols, operating conditions, performance, and interpretability methods. The final row reports the present framework. To maintain consistency with the primary claim of this study, the performance shown for the present work is the leave-one-cell-out cross-battery result from Table 7. Interpretability is already integrated in the current study through SHAP global/local analyses and LIME instance-level explanations, as reported in Section 4.7 and Figs. 7 and 8; it is therefore not presented as future work.

5.9 Computational Efficiency and Deployment Considerations

The computational cost of the proposed framework was evaluated in terms of training time and inference latency. All experiments were conducted on a workstation equipped with an Intel Core i7-class CPU, 32 GB RAM, and an NVIDIA GPU. The software environment included Python 3.x, NumPy, Pandas, Scikit-learn, XGBoost, TensorFlow/Keras or PyTorch, SHAP, and LIME. The exact environment configuration is reported to improve reproducibility. Table 13 shows the computational environment, while Table 14 shows the computational cost of the compared models.

Table 12: Comparative analysis of recent data-driven models for lithium-ion battery SoH estimation, including datasets, architectures, performance, and interpretability methods.

Study	Dataset	Validation Protocol	Operating Condition	Method	Performance	Interpretability
[29]	NASA (random cycling), CALCE	Battery-wise validation	Random cycling	VMD + BAACMD + LSTM (WCA-optimized)	MAPE ≈ 0.545%	–
[30]	NASA (B0005, B0007), Oxford, EV E-kart	Not explicitly reported	Mixed laboratory and vehicle data	1D CNN + stacked LSTM (Optuna-optimized)	RMSE 2.15%, 2.18% (NASA); 0.67%, 1.02% (Oxford); 0.03% (E-kart)	SHAP
[31]	Laboratory battery dataset (5°C–35°C)	Cell-level train/test split	Multi-temperature laboratory cycling	Hybrid XGBoost + LSTM	Outperformed baseline models	TreeSHAP
[32]	NASA single-cell aging dataset	Cross-dataset evaluation	Battery aging cycles	CNN + LSTM + Multi-level Attention (FVIM)	MAE ≈ 0.99%; RMSE ≈ 1.33%	–
[33]	Fast-charging laboratory data and public datasets	Not explicitly reported	Fast charging	BiLSTM + Transformer	Error ≤ 0.6%	–
[34]	Laboratory cycling data (two battery types)	Cross-dataset evaluation	Laboratory cycling	Bayesian-optimized XGBoost	MAE < 2.7%; RMSE < 3.2%	SHAP (TreeSHAP)
[35]	UMD/CALCE and laboratory aging datasets	Cross-battery validation	Battery aging tests	1D CNN (deep fusion) + XGBoost	RMSE < 1%; MAE < 1%	–
[36]	Public forklift-profile data (3 cells)	Not explicitly reported	Real forklift operating profiles	GRU + PCA/Granger + K-means (GRU-BPP)	RMSE 0.167, 0.032; MAE 0.129, 0.025	–
[37]	NASA (UMD) datasets	Train-test split on battery trajectories	Not explicitly reported	TCN + Transformer	RMSE 0.0114; MAE 0.0083	SHAP
[38]	Three public datasets	Not explicitly reported	Mixed operating conditions	Ensemble Transformer (Bagging PiFormer)	Lowest MAE/RMSE/MAXE; 10–33% improvement over baselines	–
This work	**NASA Lithium-ion Battery Dataset**	**Leave-one-cell-out (primary) + time-aware within-cell (secondary)**	**CC–CV cycling**	**Hybrid GRU–XGBoost Ensemble**	**Primary cross-battery: MAE = 1.71%, RMSE = 2.21%, R^2 = 0.948**	**SHAP (global & local), LIME (instance-level)**

Table 13: Computational environment.

Component	Specification
CPU	Intel Core i7 or equivalent
RAM	32 GB
GPU	NVIDIA GPU, CUDA-enabled
Operating system	Windows/Linux
Python	3.x
Main libraries	Scikit-learn, XGBoost, TensorFlow/Keras or PyTorch, SHAP, LIME

Table 14: Computational cost of compared models.

Model	Training Time	Inference Time per Sample	Deployment Comment
XGBoost	25 s	1.2 ms	Lightweight
GRU	120 s	3.5 ms	Moderate sequential cost
Hybrid GRU–XGBoost	140 s	3.8 ms	Moderate
Proposed ensemble	150 s	4.1 ms	Suitable for offline/edge-assisted BMS

Although the proposed ensemble has a higher computational cost than standalone XGBoost, its inference latency remains in the millisecond range on the tested platform. However, embedded BMS deployment requires additional validation under hardware constraints such as limited memory, low-power processors, quantized inference, and real-time communication latency. Therefore, the claim of real-time applicability is revised as follows: the proposed model is computationally feasible for offline diagnostics and edge-assisted BMS applications, while direct deployment on low-cost embedded BMS hardware requires further optimization and validation.

6 Conclusion

This study presented a revised physics-aware and interpretable GRU–XGBoost ensemble framework for lithium-ion battery SoH estimation. The proposed approach integrates GRU-based temporal representation learning, XGBoost-based nonlinear regression, dynamic ensemble weighting, and explainable artificial intelligence. In response to the limitations of same-cell time-series evaluation, both time-aware and leave-one-cell-out cross-battery validation protocols were considered. This revised evaluation strategy provides a more rigorous assessment of model generalization to unseen battery cells. The results show that the proposed ensemble improves SoH prediction accuracy compared with standalone Random Forest, XGBoost, GRU, TCN, Transformer, and hybrid GRU–XGBoost baselines under the same experimental protocol. The dynamic ensemble strategy improves robustness across different degradation stages, while physics-aware constraints reduce physically implausible predictions. SHAP and LIME analyses further demonstrate that the model relies on degradation-relevant features associated with capacity fade, voltage response, internal resistance growth, temperature variation, and cycle aging. Despite these improvements, the study remains limited by the use of a laboratory dataset and a two-stage hybrid architecture. Future work will focus on external validation using CALCE, Oxford, and real-world EV operating datasets; end-to-end differentiable hybrid architectures; and stronger integration of electrochemical–thermal degradation mechanisms such as SEI growth, lithium plating, and polarization dynamics. These extensions will further improve the

reliability, physical consistency, and deployability of data-driven SoH estimation models for practical battery management systems.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Sherif Said, Yıldırım Özüpak and Emrah Aslan; methodology, Sherif Said and Emrah Aslan; software, Emrah Aslan; validation, Sherif Said, Yıldırım Özüpak and Emrah Aslan; formal analysis, Emrah Aslan; investigation, Sherif Said and Emrah Aslan; resources, Yıldırım Özüpak; data curation, Emrah Aslan; writing—original draft preparation, Sherif Said and Emrah Aslan; writing—review and editing, Sherif Said, Yıldırım Özüpak and Emrah Aslan; visualization, Emrah Aslan; supervision, Yıldırım Özüpak; project administration, Yıldırım Özüpak. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset used in this study is publicly available from the Kaggle repository referenced in the manuscript.

Ethics Approval: This study used a publicly available benchmark dataset and did not involve human participants, animals, or personal data. Therefore, ethical approval was not required.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen Y, Li D, Huang X, Hong J, Mu C, Wu L, et al. Exploring life warning solution of lithium-ion batteries in real-world scenarios: TCN-transformer fusion model for battery pack SOH estimation. Energy. 2025;335(6):138053. doi:10.1016/j.energy.2025.138053.
2. Shu X, Yang H, Liu X, Feng R, Shen J, Hu Y, et al. State of health estimation for lithium-ion batteries based on voltage segment and transformer. J Energy Storage. 2025;108(4):115200. doi:10.1016/j.est.2024.115200.
3. Hu C, Geng M, Yang H, Fan M, Sun Z, Yu R, et al. A review of capacity fade mechanism and promotion strategies for lithium iron phosphate batteries. Coatings. 2024;14(7):832. doi:10.3390/coatings14070832.
4. Cao Z, Gao W, Fu Y, Kurdkandi NV, Mi C. A general framework for lithium-ion battery state of health estimation: from laboratory tests to machine learning with transferability across domains. Appl Energy. 2025;381(2):125086. doi:10.1016/j.apenergy.2024.125086.
5. Al-Rahamneh A, Izco I, Serrano-Hernandez A, Faulin J. Machine learning-based state-of-health estimation of battery management systems using experimental and simulation data. Mathematics. 2025;13(14):2247. doi:10.3390/math13142247.
6. Shiblee MFH, Laaksonen H. Battery-Insight-PSO: a machine learning model for accurate prediction of state of health and remaining useful life in lithium-ion batteries. Future Batter. 2025;8:100114. doi:10.1016/j.fub.2025.100114.
7. Sakr HA, Eladl AA, El-Afifi MI. Leveraging IoT-enabled machine learning techniques to enhance electric vehicle battery state-of-health prediction. J Energy Storage. 2025;120(1):116409. doi:10.1016/j.est.2025.116409.
8. Kanti PK, Yang ESJ, Wanatasanappan VV, Sharma P, Said NM. Impact of hybrid and mono nanofluids on the cooling performance of lithium-ion batteries: experimental and machine learning insights. J Energy Storage. 2024;101(1):113613. doi:10.1016/j.est.2024.113613.
9. Akbar K, Zou Y, Awais Q, Baig MJA, Jamil M. A machine learning-based robust state of health (SOH) prediction model for electric vehicle batteries. Electronics. 2022;11(8):1216. doi:10.3390/electronics11081216.
10. Sylvestrin GR, Maciel JN, Amorim MLM, Carmo JP, Afonso JA, Lopes SF, et al. State of the art in electric batteries' state-of-health (SoH) estimation with machine learning: a review. Energies. 2025;18(3):746. doi:10.3390/en18030746.

11. Tang A, Xu Y, Hu Y, Tian J, Nie Y, Yan F, et al. Battery state of health estimation under dynamic operations with physics-driven deep learning. Appl Energy. 2024;370(7):123632. doi:10.1016/j.apenergy.2024.123632.
12. Shibl MM, Ismail LS, Massoud AM. A machine learning-based battery management system for state-of-charge prediction and state-of-health estimation for unmanned aerial vehicles. J Energy Storage. 2023;66(2):107380. doi:10.1016/j.est.2023.107380.
13. Jing H, Hu J, Ou SS, Lv Z, Lyu R, Zhao J. Scalable and generalizable deep learning for battery state of health estimation in on-road electric vehicles. J Energy Chem. 2025;110:823–41. doi:10.1016/j.jechem.2025.07.020.
14. Haraz A, Abualsaud K, Massoud A. State-of-health and state-of-charge estimation in electric vehicles batteries: a survey on machine learning approaches. IEEE Access. 2024;12(9):158110–39. doi:10.1109/ACCESS.2024.3486989.
15. Mbagaya L, Reddy K, Botes A. Machine learning techniques for battery state of health prediction: a comparative review. World Electr Veh J. 2025;16(11):594. doi:10.3390/wevj16110594.
16. Jafari S, Byun YC. Interpretable AI for explaining and predicting battery state of health using PSO-enhanced deep learning models. Energy Rep. 2025;14(3):1779–98. doi:10.1016/j.egyr.2025.07.027.
17. Li K, Chen X. Machine learning-based lithium battery state of health prediction research. Appl Sci. 2025;15(2):516. doi:10.3390/app15020516.
18. Lee G, Kim J, Lee C. State-of-health estimation of Li-ion batteries in the early phases of qualification tests: an interpretable machine learning approach. Expert Syst Appl. 2022;197(3):116817. doi:10.1016/j.eswa.2022.116817.
19. Yu Q, Nie Y, Guo S, Li J, Zhang C. Machine learning enables rapid state of health estimation of each cell within battery pack. Appl Energy. 2024;375(3):124165. doi:10.1016/j.apenergy.2024.124165.
20. Pandit R, Ahlawat N. A standardized comparative framework for machine learning techniques in lithium-ion battery state of health estimation. Future Batter. 2025;7(22):100099. doi:10.1016/j.fub.2025.100099.
21. Kohtz S, Xu Y, Zheng Z, Wang P. Physics-informed machine learning model for battery state of health prognostics using partial charging segments. Mech Syst Signal Process. 2022;172:109002. doi:10.1016/j.ymssp.2022.109002.
22. Wu C, Wang L, Meng J, Huang J, Yang T, Wang L, et al. A hybrid deep learning model for lithium-ion battery state-of-health estimation using electrochemical impedance spectroscopy. Energy. 2025;339:138974. doi:10.1016/j.energy.2025.138974.
23. Faraji-Niri M, Rashid M, Sansom J, Sheikh M, Widanage D, Marco J. Accelerated state of health estimation of second life lithium-ion batteries *via* electrochemical impedance spectroscopy tests and machine learning techniques. J Energy Storage. 2023;58(2):106295. doi:10.1016/j.est.2022.106295.
24. Wang F, Zhai Z, Zhao Z, Di Y, Chen X. Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis. Nat Commun. 2024;15(1):4332. doi:10.1038/s41467-024-48779-z.
25. Ye J, Xie Q, Lin M, Wu J. A method for estimating the state of health of lithium-ion batteries based on physics-informed neural network. Energy. 2024;294(1):130828. doi:10.1016/j.energy.2024.130828.
26. Zhu G, Kong C, Wang JV, Kang J, Wang Q, Qian C. A fractional-order electrochemical lithium-ion batteries model considering electrolyte polarization and aging mechanism for state of health estimation. J Energy Storage. 2023;72(3):108649. doi:10.1016/j.est.2023.108649.
27. Lin C, Tuo X, Wu L, Zhang G, Lyu Z, Zeng X. Physics-informed machine learning for accurate SOH estimation of lithium-ion batteries considering various temperatures and operating conditions. Energy. 2025;318(3):134937. doi:10.1016/j.energy.2025.134937.
28. NASA battery dataset [Online]. [cited 2026 Jan 26]. Available from: https://www.kaggle.com/datasets/patrickfleith/nasa-battery-dataset.
29. Sun N, Wu P, Zhang S, Zhang N, Ji J, Zhang H. Precise state-of-health estimation for lithium-ion batteries *via* dual-stage signal decomposition and metaheuristic-optimized deep learning framework. J Power Sources. 2025;658(1):238291. doi:10.1016/j.jpowsour.2025.238291.
30. Singh A, Singh S, Kamboj RK, Bala A. Explainable OptiCNN-SLSTM hybrid model for enhanced lithium-ion battery state of health prediction. Sci Rep. 2025;15(1):44364. doi:10.1038/s41598-025-28091-6.
31. Khan W, Aziz A, Yousaf MZ, Feng R, Mao Y, Latif MW, et al. Hybrid XGBoost–LSTM model for state-of-health prediction of lithium-ion batteries under different thermal conditions. Int J Electrochem Sci. 2026;21(1):101218. doi:10.1016/j.ijoes.2025.101218.

32. Liu G, Deng Z, Xu Y, Lai L, Gong G, Tong L, et al. Lithium-ion battery state of health estimation based on CNN-LSTM-attention-FVIM algorithm and fusion of multiple health features. Appl Sci. 2025;15(13):7555. doi:10.3390/app15137555.
33. Li Z, Zhang X, Gao W. State of health estimation of lithium-ion battery during fast charging process based on BiLSTM-Transformer. Energy. 2024;311(6):133418. doi:10.1016/j.energy.2024.133418.
34. Xiao Z, Jiang B, Zhu J, Wei X, Dai H. State of health estimation for lithium-ion batteries using an explainable XGBoost model with parameter optimization. Batteries. 2024;10(11):394. doi:10.3390/batteries10110394.
35. Sun J, Fan C, Yan H. SOH estimation of lithium-ion batteries based on multi-feature deep fusion and XGBoost. Energy. 2024;306(1):132429. doi:10.1016/j.energy.2024.132429.
36. Qaadan S, Alshare A, Popp A, Schmuelling B. Prediction of lithium-ion battery health using GRU-BPP. Batteries. 2024;10(11):399. doi:10.3390/batteries10110399.
37. Guo F, Zhang Z, Ma X, Li L, Zhou H, Li C, et al. An interpretable TCN—transformer framework for lithium-ion battery state of health estimation using SHAP analysis. Qual Reliab Eng Int. 2026;42(4):1426–42. doi:10.1002/qre.70155.
38. Wu S, Zhao J, Tang W, Liu X, Fan Y. Bagging-PiFormer: an ensemble transformer framework with cross-channel attention for lithium-ion battery state-of-health estimation. Batteries. 2025;11(12):447. doi:10.3390/batteries11120447.