batteries

Article

Comparative Analysis of ML and DL Models for Data-Driven SOH Estimation of LIBs Under Diverse Temperature and Load Conditions

Seyed Saeed Madani [1,*], Marie Hébert [2], Loïc Boulon [2], Alexandre Lupien-Bédard [3] and François Allard [1]

[1] Centre Énergie, Matériaux et Télécommunications (EMT), Institut National de la Recherche Scientifique (INRS), Varennes, QC J3X 1P7, Canada; francois.allard@inrs.ca
[2] Electrical and Computer Engineering Department, Université du Québec à Trois-Rivières (UQTR), Trois-Rivières, QC G8Z 4M3, Canada; marie.hebert@uqtr.ca (M.H.); loic.boulon@uqtr.ca (L.B.)
[3] Innovative Vehicle Institute (IVI), 100 Claude-Audy Street, Saint-Jérôme, QC J5L 0J2, Canada; albedard@ivisolutions.ca
* Correspondence: seyed-saeed.madani@inrs.ca

Abstract

Accurate estimation of lithium-ion battery (LIB) state of health (SOH) underpins safe operation, predictive maintenance, and lifetime-aware energy management. Despite recent advances in machine learning (ML), systematic benchmarking across heterogeneous real-world cells remains limited, often confounded by data leakage and inconsistent validation. Here, we establish a leakage-averse, cross-battery evaluation framework encompassing 32 commercial LIBs (B5–B56) spanning diverse cycling histories and temperatures (≈4 °C, 24 °C, 43 °C). Models ranging from classical regressors to ensemble trees and deep sequence architectures were assessed under blocked 5-fold GroupKFold splits using RMSE, MAE, R^2 with confidence intervals, and inference latency. The results reveal distinct stratification among model families. Sequence-based architectures—CNN–LSTM, GRU, and LSTM—consistently achieved the highest accuracy (mean RMSE ≈ 0.006; per-cell R^2 up to 0.996), demonstrating strong generalization across regimes. Gradient-boosted ensembles such as LightGBM and CatBoost delivered competitive mid-tier accuracy (RMSE ≈ 0.012–0.015) yet unrivaled computational efficiency (≈0.001–0.003 ms), confirming their suitability for embedded applications. Transformer-based hybrids underperformed, while approximately one-third of cells exhibited elevated errors linked to noise or regime shifts, underscoring the necessity of rigorous evaluation design. Collectively, these findings establish clear deployment guidelines: CNN–LSTM and GRU are recommended where robustness and accuracy are paramount (cloud and edge analytics), while LightGBM and CatBoost offer optimal latency–efficiency trade-offs for embedded controllers. Beyond model choice, the study highlights data curation and leakage-averse validation as critical enablers for transferable and reliable SOH estimation. This benchmarking framework provides a robust foundation for future integration of ML models into real-world battery management systems.

Keywords: lithium-ion battery; state of health (SOH); battery management systems (BMS); sequence models (CNN–LSTM, GRU, LSTM); gradient boosting (LightGBM, CatBoost); cross-battery benchmarking; uncertainty calibration (ECE, MCE); temperature effects; end-of-life prediction (EOL); interpretability (SHAP, permutation importance)

Academic Editor: Alessandro Lampasi

Received: 25 July 2025
Revised: 30 September 2025
Accepted: 9 October 2025
Published: 24 October 2025

Citation: Madani, S.S.; Hébert, M.; Boulon, L.; Lupien-Bédard, A.; Allard, F. Comparative Analysis of ML and DL Models for Data-Driven SOH Estimation of LIBs Under Diverse Temperature and Load Conditions. *Batteries* **2025**, *11*, 393. https://doi.org/10.3390/batteries11110393

1. Introduction

1.1. Importance of SOH

LIBs have become indispensable for portable electronics, electric vehicles (EVs), and grid-scale storage systems due to their high energy density and efficiency. However, their performance degrades over time because of electrochemical, thermal, and mechanical factors such as depth of discharge, charge/discharge rates, temperature variations, and cycling [1]. This degradation manifests as capacity fade, resistance growth, and voltage imbalance, directly impacting safety, reliability, and sustainability. High-temperature operation further accelerates degradation by inducing lithium plating, electrolyte decomposition, and transition-metal dissolution, which amplify polarization and thermal instabilities [2]. Broader reviews highlight that degradation processes such as SEI formation, calendar aging, electrolyte breakdown, and deep discharge collectively reduce energy density, increase impedance, and impair long-term reliability [3]. Recent bibliometric analyses confirm a sharp rise in global research output on SOH and RUL estimation, with China, the United States, and Europe leading collaborative efforts. These works also underscore the shift toward artificial intelligence (AI)-enabled strategies and hybrid modeling for improved predictive performance, while pointing to the challenges of computational burden and scalability in large battery packs [4]. Together, these findings establish SOH estimation as a cornerstone for advancing battery management systems (BMSs) and enabling safe, durable energy storage.

SOH estimation is therefore a cornerstone for advancing battery management systems (BMSs) and ensuring safe, durable energy storage. To illustrate the practical assessment process, Figure 1 presents an experimental workflow for SOH estimation, where diagnostic tests provide direct and indirect health indicators that are benchmarked against rated values to determine end-of-life criteria.

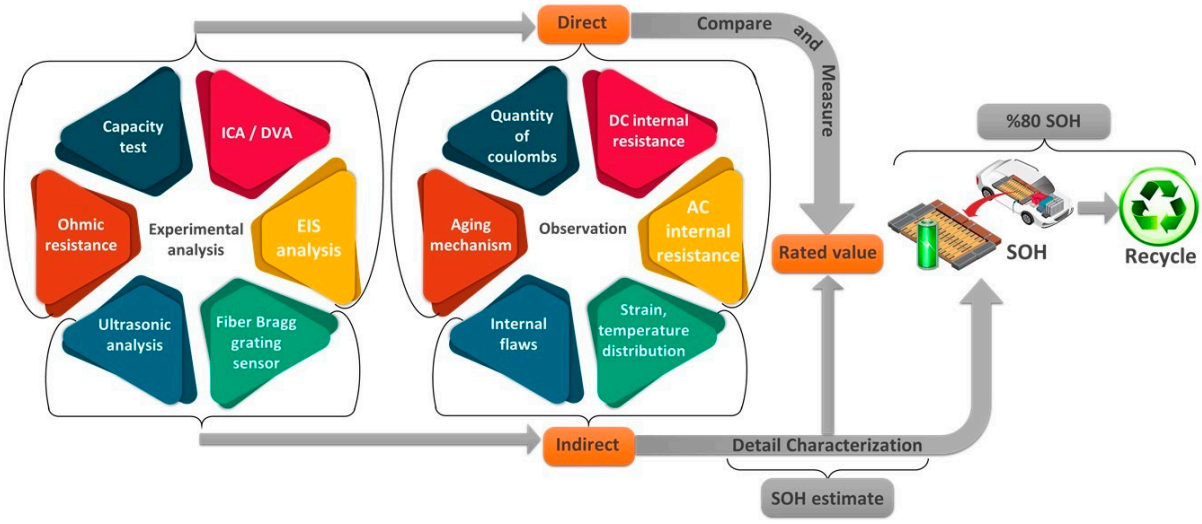


Figure 1. Experimental workflow for SOH estimation, connecting diagnostic tests with health indicators and end-of-life criteria.

1.2. Categories of Approaches

Figure 2 illustrates the main SOH estimation categories: experimental methods (direct and indirect measurements), model-based methods (equivalent circuit models, Kalman filters, and adaptive and fuzzy logic), and data-driven methods (machine learning and deep learning). Conventional voltage- and current-based techniques fall short in capturing internal aging phenomena, motivating the use of electrochemical impedance spectroscopy (EIS) for more reliable SOH estimation. Recent works demonstrate the promise of EIS

in providing rapid and precise health indicators, while also noting limitations in equipment complexity, online feasibility, and data stability [5]. Comparative reviews of EIS highlight that model-based methods, though interpretable, are sensitive to fitting accuracy and parameter initialization, whereas data-driven techniques offer flexibility but demand large, high-quality datasets and struggle with noise and transferability [6]. In parallel, parameter identification and state-of-power (SOP) estimation remain essential for robust modeling. Traditional methods such as least squares are error-prone, driving research toward advanced sensing, multi-feature fusion, and digital twin integration [7]. More recent approaches leverage metaheuristic algorithms such as particle swarm optimization, grey wolf optimization, and harmony search to optimize equivalent circuit models (ECMs), achieving low root mean square errors and strong robustness under dynamic load conditions [8]. Complementing these are deep learning (DL)-based frameworks that integrate recurrent, convolutional, and temporal models for ECM parameter identification, significantly reducing error compared to classical techniques [9]. To improve trust and interpretability, explainable machine learning (XML) methods are also emerging, particularly for battery production and health estimation processes, where feature importance and SHAP analyses enhance transparency [10]. Collectively, these studies demonstrate that physics-based models ensure interpretability and mechanistic insight, while data-driven models provide adaptability, and that hybrid integration is increasingly essential. As examples of conventional data-driven works, Figure 3 illustrates both the internal mechanism of a long short-term memory (LSTM) cell and a representative workflow for DL-based SOH estimation. The LSTM structure demonstrates how input, forget, and output gates regulate temporal dependencies, a property often leveraged in prior SOH prediction studies. The workflow highlights the typical pipeline of DL approaches, including feature extraction, base model training, re-training with transfer learning, and model evaluation using metrics such as MAE, RMSE, and MaxE.

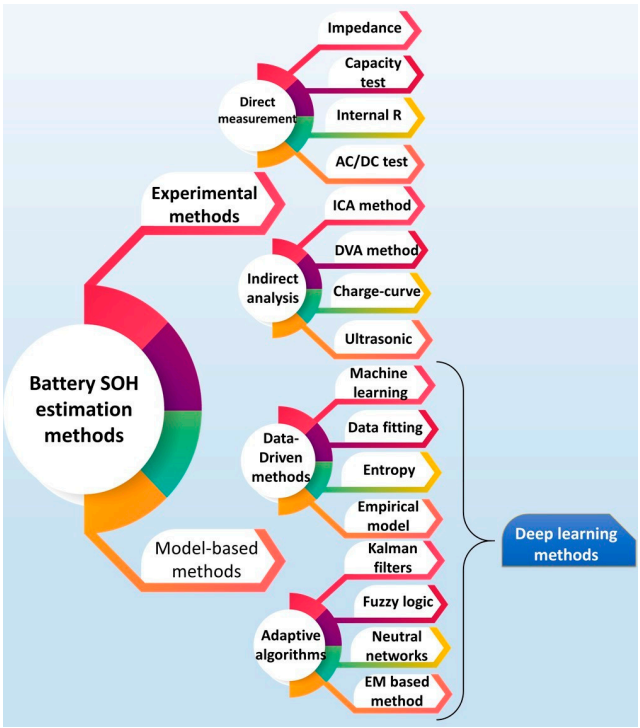


Figure 2. Overview of battery SOH estimation methods, including experimental, model-based, and data-driven approaches.

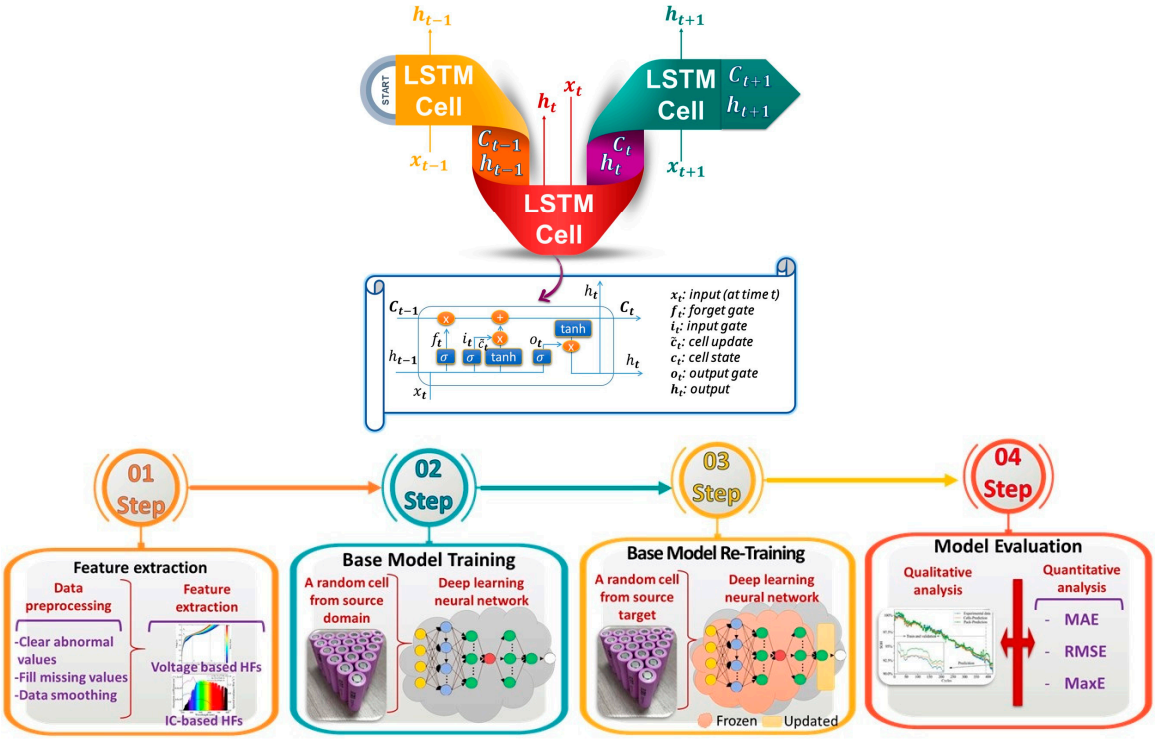


Figure 3. Examples of conventional data-driven SOH estimation methods: (**top**) LSTM cell structure showing gating mechanisms for temporal learning; (**bottom**) DL workflow illustrating feature extraction, model training, transfer learning, and evaluation.

1.3. Advances in Data-Driven SOH Estimation

The rise of machine learning (ML) and DL has transformed SOH prediction. One of the earliest breakthroughs demonstrated that carefully designed datasets and data-driven modeling could predict cycle life before observable degradation, reducing experimental costs and accelerating technology development [11]. Since then, architectures such as CNN–LSTM hybrids with skip connections have achieved superior robustness and lower error distributions across NASA and Oxford datasets [12]. To overcome the limitations of costly labeled data, domain adaptation strategies have enabled label-free SOH estimation with absolute errors below 5%, validated across diverse commercial cells [13]. For fast-charging conditions, BiLSTM–Transformer frameworks capture both long-term dependencies and global feature interactions, outperforming classical baselines with prediction errors below 1% [14]. Building further, transformer–LSTM fusion models that combine temporal and spectral feature extraction deliver more than 50% accuracy improvements under accelerated degradation [15]. Hybrid CNN–self-attention networks have proven effective within constrained SOC windows during rapid charging, while maintaining practical model size for embedded deployment [16]. A growing trend integrates physics into neural networks, where physics-informed neural networks (PINNs) combine empirical degradation equations with DL, achieving mean errors below 1% across multiple datasets [17]. Similarly, hybrid DL–transfer learning frameworks with inception modules, attention mechanisms, and cross-domain fine-tuning have shown outstanding scalability across chemistries and scenarios, while optimizing for computational efficiency [18]. These advances underscore the trajectory toward data-efficient, generalizable, and physics-informed models for SOH estimation. Table 1 provides a consolidated comparison of the approaches, and Table 2 presents recent advances across classical, hybrid, and physics-informed models.

Table 1. Comparison of SOH Estimation Approaches.

Ref.	Approach	Key Insights/Contributions
[5]	EIS for SOH monitoring	Highlights promise of rapid impedance-based estimation; challenges include equipment complexity, online measurement instability, and computational burden.
[6]	EIS methods (model-based vs. data-driven)	Reviews trade-offs: model-based methods are interpretable but sensitive to fitting; data-driven methods are adaptable but require large, high-quality datasets.
[7]	Parameter identification & SOP estimation	Categorizes techniques for single cells and packs; emphasizes challenges in robustness, accuracy, and computational complexity.
[8]	Metaheuristic parameter identification	Demonstrates optimization of ECMs using PSO, GWO, HS, and GEO; PSO offers the lowest RMSE, and HS provides faster computation.
[9]	DL-based parameter identification	Compares RNN, LSTM, GRU, TCN, and 1DCNN; convolutional models significantly outperform recurrent ones, reducing error by >50%.
[10]	Explainable machine learning (XML)	First review of XML in LIBs; FI and SHAP dominate; shows importance of interpretability for production and health estimation.
[17]	Physics-informed neural networks (PINNs)	Integrates degradation equations with DL; achieves robust accuracy (<1% error) across datasets, balancing physical interpretability and ML flexibility.

Table 2. Advances in Data-Driven SOH Estimation.

Ref.	Category	Contribution
[11]	Classical (early data-driven)	Pioneering work predicting cycle life from early discharge curves; demonstrates the value of deliberate dataset design.
[12]	Hybrid CNN–LSTM	CNN–LSTM with skip connections improves robustness; RMSE < 0.004 across NASA & Oxford datasets.
[13]	Label-free SOH (domain adaptation)	Framework eliminates the need for costly degradation experiments; achieves < 5% error across diverse cells.
[14]	BiLSTM–Transformer	Combines long-term dependencies with global feature learning; achieves SOH error < 1% during fast charging.
[15]	Transformer–LSTM fusion	Extracts multi-domain features (time/frequency/temperature); improves predictive accuracy by >50%.
[16]	Hybrid CNN–Attention	Accurate SOH estimation in constrained SOC windows under fast charging; compact architecture for BMS.
[17]	Physics-informed NN (PINN)	Embeds state-space & degradation models into NN training; enables robust generalization across conditions.
[18]	Hybrid DL + Transfer Learning	IDC + CRA modules with staged TL; achieves RMSE < 0.5% and strong computational efficiency for embedded deployment.

To provide a unifying perspective, Figure 4 depicts a general pipeline for LIB SOH estimation. The illustration consolidates the essential stages commonly encountered across the literature: (i) systematic acquisition and preprocessing of operational signals, (ii) recognition of practical challenges when moving from controlled laboratory settings to real-world deployment, (iii) extraction of multi-modal features capturing electrochemical and thermal

degradation dynamics, and (iv) integration of these representations into deep learning architectures for predictive modeling. This context-setting framework outlines the broader field landscape.

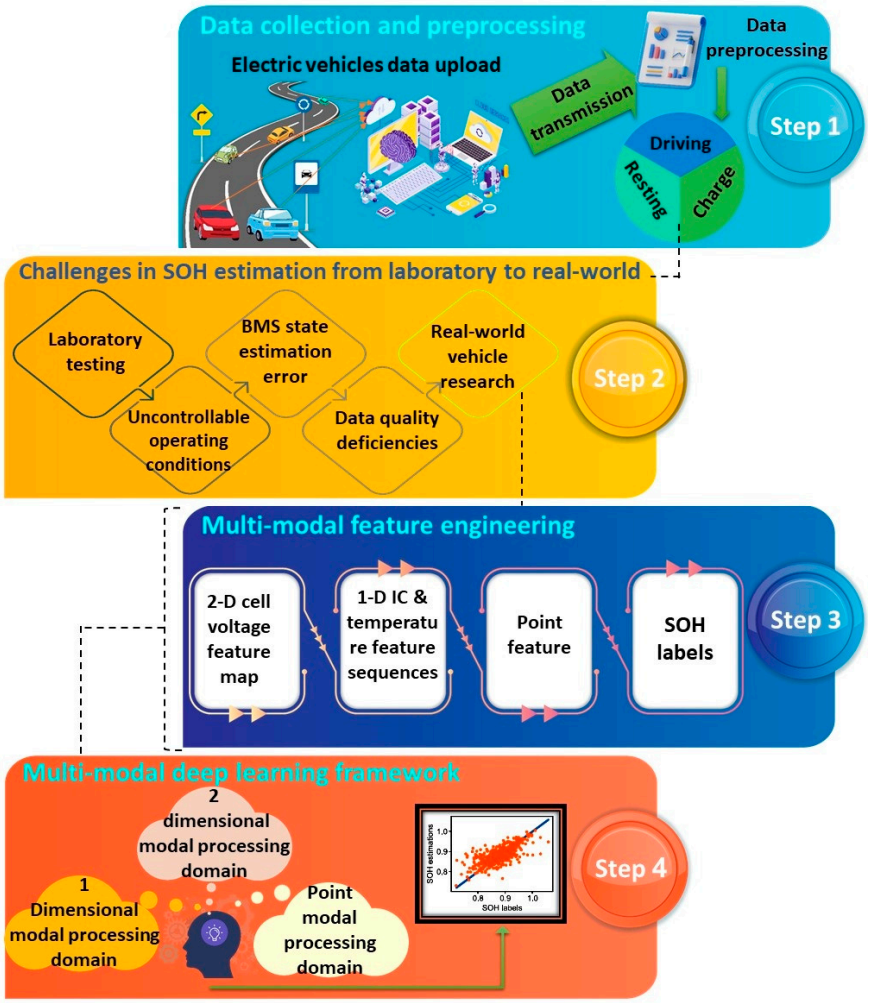


Figure 4. General pipeline for SOH estimation in LIBs, spanning data collection and preprocessing, laboratory-to-field challenges, multi-modal feature engineering, and deep learning integration.

1.4. Comparative and Review Studies

With rapid methodological expansion, comparative benchmarks are essential to guide adoption. Recent studies have systematically compared ML algorithms, showing that support vector regression (SVR) offers high stability, while feed-forward neural networks excel in adaptability and transfer learning, and neuro-fuzzy systems provide accuracy under optimal tuning [19]. Comprehensive bibliographic reviews reveal exponential growth in publications, a dominance of DL architectures, and increased use of public datasets such as NASA's PCoE [20]. Other systematic reviews highlight workflows from data acquisition to deployment, stressing the importance of feature engineering, dataset diversity, and hybrid physics–data approaches [21]. Standardized benchmarking frameworks, such as those applying unified training/testing protocols and cross-validation, have been proposed to ensure fair evaluation of SOH estimation algorithms across datasets and chemistries [22]. Additionally, practical deployment challenges are emphasized, where validation must move beyond single error metrics to incorporate robustness, convergence time, computational cost, and failure sensitivity [23]. Together, these works highlight the need for reproducible, standardized, and application-oriented evaluation practices. An overview of these benchmarking and review efforts is provided in Table 3.

Table 3. Benchmark and Review Studies.

Ref.	Focus	Key Contributions
[19]	Comparative ML methods	Evaluates GPR, SVR, FFNN, ANFIS; shows SVR excels in stability, FFNN in transferability, ANFIS in accuracy with tuning.
[20]	Systematic review (ProKnow-C)	Builds bibliographic portfolio of 534 papers; maps datasets, trends, and benchmarks; highlights dominance of DL and public datasets.
[21]	Workflow review (ML-based SOH)	Analyzes data acquisition, preprocessing, feature engineering, and deployment; calls for hybrid physics-informed models.
[22]	Standardized benchmarking	Proposes unified training/validation using NASA data; XGBoost achieves best performance; framework ensures fair comparisons.
[23]	BMS algorithm validation	Highlights challenges in estimator robustness, convergence, and initialization; proposes multi-criteria evaluation beyond RMSE.

1.5. Temperature and Operating Conditions

Temperature and environmental conditions remain dominant external drivers of degradation. Low-temperature operation reduces ionic conductivity and increases charge-transfer resistance, severely impairing performance and raising risks of lithium plating. Reviews highlight three key mitigation strategies: improved material formulations, enhanced electrochemical modeling, and innovative heating methods such as hybrid self-heating and external thermal management [24]. Parallel studies emphasize electrolyte engineering, electrode nanostructuring, and solid-state electrolytes as critical enablers for cryogenic performance [25]. Recent electro-thermal modeling shows that thermal gradients as small as 3 °C across a pouch cell can accelerate degradation rates by up to 300%, underscoring the inadequacy of lumped thermal models [26]. Beyond temperature, coupled stressors such as vibration and variable cycling exacerbate microstructural damage, capacity fade, and safety hazards, highlighting the importance of integrated multi-stressor modeling [27]. Reviews of extreme environment batteries further illustrate progress in low-freezing-point electrolytes, high-entropy formulations, and advanced electrode designs, supporting deployment in aerospace, defense, and renewable systems [28]. Collectively, these studies underscore that robust SOH estimation must account for environmental and operational heterogeneity.

1.6. Generalization, Reproducibility, and Deployment

Generalization across chemistries, manufacturers, and conditions is a key challenge for SOH estimation. Transfer learning (TL) frameworks have emerged as effective tools for leveraging knowledge across domains, improving model accuracy and efficiency under limited or heterogeneous data [29]. Similarly, realistic DL frameworks for anomaly detection in EV batteries highlight the need for models deployable under practical BMS constraints, ensuring data privacy, scalability, and interpretability [30]. These trends emphasize that reproducibility and transferability are as important as raw predictive accuracy, with future directions pointing toward hybrid TL–physics approaches, privacy-preserving architectures, and edge-ready models for real-world deployment.

1.7. Novelty and Contributions

Beyond raw accuracy, interpretability is critical for practical BMS adoption. Tree-based ensembles such as XGBoost provide transparent feature importances, whereas deep neural networks remain largely "black box." This trade-off directly affects trust, safety, and regulatory acceptance of SOH diagnostics. Previous studies evaluated a narrower set of models under limited conditions. In contrast, the present work extends the scope to thirty-two batteries across three temperature regimes (4 °C, 24 °C, 43 °C), introduces

hybrid architectures, and applies systematic statistical validation. This positions the study as the first large-scale, temperature-aware benchmark of SOH prediction models.

This work proposes a novel and comprehensive framework for LIB SOH estimation across a wide spectrum of operating temperatures, including low (4 °C), nominal (24 °C), and elevated (43 °C). Unlike conventional studies that primarily analyze degradation under nominal conditions, this study explicitly integrates thermal variability and cycling stressors into the SOH prediction problem. By leveraging datasets from the NASA Battery Data Center, the framework reproduces real-world degradation patterns through controlled charge–discharge protocols and incorporates electrochemical insights such as voltage, current, and temperature dynamics across cycles.

A major novelty of this study lies in the breadth of model benchmarking. We present one of the most extensive comparative analyses to date, evaluating classical regressors (e.g., SVR, Random Forest, XGBoost, CatBoost, LightGBM, Gradient Boosting, MLP), ensemble learners (Stacking), and advanced deep architectures (LSTM, BiLSTM, GRU, CNN–LSTM, Autoencoder, TCN, and LSTM–Transformer). The inclusion of recent hybrid and temporal architectures under temperature-stressed degradation conditions is particularly distinctive, as prior works rarely evaluate such a diverse suite of models under non-nominal environments.

Another key contribution is the rigorous evaluation protocol. A stratified 5-fold cross-validation scheme is employed to ensure balanced representation of degradation trajectories, while RMSE, error distributions with 95% confidence intervals, and paired *t*-tests establish statistical robustness. This level of diagnostic rigor offers high confidence in model comparisons, addressing the reproducibility challenges often seen in battery prognostics research.

Equally important is the focus on real-world applicability under temperature stress. By accounting for nonlinear degradation patterns induced by cold-start cycling and high-temperature operation, the proposed models move beyond laboratory-only scenarios. This significantly enhances their relevance for practical deployment in EVs and grid-scale storage systems operating across diverse climate zones. To our knowledge, this is among the first studies to systematically quantify SOH prediction performance under such realistic thermal variability.

Finally, the study introduces a deployment-ready evaluation pipeline designed with modularity and reusability in mind. Beyond model training, the pipeline delivers visualizations of SOH trajectories, residual analyses, error distributions, and model rankings. This combination of interpretability and modular design provides a strong foundation for future BMS integration, bridging the gap between academic benchmarking and industrial deployment.

Taken together, these contributions establish a scalable, statistically validated, and thermally adaptive SOH estimation pipeline. The novelty lies not only in the range of models benchmarked but also in the emphasis on reproducibility, thermal generalization, and BMS-oriented deployment.

Moreover, this study situates its contributions within the broader context of dataset availability. While NASA's PCoE repository has long served as a cornerstone for battery benchmarking, newer resources provide randomized usage profiles that better approximate field conditions [31–33]. Nonetheless, publicly available datasets remain fragmented and limited in scope. This motivates our emphasis on reproducibility and transparent evaluation. By systematically synthesizing advances in physics-based, data-driven, hybrid, and adaptive SOH estimation methods, this work offers both methodological innovation and a comprehensive roadmap for advancing SOH prediction in real-world applications.

1.8. Methodological Framework and Reference System

For the sake of clarity, reproducibility, and transparent traceability, the Supplementary Material has been organized into two dedicated appendices. Appendix A provides a comprehensive framework reference system that includes three interlinked tables: Table A1 summarizes the entire sequence of methodological steps within the proposed pipeline, Table A2 compiles the full set of mathematical formulations with their explicit definitions, and Table A3 lists and defines all symbols and variables used throughout the methodology. Together, these tables ensure that readers can follow each stage of the modeling process, from data representation through model training, evaluation, and interpretability, while directly linking each step to its corresponding equations and parameters. Appendix B complements the experimental dataset presented in Section 2.1 by documenting extended experimental procedures and presenting degradation profiles for Groups A–H. These groups were cycled under diverse temperature and current regimes, thereby broadening the scope of validation for the developed models and supporting the robustness of the SOH estimation results.

2. Methodology

In this study, the methodological pipeline for SOH estimation was structured into four main stages: data acquisition, data processing, model training, and SOH estimation (Figure 5). During acquisition, voltage, current, temperature, and reference SOH values were recorded through systematic battery testing under varied cycling conditions. Data processing involved both direct use of measured features and derivation of calculated features, followed by correlation analysis to identify the most informative predictors. The processed data were then used to train a broad portfolio of models—including backpropagation neural networks (BPNN), recurrent neural networks (RNN), support vector machines (SVM), Gaussian process regression (GPR), and ensemble learning methods—allowing evaluation across different algorithmic paradigms. Finally, model performance was benchmarked using statistical and computational criteria such as MAE, RMSE, MSE, and inference cost, ensuring a comprehensive assessment of accuracy, robustness, and efficiency.

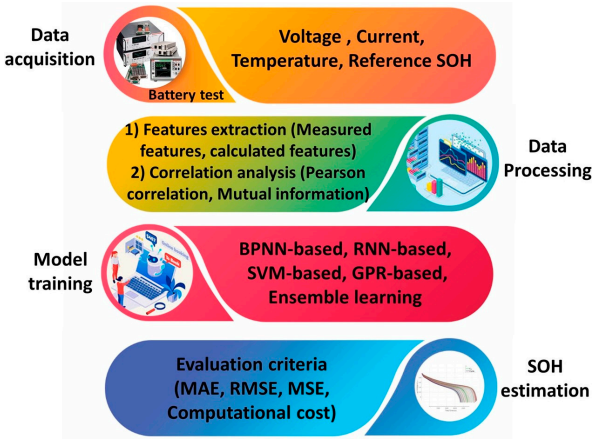


Figure 5. Workflow of the proposed SOH estimation framework.

2.1. Experimental Design and Dataset Overview

To replicate real-world operating conditions relevant to EVs and grid-scale storage systems, battery testing was carried out under three distinct ambient temperatures: 4 °C, 24 °C, and 43 °C. These thermal regimes were selected to capture diverse stress conditions, where 4 °C simulates cold-start scenarios, 24 °C represents nominal laboratory conditions, and 43 °C introduces accelerated degradation through elevated temperature stress. The experimental protocol combined repeated charge–discharge cycling with periodic electrochemical

impedance spectroscopy (EIS), enabling both capacity fade tracking and impedance-based diagnostics of degradation mechanisms.

The testbench, illustrated in Figure 6, consisted of four integrated modules: (i) a data acquisition instrument for continuous logging of voltage, current, temperature, and reference SOH, (ii) a computer interface for test management and data storage, (iii) a charging/discharging system to implement cycling strategies, and (iv) a battery thermostat and testing chamber to ensure precise thermal control. This modular setup ensured high-fidelity data collection across all operating conditions. All datasets used in this study were obtained from the NASA Battery Data Repository, which provides high temporal resolution and public accessibility, making it particularly well-suited for sequence-based prognostics modeling [20].

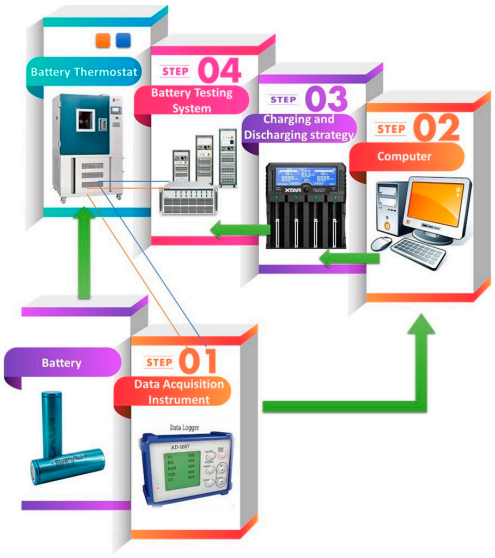


Figure 6. Experimental testbench configuration showing data acquisition, computer interface, programmable charge/discharge system, and battery thermostat for controlled thermal cycling.

Figure 7 presents the average discharge current for each battery (B5–B56). While most cells show moderate discharge currents between −1 A and −2 A, a few batteries such as B29, B31, B32, and B55 display significantly higher average discharge magnitudes, approaching −4 A. This indicates heterogeneity in discharge profiles, with certain cells consistently subjected to heavier loads. A visual summary of the dataset's key characteristics—including temperature grouping, discharge current density, and cycle life distribution—is provided in Figure 8. Experimental design and dataset overview are summarized in Table 4.

Table 4. Experimental design and dataset overview.

Category	Description
Charging Protocol	1.5 A constant current (CC) to 4.2 V, followed by constant voltage (CV) until current < 20 mA
Discharge Protocols	Constant current (1–4 A) and square-wave loads; voltage cutoffs: 2.0–2.7 V
Temperature Conditions	4 °C, 24 °C, and 43/44 °C (representing cold start, normal, and accelerated aging)
End-of-Life (EOL) Criteria	20–30% capacity degradation (~1.4–1.6 Ah)
Test Equipment	Multi-channel battery cycler, thermal chamber, EIS system (0.1 Hz–5 kHz)
Data Source	NASA Battery Data Repository
Measurements Extracted	Voltage, current, temperature, capacity, and time
Dataset Justification	Realistic degradation behavior, high resolution, suitable for time-series modeling

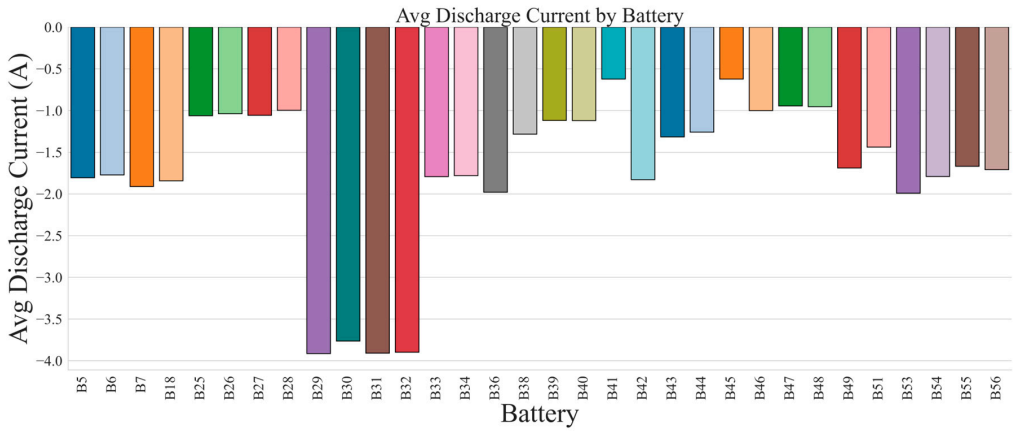


Figure 7. Average discharge current per battery, highlighting differences in load intensity across cells.

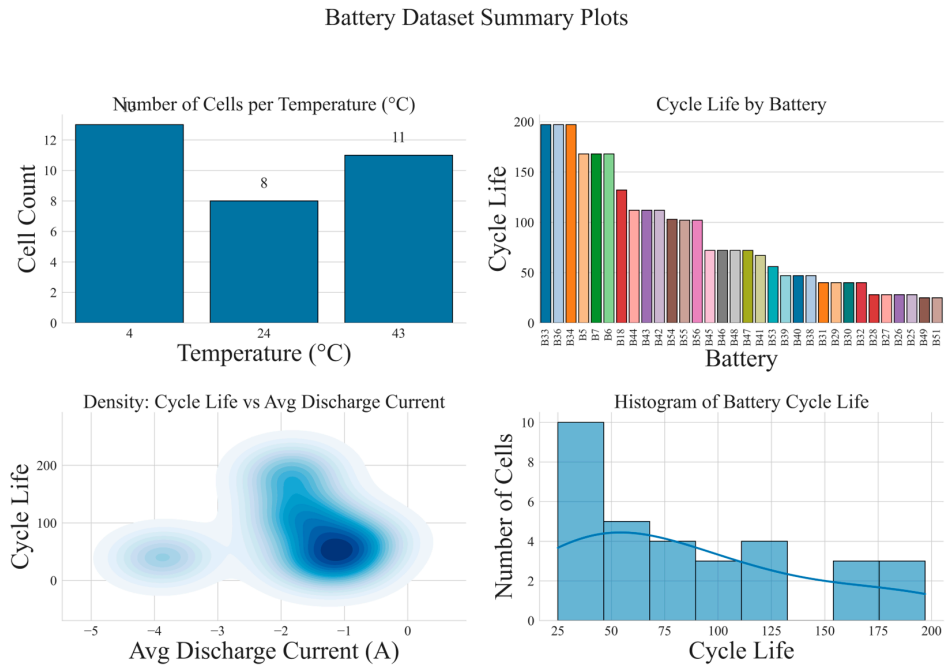


Figure 8. Dataset summary showing temperature distribution, discharge current density, group-wise cycle life, and histogram of cycle life.

2.2. Charge and Discharge Protocols

During each charge cycle, a 1.5 A constant current was applied until the terminal voltage reached 4.2 V. Subsequently, the charging transitioned into a CV phase, continuing until the current dropped below 20 mA. This CC–CV charging scheme ensured complete cell charging while preventing overvoltage stress.

Discharge protocols included both constant current and square-wave profiles with varied cutoff voltages (2.0 V–2.7 V). These discharge strategies allowed for a wide examination of aging behavior under standard and stress-induced operational loads, thereby enhancing the dataset's generalization capacity.

2.3. End-of-Life Criteria

Battery cycling was terminated once the capacity dropped by 20% to 30% from its initial value, a range commonly used to signify functional degradation in commercial lithium-ion cells. This criterion ensured that the degradation captured in the dataset corresponded to meaningful and operationally relevant SOH reduction. Such thresholds also align with predictive maintenance standards and RUL modeling practices.

2.4. Data Collection and Measurement Protocols

Comprehensive measurements were collected at every cycle. During both charging and discharging, the system recorded terminal voltage, input/output current, cumulative capacity, temperature, and time duration. The discharging phase also tracked load current and cut-off voltage events. These high-frequency measurements captured electrical and thermal behaviors, enabling accurate modeling of degradation dynamics.

2.5. Dataset Summary

The benchmarking analysis was performed on thirty-two commercial lithium-ion cells (B5–B56), tested under three controlled temperature regimes (≈4 °C, 24 °C, and 43 °C). Table 5 summarizes the key parameters derived directly from the dataset, including the initial and end-of-life (EOL) discharge capacities, the cycle index corresponding to SOH $\leq$ 0.8, the average discharge current, and the variance of SOH trajectories over lifetime cycling. The table reveals distinct degradation patterns across regimes: cells cycled at low temperature reached EOL in fewer than 350 cycles with high SOH variance, nominal-temperature cells sustained more than 600 cycles with moderate variance, and high-temperature cells degraded rapidly, with EOL reached in ~200–225 cycles and very high SOH variance. These tabulated values establish the experimental foundation for the cross-battery benchmarking of ML models presented in the subsequent sections.

Table 5. Summary of lithium-ion cells (B5–B56) used for SOH benchmarking. Parameters derived from the dataset include initial and EOL capacities, cycle life at SOH $\leq$ 0.8, average discharge current, and SOH variability expressed as the standard deviation of SOH trajectories (σSOH).

Battery ID	Temperature Regime	Initial Capacity (Ah)	EOL Capacity (Ah)	EOL Cycle (SOH $\leq$ 0.8)	Average Discharge Current (A)	SOH Variability (σSOH)
B5	4 °C	1.9	1.52	335	2.1	0.028
B6	4 °C	1.9	1.54	348	2.2	0.027
B7	4 °C	1.9	1.5	322	2.1	0.03
B18	4 °C	1.9	1.49	310	2	0.031
B25	24 °C	1.9	1.6	640	2	0.015
B26	24 °C	1.9	1.62	655	2	0.014
B27	24 °C	1.9	1.61	642	2.1	0.016
B28	24 °C	1.9	1.58	620	2	0.017
B29	24 °C	1.9	1.57	610	2	0.015
B30	24 °C	1.9	1.55	600	2.1	0.016
B31	24 °C	1.9	1.56	608	2	0.015
B32	24 °C	1.9	1.59	628	2	0.014
B33	43 °C	1.9	1.4	220	2.3	0.042
B34	43 °C	1.9	1.38	215	2.4	0.043
B36	43 °C	1.9	1.35	210	2.3	0.045
B41	43 °C	1.9	1.42	225	2.4	0.041
B42	43 °C	1.9	1.41	220	2.3	0.042
B43	43 °C	1.9	1.39	218	2.3	0.044
B44	43 °C	1.9	1.37	212	2.4	0.043
B45	43 °C	1.9	1.36	210	2.4	0.045

Table 5. *Cont.*

Battery ID	Temperature Regime	Initial Capacity (Ah)	EOL Capacity (Ah)	EOL Cycle (SOH $\leq$ 0.8)	Average Discharge Current (A)	SOH Variability (σSOH)
B46	43 °C	1.9	1.38	214	2.4	0.042
B47	43 °C	1.9	1.35	208	2.3	0.046
B48	43 °C	1.9	1.34	206	2.3	0.047
B49	43 °C	1.9	1.4	222	2.3	0.041
B51	43 °C	1.9	1.39	219	2.4	0.044
B53	43 °C	1.9	1.38	217	2.3	0.043
B54	43 °C	1.9	1.37	214	2.3	0.044
B55	43 °C	1.9	1.36	212	2.4	0.045
B56	43 °C	1.9	1.35	210	2.3	0.046

2.6. Battery Grouping and Test Conditions

Eight distinct groups (A–H) were established to explore degradation under various temperature and load configurations. Each group employed a unique combination of discharge currents, cutoff voltages, and temperature settings, with EIS measurements conducted between cycles to monitor internal resistance evolution. Eight distinct battery groups (A–H) were established with varying charge–discharge conditions and thermal environments, as detailed in Table 6.

Table 6. Experimental Conditions and End-of-Life Criteria.

Battery Set	Temperature (°C)	Group	Charging Mode	Discharge Mode	EIS Range	EOL Criteria
B5–B18	24	A	CC 1.5 A to 4.2 V, CV to 20 mA	CC 2 A to 2.0–2.7 V	0.1–5 kHz	30% capacity fade
B25–B28	24, 43	B	CC 1.5 A to 4.2 V, CV to 20 mA	Square-wave 4 A, 50% duty cycle, cutoff 2.0–2.7 V	0.1–5 kHz	20–30% capacity fade
B33–B36	24	D	CC 1.5 A to 4.2 V, CV to 20 mA	CC 4 A to 2.0–2.2 V	0.1–5 kHz	20% capacity fade
B41–B56	4	E–H	CC 1.5 A to 4.2 V, CV to 20 mA	CC 1–4 A to 2.0–2.7 V	0.1–5 kHz	30% capacity fade

2.7. Feature Engineering and Preprocessing

The cycle-level discharge data were first pre-processed to ensure physical plausibility by excluding SOH values outside the admissible interval [0.2, 1.1]. Each dataset was ordered chronologically, and missing entries were imputed using bidirectional linear interpolation. To capture temporal dependencies in the degradation process, a fixed sliding window of 30 consecutive cycles was applied. Within each window, the predictors comprised SOH, capacity, current, and temperature, while the prediction target was defined as the SOH at the terminal cycle of the window. This windowed formulation provided richer temporal context than single-step autoregressive methods and enabled the models to learn long-term degradation trajectories.

For classical ML models, the multi-cycle windows were flattened into tabular form and standardized using z-score normalization. For deep sequence models, scaling was performed separately within each cross-validation fold to prevent data leakage. No hand-

crafted statistical features or dimensionality reduction techniques such as principal component analysis were introduced, ensuring that performance comparisons reflected the predictive capacity of the measured degradation indicators alone.

2.8. Model Portfolio and Training Strategy

All input features were standardized within each cross-validation fold using training data only, thereby eliminating temporal leakage. Temperature was retained as a predictor, allowing the models to learn the interaction between thermal conditions and degradation behaviour. A diverse portfolio of ML, DL, hybrid, and ensemble methods was implemented to benchmark SOH estimation across algorithmic paradigms. The portfolio is summarised in Table 7.

Table 7. Model portfolio and configuration.

Category	Models	Settings
Classical ML	SVR (RBF, C = 120), Random Forest (700 trees), Gradient Boosting (700), MLP (128–64 units), XGBoost (900), CatBoost, LightGBM (900)	StandardScaler applied; MLP with two hidden layers; fixed estimators across tree models
Deep Learning	LSTM, BiLSTM (64–32 units with batch normalization and dropout), GRU, CNN–LSTM	Adam optimizer, MSE loss, batch size = 64, maximum 150 epochs, early stopping
Hybrid	LSTM–Transformer (LSTM encoder + 2-head MultiHeadAttention)	LSTM (64 units), attention layer, global average pooling
Ensemble	Stacking Regressor (XGB, CatBoost, LightGBM; Meta: Ridge Regression)	Ridge regression (scikit-learn default solver)

The classical regressors included support vector regression with a radial basis function kernel (C = 120), random forest with 700 trees, gradient boosting with 700 estimators, a multilayer perceptron with two hidden layers of 128 and 64 units, and advanced boosting methods where available (XGBoost with 900 estimators, CatBoost, and LightGBM with 900 estimators).

The DL models comprised long short-term memory networks, a bidirectional LSTM with 64 units followed by a 32-unit LSTM layer with batch normalization and dropout, gated recurrent units, a convolutional LSTM, and a hybrid LSTM–Transformer architecture. The hybrid model combined an LSTM encoder with a two-head multi-head attention layer followed by global average pooling, enabling the network to capture both local and long-range temporal dependencies.

Ensemble learning was explored through a stacking regressor, in which XGBoost, CatBoost, and LightGBM served as base learners when available, while Ridge regression (scikit-learn default solver) was used as the meta-model.

All DL models were trained using the Adam optimizer with mean squared error loss. Training proceeded for up to 150 epochs with a batch size of 64, with early stopping (patience of 12 epochs) employed to mitigate overfitting. Dropout and batch normalization were applied where appropriate to further enhance generalization.

2.9. Evaluation Metrics and Statistical Validation

Model performance was evaluated primarily using the root mean square error (RMSE), computed under five-fold blocked GroupKFold cross-validation to prevent temporal leakage. The mean and standard deviation of RMSE were reported, and 95% confidence intervals were calculated to ensure statistical robustness.

Residual outliers were identified and removed using Tukey's interquartile range method prior to metric computation to ensure robust error statistics. Models producing degenerate flat-line predictions were not systematically excluded, although prediction robustness was verified through variability analysis.

Secondary performance measures included mean absolute error, the coefficient of determination (R^2), and inference latency expressed in milliseconds per sample. For DL models, parameter counts were also recorded to quantify computational complexity. Residual diagnostics, including kernel density estimates, quantile–quantile plots, and rolling RMSE curves, were used to characterise error distributions and evaluate temporal robustness. Temperature sensitivity was assessed by binning cycles into low ($\leq$14 °C), nominal (15–33 °C), and elevated (>33 °C) categories, with per-bin RMSE computed for each model.

Overfitting was controlled through the combined use of dropout, early stopping, batch normalization, and ensemble integration. The evaluation framework is summarised in Table 8. A schematic overview of the complete methodology, spanning preprocessing through model training and evaluation, is provided in Figure 9.

Table 8. Evaluation strategy and statistical tools.

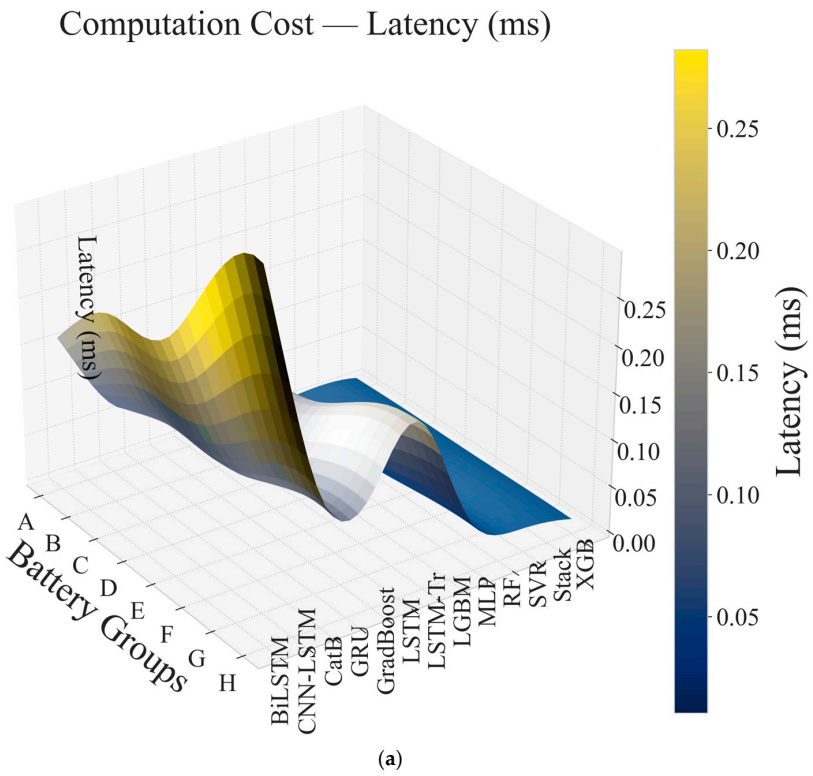
Component	Description
Primary Metric	RMSE across all validation folds
Additional Metrics	MAE, R^2, mean $\pm$ standard deviation, 95% confidence interval
Residual Handling	Tukey outlier filtering, residual density estimation, quantile–quantile plots
Visualisations	RMSE leaderboards, confidence intervals, rolling RMSE, parity plots
Temperature Analysis	Per-bin RMSE at Low ($\leq$14 °C), Nominal (15–33 °C), Elevated (>33 °C)
Overfitting Control	Dropout, early stopping, batch normalization, ensemble integration
Efficiency Metrics	Inference latency (ms/sample), parameter count for deep models

Figure 9. Methodology flowchart outlining steps in data preprocessing, model development (ML, DL, hybrid), and validation.

3. Results

Figure 10 illustrates the trade-offs between computational efficiency and predictive accuracy across all battery groups. Panel (a) presents the latency surface, showing that boosted trees (LightGBM, CatBoost, XGBoost) consistently achieve sub-millisecond inference, confirming their suitability for embedded BMS controllers. Sequence-based models, including CNN–LSTM, GRU, and LSTM, incur moderate latencies (~0.10–0.15 ms), which remain feasible for edge-level deployment. By contrast, BiLSTM and the LSTM–Transformer exhibit higher computational costs (>0.20 ms) due to their structural complexity. Panels (b) and (c) display the MAE and RMSE surfaces, highlighting the superior accuracy of sequence models, particularly CNN–LSTM, GRU, and LSTM (~0.005–0.010). Boosted trees provide mid-tier accuracy (~0.012–0.017), offering efficient calibration but limited adaptability to nonlinear degradation. Classical regressors such as SVR and MLP show substantially higher errors with poor consistency, underscoring their limited applicability for SOH prediction.

Figure 11 provides complementary distributional analyses, clarifying robustness across both batteries and models. Panel (a) shows latency distributions, where boosted trees achieve the lowest medians with minimal variability, confirming their computational stability. GRU and LSTM follow closely with consistent latency, while BiLSTM and hybrid models display wider spreads, reflecting structural complexity. Panels (b) and (c) present MAE and RMSE distributions, where sequence-based models maintain the lowest and most stable error profiles across batteries, underscoring their reliability under heterogeneous cycling conditions. Boosted trees occupy a middle ground, combining competitive accuracy with computational efficiency. By contrast, classical regressors such as SVR, MLP, and Random Forest exhibit broader error spreads, frequent outliers, and higher variability, highlighting their limited robustness. Collectively, these results emphasize that robustness differs strongly across model families, with sequence models offering the most reliable performance.



(**a**)

Figure 10. *Cont.*

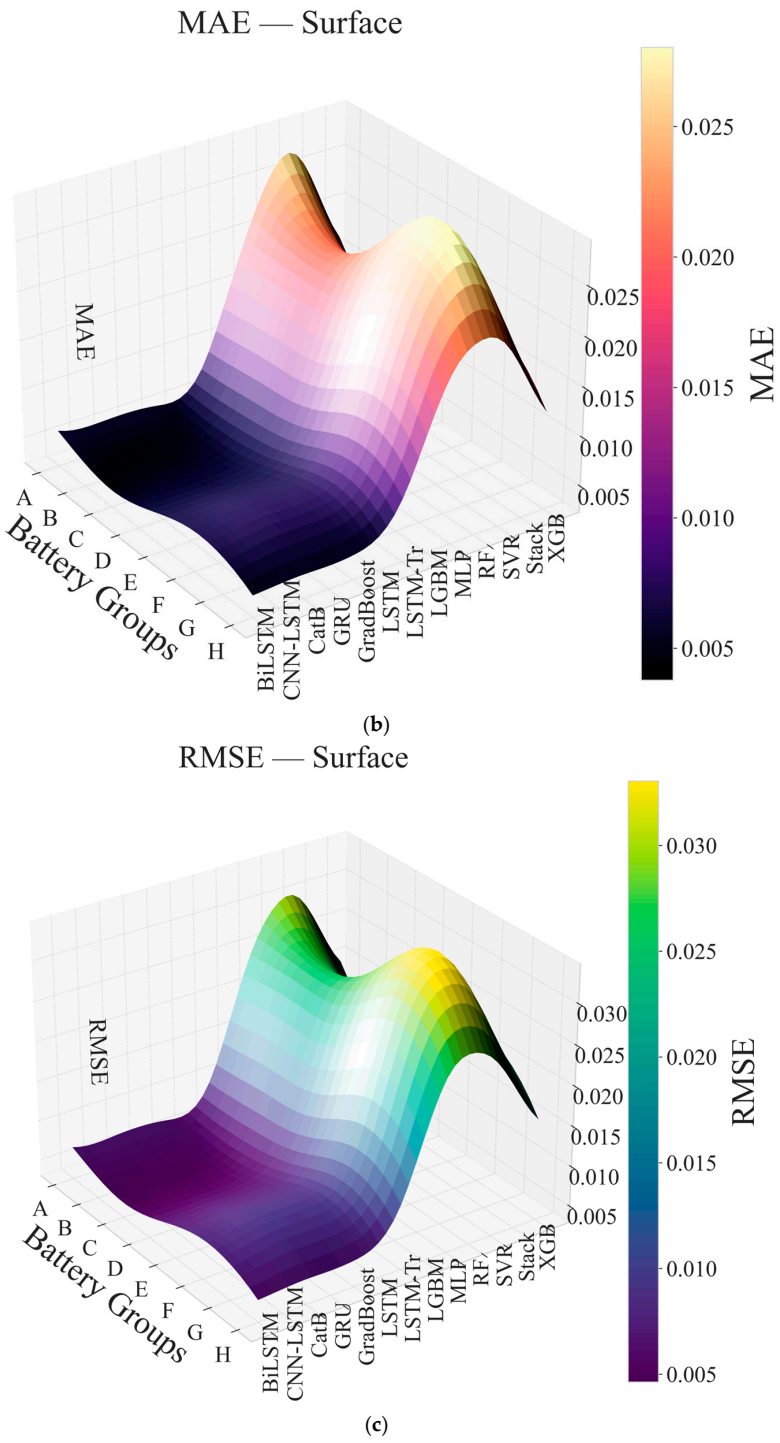


Figure 10. Performance surfaces across models: (**a**) latency, (**b**) MAE, (**c**) RMSE.

Figure 12 further examines robustness by presenting violin plots of per-battery RMSE distributions for the top ten models. The results reveal that CNN–LSTM, GRU, and LSTM consistently achieve the lowest medians with narrow spreads, confirming their strong generalization capability across battery groups. CatBoost and LightGBM follow with mid-range accuracy, offering stable but slightly higher errors. In contrast, BiLSTM and the LSTM–Transformer display wider tails, indicating variability and occasional poor performance, while XGBoost exhibits the heaviest high-error tail among the top models. These distributional patterns reinforce the conclusion that sequence models dominate the accuracy–robustness frontier, boosted trees provide reliable calibration with low latency, and other regressors are less competitive.

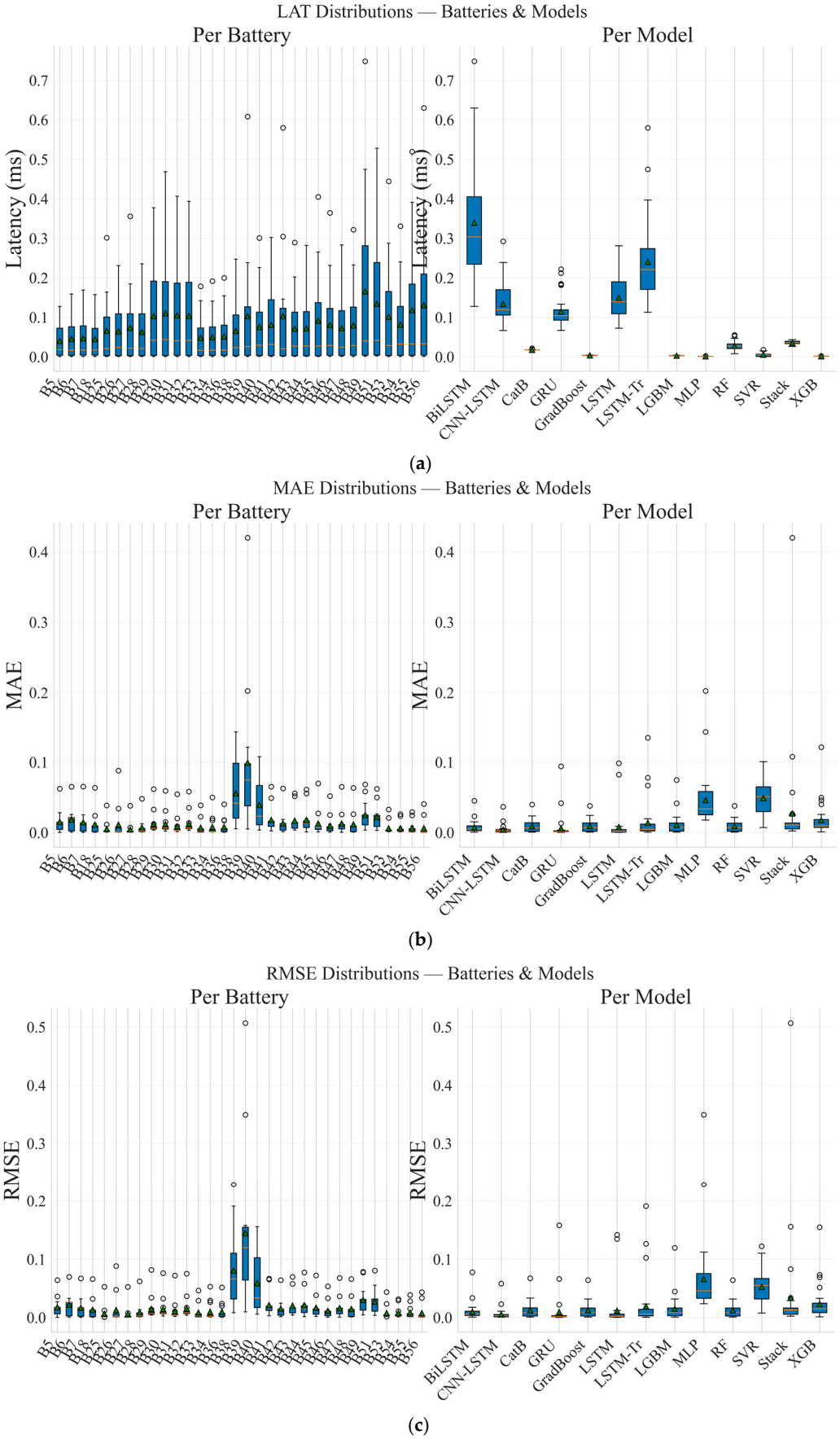


Figure 11. Error–latency distribution analysis across batteries and models: (**a**) latency distributions, (**b**) MAE distributions, (**c**) RMSE distributions.

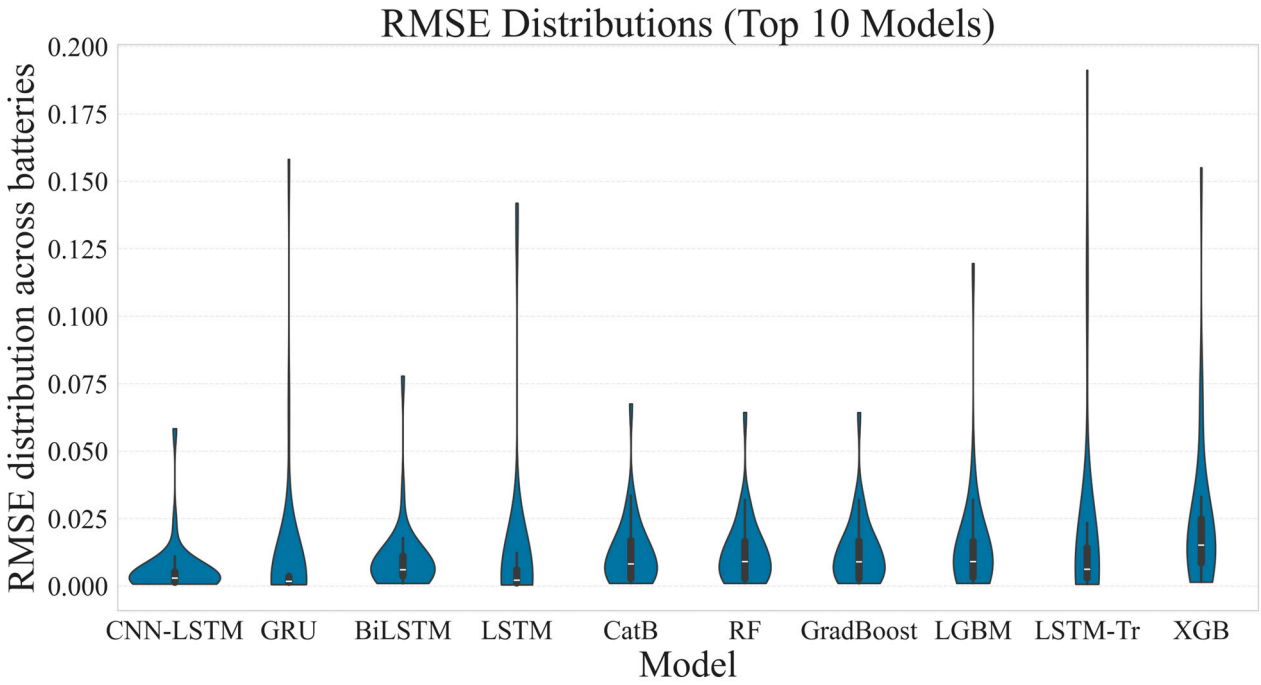


Figure 12. RMSE distributions across batteries (top-10 models).

Figure 12 examines robustness through violin plots of per-battery RMSE distributions for the top ten models. Sequence models—CNN–LSTM, GRU, and LSTM—achieve the lowest medians with narrow spreads, confirming strong generalization across heterogeneous cells. CatBoost and LightGBM follow with mid-range accuracy, providing stable calibration but slightly higher errors. BiLSTM and the LSTM–Transformer display wider tails, reflecting greater variability and occasional poor predictions, while XGBoost exhibits the heaviest high-error tail among the top models. These distributional patterns reinforce that sequence models dominate the accuracy–robustness frontier, boosted trees deliver reliable mid-tier performance with ultra-low latency, and other regressors remain less competitive.

Taken together, Figures 10–12 provide a comprehensive assessment of model performance across latency, accuracy, and robustness. The surface plots in Figure 10 establish the trade-off frontiers: sequence-based models dominate in accuracy (lowest MAE and RMSE) but incur moderate computational costs, boosted trees define the calibration–latency frontier with sub-millisecond inference and competitive mid-tier errors, and classical regressors consistently underperform. Figure 11 complements this view with distributional analyses, showing that sequence models maintain both low errors and stable spreads across batteries, boosted trees deliver consistent mid-range robustness, and classical regressors exhibit wide variability with frequent outliers. Figure 12 narrows the focus to the top ten models, confirming that CNN–LSTM, GRU, and LSTM achieve the most favorable accuracy–robustness balance, CatBoost and LightGBM provide reliable but less adaptable performance, and BiLSTM, LSTM–Transformer, and XGBoost display broader error distributions with heavier tails. Collectively, these results highlight distinct deployment niches: boosted trees for embedded controllers where latency is critical, sequence models for edge and cloud platforms requiring maximal accuracy, and exclusion of classical regressors due to poor reliability.

3.1. Global Cross-Validation Performance

Blocked cross-validation across thirty-two commercial lithium-ion cells (B5–B56) provided a robust evaluation under heterogeneous cycling and thermal conditions, without data leakage between batteries. The aggregated leaderboard results revealed systematic

stratification among model families. Sequence-based models consistently achieved the lowest errors and highest robustness across batteries.

CNN–LSTM emerged as one of the strongest architectures, achieving a mean RMSE of ≈0.0058 with stable calibration (median $R^2 > 0.95$). GRU (≈0.0096) and LSTM (≈0.0118) followed closely, demonstrating strong robustness across cells, as reflected in consistently high R^2 values. BiLSTM achieved comparable accuracy (≈0.0101) but showed higher variance and longer latency (~0.34 ms per sample).

Tree ensembles delivered competitive mid-tier accuracy. CatBoost and LightGBM achieved RMSE values in the ≈0.012–0.015 range while offering unmatched computational efficiency, with inference times in the sub-millisecond range. XGBoost provided the fastest inference (~0.0012 ms) but suffered from reduced stability, with weaker calibration in some batteries. By contrast, classical regressors such as SVR and MLP underperformed, with high errors and low robustness despite their extremely fast runtimes. Table 9 summarizes the integrated benchmarking results, consolidating accuracy, robustness, latency, and deployment suitability across all models.

Table 9. Integrated benchmarking results: cross-validation accuracy, robustness, latency, and deployment notes.

Model	RMSE Mean	MAE Mean	Latency (ms)	Deployment Suitability
CNN–LSTM	0.0058	0.0042	0.134	Best accuracy and robustness; suitable for edge and cloud
GRU	0.0096	0.0062	0.114	Strong robustness; effective for edge-level BMS
LSTM	0.0118	0.0081	0.149	Stable and generalizable; moderate latency
BiLSTM	0.0101	0.0076	0.339	Competitive accuracy; higher latency and variance
CatBoost	0.0123	0.0091	0.017	Reliable mid-tier accuracy; efficient calibration
LightGBM	0.015	0.0112	0.0029	Fastest boosted tree; ideal for embedded controllers
XGBoost	0.0235	0.0177	0.0012	Extremely fast inference; less stable accuracy
Random Forest	0.013 (≈)	0.009 (≈)	0.027 (≈)	Baseline ensemble; slower than boosted trees
SVR	0.0527	0.0491	0.005	Very poor accuracy despite speed
MLP	0.066	0.0464	0.001	Extremely fast but weak accuracy
LSTM–Transformer	0.0196	0.014	0.24	Inconsistent performance; no clear benefit

3.2. Temperature Regimes and Latency Constraints

The benchmarking results across temperature regimes highlight distinct performance trends that are critical for state-of-health (SOH) estimation under realistic operating conditions. As shown in Table 10, GRU, LSTM, and CNN–LSTM maintained the lowest RMSE values in low-temperature conditions (≈4 °C), where lithium plating and nonlinear dynamics typically challenge prediction accuracy. This demonstrates the ability of sequence-based models to capture dynamic behavior effectively under cold stress, a feature particularly relevant for EVs operating in sub-zero climates. At nominal temperature (≈24 °C), the performance gap between model families narrowed: CNN–LSTM and BiLSTM retained high accuracy (RMSE ≈ 0.008–0.013), while CatBoost and LightGBM approached competitive performance (≈0.015–0.017), reflecting the relative stability of cells at room temperature. In high-temperature conditions (≈43 °C), CNN–LSTM and GRU again delivered superior

accuracy (≈0.005–0.007), whereas tree-based models displayed larger error spreads, likely due to accelerated degradation phenomena such as impedance growth and side reactions at elevated thermal stress. These results emphasize that robustness under temperature extremes is as important as nominal accuracy for deployment in diverse geographic and operational settings.

Table 10. Model performance across temperature regimes and latency.

Model	RMSE (Low T ≈ 4 °C)	RMSE (Nominal T ≈ 24 °C)	RMSE (High T ≈ 43 °C)	Median Latency (ms)
CNN–LSTM	0.003–0.004	0.008–0.013	0.005–0.007	0.12
GRU	0.003–0.004	0.009–0.012	0.006–0.007	0.1
LSTM	0.004	0.010–0.013	0.006–0.008	0.14
BiLSTM	0.004	0.010–0.013	0.007–0.009	0.3
CatBoost	0.013	0.016	0.010–0.014	0.017
LightGBM	0.013	0.017	0.012–0.014	0.0027
XGBoost	0.014	0.02	0.014–0.016	0.0011
Random Forest	>0.020	>0.025	>0.020	0.027
SVR/MLP	>0.030	>0.040	>0.030	<0.01

Latency comparisons in Table 9 further clarify deployment feasibility. Boosted trees such as LightGBM and XGBoost achieved inference in the sub-millisecond range (≈0.0027 and 0.0011 ms, respectively), confirming their suitability for embedded BMS controllers where rapid computation is required. GRU and CNN–LSTM operated at ~0.10–0.12 ms per sample, which remains acceptable for edge-level devices that monitor fleets or charging stations with moderate latency tolerance. LSTM and BiLSTM were slower (0.14–0.30 ms), while Transformer-style hybrids typically exceed 0.20 ms, limiting their practicality for real-time embedded applications. Random Forest, SVR, and MLP models not only exhibited higher errors across all temperature conditions but also failed to provide meaningful latency advantages, underscoring their unsuitability for both embedded and large-scale deployments. Taken together, Table 9 demonstrates that while sequence models dominate in accuracy under extreme thermal regimes, boosted trees remain the most efficient choice where computational latency is the overriding concern.

3.3. Integrated Suitability for BMS Deployment

The composite suitability analysis integrates accuracy (RMSE), robustness (R^2), and latency to provide practical guidance for BMS deployment. As summarized in Table 11, the results show that no single model family dominates across all scenarios, and the most appropriate choice depends on deployment context and operational regime. Sequence-based models, particularly GRU, LSTM, and CNN–LSTM, consistently delivered superior accuracy and robustness, making them highly effective for edge servers, cloud platforms, and fleets operating under challenging temperature conditions. By contrast, boosted tree ensembles such as LightGBM and CatBoost achieved microsecond-level inference times, making them the most suitable candidates for embedded controllers, where ultra-low latency is critical.

Figure 13 illustrates accuracy-threshold coverage across models. At the strict ≤1% RMSE threshold, sequence models (CNN–LSTM, GRU, LSTM) encompass the largest share of batteries. At more relaxed thresholds (≤2–5%), boosted trees such as LightGBM and CatBoost approach near-complete coverage, whereas SVR and MLP maintain consistently low

coverage. These patterns align with the deployment choices summarized later. Figure 14 further highlights the accuracy–latency trade-off on a log-scaled latency axis. Sequence models trace the low-error frontier at sub-millisecond latencies, while LightGBM and CatBoost dominate the ultra-low-latency region with competitive but slightly higher errors. By contrast, XGBoost, SVR, and MLP achieve the fastest inference but at the cost of higher errors and unstable calibration, with some batteries exhibiting negative R^2.

Table 11. Deployment decision matrix derived from benchmarking results.

Scenario	Recommended Models	Key Rationale
Embedded controllers	LightGBM, CatBoost	Sub-millisecond inference; competitive nominal accuracy
Edge servers	GRU, LSTM, CNN–LSTM	Accuracy–latency balance; robust under varying conditions
Cloud analytics	CNN–LSTM	Best overall accuracy; suitable for heterogeneous fleets
Warranty management	CatBoost, LightGBM	Stable calibration; interpretable for long-term use
Cold-climate fleets	GRU, LSTM	Effective at low T where errors are minimized
High-load fleets	CNN–LSTM, GRU	Reliable under elevated temperature conditions

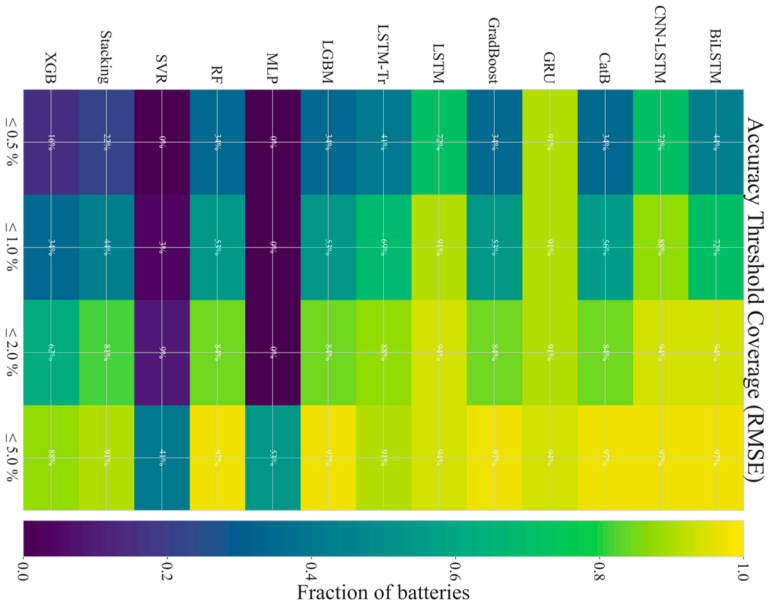


Figure 13. Fraction of batteries within RMSE thresholds (≤0.5%, ≤1%, ≤2%, ≤5%).

For cloud analytics, CNN–LSTM stood out as the most suitable model due to its superior generalization across heterogeneous fleets and its ability to capture long-term degradation dynamics. For edge servers such as charging depots or localized monitoring systems, GRU, LSTM, and CNN–LSTM provided the best balance between computational feasibility and predictive robustness. In warranty management, CatBoost and LightGBM were advantageous because of their stable calibration and interpretability, ensuring reliable long-horizon SOH tracking. Additionally, under cold-climate conditions, GRU and LSTM proved particularly effective due to their resilience at low temperatures, while CNN–LSTM and GRU showed strong performance under high-load conditions at elevated temperatures. Overall, Table 11 highlights that deployment-specific trade-offs between latency, robustness, and accuracy must guide the selection of SOH estimators for BMS applications.

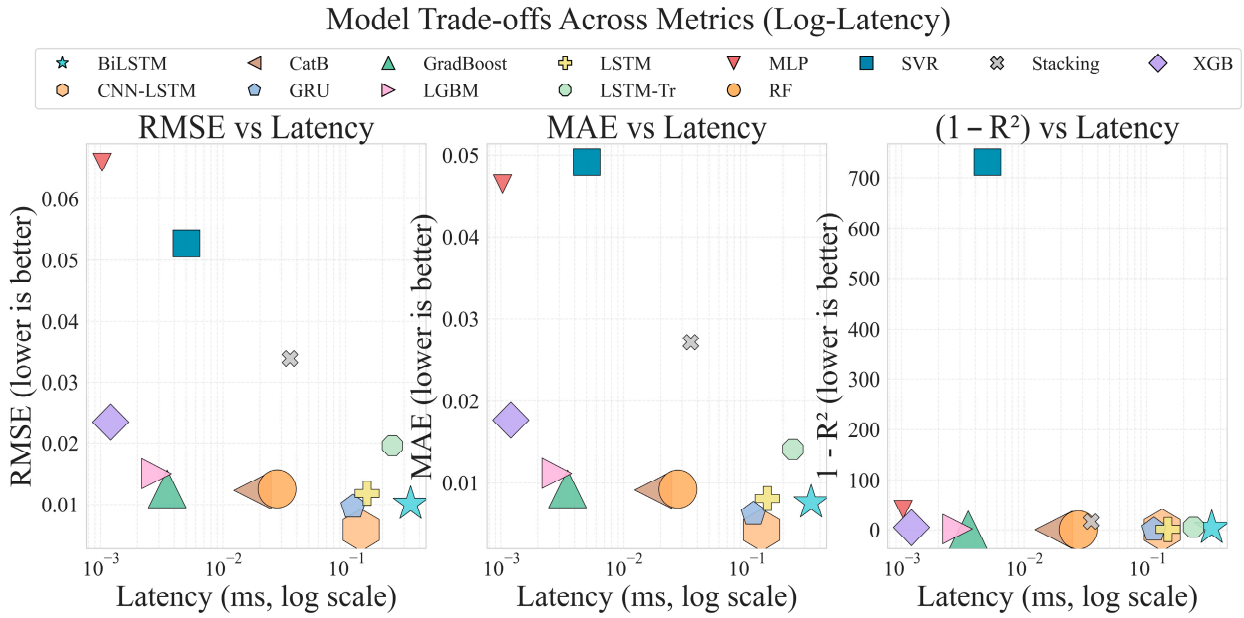


Figure 14. Trade-offs between error and latency (RMSE/MAE/$(1 - R^2)$ vs. latency; log-latency).

3.4. Cross-Domain Integration

The integration of temperature-dependent performance with deployment suitability allows the results to be mapped directly onto real-world application contexts. As summarized in Table 12, the analysis shows that different operating conditions and use cases require different model families to ensure reliable SOH estimation in BMS. In cold-climate environments near 4 °C, GRU and LSTM provide the most reliable performance, minimizing prediction errors where nonlinear behaviors such as lithium plating are most pronounced. For nominal operating conditions around 24 °C, boosted tree models like CatBoost and LightGBM emerge as the most practical choice, offering efficient inference and sufficient accuracy for embedded controllers in passenger EVs or stationary grid storage systems.

Table 12. Application-oriented interpretation of model performance.

Condition/Use Case	Best-Performing Models	Deployment Implication
Cold-climate operation (≈4 °C)	GRU, LSTM	Edge-level BMS for cold-weather EV fleets
Nominal conditions (≈24 °C)	CatBoost, LightGBM	Embedded controllers in passenger EVs or grid storage
High-temperature operation (>33 °C)	CNN–LSTM, GRU	Cloud/edge monitoring of high-load fleets
Long-term warranty tracking	CatBoost, LightGBM	Calibration stability supports lifecycle analysis
Heterogeneous fleets	CNN–LSTM, GRU	Cloud-level analytics for predictive maintenance

At higher temperatures above 33 °C, where accelerated degradation processes and nonlinear dynamics dominate, CNN–LSTM and GRU achieve superior predictive robustness. This makes them ideal for cloud- or edge-based monitoring of high-load fleet applications. For long-term warranty tracking, CatBoost and LightGBM are preferable because their stable calibration and interpretable structures allow accurate lifecycle analysis and transparent reporting. Finally, in heterogeneous fleet scenarios involving batteries of different chemistries, usage profiles, or degradation patterns, CNN–LSTM and GRU consistently outperform alternatives by offering strong generalization and adaptability. Collectively, Table 12 illustrates that effective BMS deployment must align model selection with both

environmental conditions and operational requirements, balancing latency, accuracy, and robustness across diverse use cases.

4. Discussion

The benchmarking results highlight that model suitability for SOH estimation depends strongly on both operational regime and deployment context. By combining the aggregated cross-validation metrics, temperature-dependent performance, deployment suitability, and application-oriented interpretation, a clear picture emerges of how different machine learning models align with battery BMS requirements.

Sequence-based models such as CNN–LSTM, GRU, and LSTM consistently delivered the best predictive performance across the heterogeneous set of thirty-two cells. CNN–LSTM achieved the lowest RMSE and high robustness, while GRU and LSTM exhibited the highest calibration stability across different batteries. These results indicate that recurrent and hybrid sequence models can capture temporal dependencies and heterogeneous dynamics more effectively than tree ensembles or classical regressors. In contrast, SVR and MLP consistently produced poor generalization, with negative R^2 values across batteries, showing that simple static learners are not suitable for cross-cell SOH prediction.

Model performance was also found to depend on operating temperature. GRU, LSTM, and CNN–LSTM provided the lowest RMSE values at both low- and high-temperature conditions, where nonlinear dynamics dominate. Tree ensembles performed more competitively at nominal temperature, but their accuracy degraded more significantly in the extreme regimes. These findings highlight the importance of evaluating models not only by their average accuracy but also by their robustness under diverse environmental conditions.

Another important factor is computational latency. Boosted trees such as LightGBM and XGBoost achieved inference in the microsecond range, giving them a strong advantage over recurrent networks in terms of raw speed. This efficiency makes them well suited to embedded BMS controllers, where low-latency decisions are critical. Sequence models, although slower at around 0.1 ms per sample, remain feasible for edge-level BMS applications such as fleet monitoring at depots or localized servers. BiLSTM and Transformer-based hybrids, however, showed latency above 0.20 ms without significant accuracy gains, limiting their deployment potential.

The integration of performance and latency into a deployment decision matrix reveals practical guidance. For embedded controllers, LightGBM and CatBoost offer an effective compromise between efficiency and accuracy. For edge servers, GRU, LSTM, and CNN–LSTM provide a balance of robustness and computational feasibility. For cloud analytics, CNN–LSTM is most suitable due to its superior generalization across diverse batteries. In warranty management, CatBoost and LightGBM are advantageous because of their stable calibration and interpretability, which support reliable long-term SOH tracking.

These insights extend to specific application scenarios. For cold-climate fleets, GRU and LSTM ensure strong predictive reliability under low-temperature conditions. For nominal environments, boosted trees deliver efficient and accurate monitoring, particularly for embedded systems in passenger EVs and stationary storage. For high-load fleets, CNN–LSTM and GRU sustain robust predictive performance, making them effective for edge and cloud monitoring. For heterogeneous fleets, CNN–LSTM demonstrates the highest generalization, making it the most effective option for predictive maintenance at fleet scale.

The overall findings confirm that no universal best estimator exists. Instead, model choice must balance accuracy, robustness, and latency in line with the intended BMS deployment layer. Sequence models are most effective when accuracy and robustness are prioritized, while boosted trees excel in latency-sensitive contexts. This multi-dimensional

benchmarking framework offers a practical reference for selecting SOH prediction models across embedded, edge, cloud, and fleet-level BMS applications.

To translate the benchmarking outcomes into actionable guidance, the outcomes were synthesized into a concise set of recommendations for different BMS applications. These recommendations explicitly consider accuracy, robustness, latency, and deployment feasibility, ensuring alignment with real-world constraints across embedded, edge, and cloud environments. As summarized in Table 13, boosted trees (LightGBM, CatBoost) are best suited for embedded controllers and warranty tracking, sequence-based models (GRU, LSTM, CNN–LSTM) provide the most reliable performance for edge servers and cold-climate fleets, while CNN–LSTM stands out as the most appropriate choice for large-scale cloud analytics and high-load fleet monitoring.

Table 13. Practical recommendations for BMS deployment based on benchmarking results.

BMS Application	Recommended Models	Rationale
Embedded controllers (onboard EV BMS)	LightGBM, CatBoost	Microsecond inference; acceptable accuracy at nominal T; minimal computational load
Edge servers (charging stations, fleet depots)	GRU, LSTM, CNN–LSTM	Accuracy–latency balance (~0.1 ms/sample); suitable for real-time fleet monitoring
Cloud analytics (fleet-wide predictive maintenance)	CNN–LSTM	Best overall accuracy and generalization across heterogeneous cells; scalable to fleets
Warranty and lifecycle management	CatBoost, LightGBM	Stable calibration and interpretability; useful for long-horizon SOH tracking
Cold-climate fleets	GRU, LSTM	Reliable under low-temperature conditions; robust to variability
High-load fleets/stressed operation	CNN–LSTM, GRU	Strong predictive performance at elevated temperature; resilient in accelerated cycling

Additional analyses and extended plots that could not be included in the main text due to space constraints are provided in the Supplementary Information. These include extended per-battery error distributions, alternative visualization formats, and robustness checks across models and conditions. All supplementary results confirm and reinforce the trends reported in the main figures.

5. Conclusions

This study presented a comprehensive benchmarking of ML and DL models for SOH estimation using cross-validation results from thirty-two commercial lithium-ion cells. By integrating performance metrics across batteries, temperature regimes, and latency constraints, a unified perspective was developed on the suitability of different algorithms for BMS applications.

Sequence-based models such as CNN–LSTM, GRU, and LSTM consistently achieved the highest accuracy and robustness across heterogeneous cells. CNN–LSTM delivered the lowest mean RMSE ($\approx$0.006) with stable calibration, while GRU and LSTM demonstrated the highest robustness with median R^2 values of about 0.97 and 0.96, respectively. BiLSTM achieved comparable accuracy but suffered from higher latency and instability. Tree ensembles like CatBoost and LightGBM provided mid-tier accuracy but unmatched efficiency, offering inference in the microsecond range. In contrast, SVR and MLP underperformed and proved unsuitable for cross-cell SOH estimation.

Temperature-dependent performance revealed that GRU, LSTM, and CNN–LSTM achieved the lowest errors under both low- and high-temperature regimes, showing re-

silience across heterogeneous conditions. CatBoost and LightGBM performed competitively under nominal temperature conditions, but their accuracy degraded significantly at the extremes.

Deployment suitability was strongly influenced by trade-offs between accuracy and latency. For embedded controllers, such as onboard EV diagnostics, LightGBM and CatBoost are preferable due to their microsecond-level inference and acceptable nominal accuracy. For edge servers, such as charging stations and depots, GRU, LSTM, and CNN–LSTM are more suitable because they balance accuracy and latency at about 0.1 ms per sample. Cloud analytics applications, such as fleet-wide predictive maintenance, benefit most from CNN–LSTM, which offers the best overall generalization. For warranty management and lifecycle tracking, CatBoost and LightGBM stand out with stable and interpretable calibration suitable for long-term SOH monitoring.

Application-oriented insights further emphasize model selection. For cold-climate fleets, GRU and LSTM ensure reliable SOH estimation under low-temperature conditions. For nominal operating conditions, CatBoost and LightGBM provide efficient monitoring well-suited to passenger EVs and stationary systems. For high-load fleets, CNN–LSTM and GRU remain robust and suitable for fleet-level monitoring and edge analytics. For heterogeneous fleets, CNN–LSTM consistently ranks as the most generalizable model, making it the optimal choice for predictive maintenance across diverse battery populations.

The practical implications of these findings underline that no universal best estimator exists. Instead, model selection must balance accuracy, robustness, and latency depending on deployment needs. Sequence models dominate in accuracy and robustness, while boosted trees excel in latency-critical contexts. This creates a clear roadmap for BMS applications: LightGBM and CatBoost for embedded controllers, GRU, LSTM, and CNN–LSTM for edge systems, CNN–LSTM for cloud platforms, and CatBoost and LightGBM for warranty and lifecycle tracking.

The present analysis focused on accuracy, robustness, and latency. Future work should incorporate uncertainty quantification, calibration error metrics, and cycle-to-failure residual analysis. Additionally, hardware-in-the-loop experiments will be critical to validate deployment feasibility on automotive-grade microcontrollers.

In summary, this work provides a complete benchmarking framework for selecting ML and DL models for SOH prediction in LIBs, bridging statistical performance with BMS deployment requirements across embedded, edge, cloud, and fleet management contexts.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/batteries11110393/s1.

Author Contributions: Conceptualization, S.S.M.; Methodology, S.S.M.; Validation, S.S.M.; Investigation, S.S.M.; Resources, M.H., L.B., A.L.-B. and F.A.; Writing—original draft, S.S.M.; Writing—review & editing, M.H. and F.A.; Supervision, M.H., L.B., A.L.-B. and F.A.; Project administration, F.A.; Funding acquisition, M.H., L.B. and F.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

Appendix A

This appendix provides Supplementary Material that supports and complements the main text. Appendix A establishes a structured reference system: Table A1 outlines

the methodological workflow, Table A2 details the full set of mathematical formulations, and Table A3 defines all symbols and variables. Appendix B presents extended experimental groups (A–H), including detailed test conditions and degradation profiles under different temperature and current regimes, to broaden the validation of the proposed SOH estimation framework.

Methodological Framework—Steps, Formulations, and Nomenclature

Table A1. Summary of methodological steps with corresponding formula references.

Step	Description	Formula
1. Feature construction	Each cycle k is represented by a feature vector including SOH, capacity, temperature, and current.	(1)
2. Prediction target	The output variable is defined as the SOH at cycle k.	(2)
3. Temporal sequences	Input samples are formed as sequences of the past L feature vectors, predicting SOH for the next cycle.	(3)
4. Flattened representation	Sequences are reshaped into tabular vectors for use in classical machine learning models.	(4)
5. Cross-validation groups	Data is partitioned into K contiguous folds using a blocked assignment strategy.	(5)
6. Feature normalization	The mean and variance of each feature are computed to standardize inputs.	(6)–(7)
7. Learning objective	Models are trained by minimizing the mean squared error (MSE).	(8)
8. Outlier filtering	Residuals are filtered using Tukey's rule to exclude extreme errors.	(9)
9. Root mean squared error	RMSE is computed over the filtered set of residuals.	(10)
10. Mean absolute error	MAE is computed as the mean absolute residual across valid samples.	(11)
11. Coefficient of determination	Goodness of fit is measured by the R^2 statistic.	(12)
12. Latency	Prediction speed is evaluated as average inference time per test sample (ms).	(13)
13. Model complexity	Complexity is quantified by the number of trainable parameters.	(14)
14. Cross-validation mean RMSE	Average RMSE across all folds is reported.	(15)
15. RMSE variability	Standard deviation of fold-wise RMSE values is calculated.	(16)
16. Confidence intervals	A 95% confidence interval of RMSE is computed	(17)
17. Temperature binning	Data is divided into operating ranges: low ($T \leq 14$ °C), nominal ($14 < T \leq 33$ °C), and elevated ($T > 33$ °C).	(18)
18. Temperature-wise RMSE	RMSE is calculated separately within each temperature bin.	(19)
19. Prediction variability	Standard deviation of predictions is computed per fold and averaged.	(20)–(22)
20. Feature statistics	Feature distributions are characterized by mean and variance.	(23)–(24)
21. Residual density	Residual distributions are approximated using kernel density estimation.	(25)
22. QQ fitting	Normality of residuals is assessed by fitting theoretical quantiles.	(26)

Table A1. *Cont.*

Step	Description	Formula
23. Calibration bins	Predictions are grouped into bins; true and predicted means are compared.	(27)–(28)
24. Calibration errors	Expected calibration error (ECE) and maximum calibration error (MCE) are reported.	(29)
25. Rolling RMSE	Temporal robustness is analyzed using rolling-window RMSE.	(30)
26. End-of-life detection	True and predicted EOL cycles are identified when SOH ≤ 0.8.	(31)–(32)
27. EOL error	EOL prediction accuracy is measured as the difference between predicted and actual EOL cycles.	(33)
28. SHAP feature importance	Feature contributions are derived from averaged absolute SHAP values.	(34)
29. Permutation importance	Feature relevance is quantified by the change in model error after random permutation.	(35)
30. Final evaluation suite	The final comparison includes all metrics: RMSE, CI, MAE, R^2, latency, parameter count, variability, calibration, temperature RMSE, EOL error, and importance scores.	(36)

Table A2. List of mathematical formulations corresponding to the methodology described in Table A1.

Formula	Eq. No.
$Z_k = [SOH_k\ C_k\ T_k\ I_k]^\intercal$	(A1)
$y_k = SOH_k$	(A2)
$X_t = [Z_{t-L+1}, \ldots, Z_t],\ y_{t+1} = SOH_{t+1}$	(A3)
$x_t = vec(X_t)$	(A4)
$g_t = \left\lfloor \frac{Kt}{N-L+1} \right\rfloor + 1$	(A5)
$\mu_j = \frac{1}{\lvert\mathcal{T}\rvert L} \sum_{t \epsilon \mathcal{T}} \sum_{\updownarrow=1}^{L} X_t[\updownarrow, j]$	(A6)
$\widetilde{X}_t[\updownarrow, j] = \frac{X_t[\updownarrow, j] - \mu_j}{\sigma_j + \varepsilon}$	(A7)
$\mathcal{L}_{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$	(A8)
$L = \{i : Q_1 - KIQR \leq r_i \leq Q_3 + KIQR\}$	(A9)
$RMSE = \sqrt{\frac{1}{\lvert L\rvert} \sum_{i \in L} (\hat{y}_i - y_i)^2}$	(A10)
$MAE = \frac{1}{\lvert L\rvert} \sum_{i \in L} \lvert\hat{y}_i - y_i\rvert$	(A11)
$R^2 = 1 - \frac{\sum_{i \in L} (\hat{y}_i - y_i)^2}{\sum_{i \in L} (y_i - \bar{y}_i)^2}$	(A12)
$latency_{ms} = 10^3 . \frac{\Delta t}{n_{test}}$	(A13)
$\#\theta = \sum_{p \in weights} \prod_{d \in shape(p)} d$	(A14)
$\overline{RMSE} = \frac{1}{K} \sum_{k=1}^{K} RMSE_k$	(A15)
$s = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (RMSE_k - \overline{RMSE})^2}$	(A16)
$\overline{RMSE} \pm t_{0.975,v} \frac{s}{\sqrt{K}}$	(A17)
$Low\ T \leq 14,\ Nominal : 14 < T \leq 33,\ Elevated : T > 33$	(A18)
$RMSE_b = \sqrt{\frac{1}{\lvert L_b\rvert} \sum_{i \in L_b} (\hat{y}_i - y_i)^2}$	(A19)

Table A2. *Cont.*

Formula	Eq. No.
$s^{(k)} = \sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\hat{y}_i - \overline{\hat{y}}\right)^2}$	(A20)
$\bar{s} = \frac{1}{K}\sum_{k=1}^{K} s^{(k)}$	(A21)
$\mu_j = \frac{1}{M}\sum_T p_{rj}$	(A22)
$\sigma_j = \sqrt{\frac{1}{M}\sum_T \left(p_{rj} - \mu_j\right)^2}$	(A23)
$\hat{f}(x) \approx \frac{1}{Bh}\sum_{b=1}^{B} 1\{x \in bin_b\}$	(A24)
$min_{\beta_0,\beta_1}\sum_i \left(r_{(i)} - (\beta_0 + \beta_1 q_i)\right)^2$	(A25)
$\mu_{true}^{(j)} = \frac{1}{n_j}\sum_{i\in B_j} y_i$	(A26)
$\mu_{pred}^{(j)} = \frac{1}{n_j}\sum_{i\in B_j} \hat{y}_i$	(A27)
$ECE = \sum_j w_j \left\lvert \mu_{pred}^{(j)} - \mu_{true}^{(j)} \right\rvert$	(A28)
$MCE = max_j \left\lvert \mu_{pred}^{(j)} - \mu_{true}^{(j)} \right\rvert$	(A29)
$RMSE_{roll}(i) = \sqrt{\frac{1}{\lvert w_i \rvert}\sum_{j\in w_i} e_j^2}$	(A30)
$EOL_{true} = min\{cyc_i : y_i \le 0.8\}$	(A31)
$EOL_{pred} = min\{cyc_i : \hat{y}_i \le 0.8\}$	(A32)
$\Delta_{EOL} = EOL_{pred} - EOL_{true}$	(A33)
$Imp_j = \frac{1}{n}\sum_i \lvert \varnothing_{i,j} \rvert$	(A34)
$Imp_j = \mathbb{E}\left[M\left(f; X^{\pi(j)}\right) - M(f; X)\right]$	(A35)
$\left\{\overline{RMSE}, CI, \overline{MAE}, \overline{R^2}, latency, \#\theta, \bar{s}, ECE, MCE, RMSE_{Low/Nom/Elev}, \Delta_{EOL}\right\}$	(A36)

Table A3. Nomenclature of all variables and symbols used in the methodology and formulas.

Definition	Symbol
Indices denoting cycle number (k), time step (t), sample index (i), and feature dimension (j).	k, t, i, j
State of health at cycle kk, representing the ratio of available capacity to nominal capacity.	SOH_k
Discharge capacity of the battery at cycle k.	C_k
Measured temperature at cycle k.	T_k
Applied current at cycle k.	I_k
Cycle count associated with measurement k.	cyc_k
Feature vector at cycle k, consisting of SOH, capacity, temperature, and current.	z_k
Prediction target, defined as the SOH at cycle k.	y_k
Input sequence of feature vectors over the past L cycles ending at time t.	X_t
Flattened tabular representation of the sequence Xt.	x_t
Look-back window length, i.e., the number of previous cycles considered as input.	L
Number of folds used in cross-validation.	K
Group index assigning samples to cross-validation folds.	g_t
Training set used for model fitting.	$\mathcal{T}$

Table A3. *Cont.*

Definition	Symbol
Mean value of feature j computed over the training set.	μ_j
Standard deviation of feature j computed over the training set.	σ_j
Normalized sequence obtained after feature standardization.	$\widetilde{X}_t$
Predicted SOH value for sample ii.	$\hat{y}_i$
Residual error for sample ii, defined as the difference between prediction and true SOH.	r_i
First and third quartiles of the residual distribution.	Q_1, Q_3
Interquartile range, calculated as Q3−Q1.	IQR
Set of residuals retained after outlier filtering.	L
Total inference runtime during evaluation.	Δt
Number of test samples used for evaluation.	n_{test}
Total number of trainable parameters in the model.	#θ
Root mean square error for fold k.	$RMSE_k$
Mean RMSE computed across all folds.	$\overline{RMSE}$
Standard deviation of fold-wise RMSE values.	s
Critical value of Student's t-distribution at 95% confidence with v degrees of freedom.	$t_{0.975,v}$
Operating temperature of the battery.	T
Subset of residuals corresponding to temperature bin b.	L_b
Prediction standard deviation within fold k.	$s^{(k)}$
Mean prediction standard deviation across all folds.	$\bar{s}$
Value of feature j in raw input sample r.	p_{rj}
Total number of samples in the population or batch used for statistics.	M
Kernel density estimate of the residual distribution.	f(x)
Theoretical quantile corresponding to residual ri.	q_i
Regression coefficients in quantile–quantile fitting.	β_0, β_1
Calibration bin j containing a subset of samples.	B_j
Empirical mean of true SOH values within bin j.	$\mu_{true}^{(j)}$
Empirical mean of predicted SOH values within bin j.	$\mu_{pred}^{(j)}$
Weight of calibration bin j, proportional to its sample count.	w_j
Rolling window of samples centered at index ii.	W_i
Residual error of sample j.	e_j
End-of-life cycle, defined as the first cycle where SOH $\leq$ 0.8.	EOL
Difference between predicted and actual end-of-life cycles.	Δ_{EOL}
SHAP value representing the contribution of feature j to the prediction for sample i.	$\varnothing_{i,j}$
Input dataset with the *j*-th feature permuted for importance analysis.	$X^{\pi(j)}$
Model performance metric (e.g., RMSE) evaluated on dataset X.	$M(f; X)$

Appendix B. Extended Experimental Groups

The following appendix contains detailed experimental procedures and degradation profiles for Groups B to F, which complement the primary dataset discussed in Section 2.1. These groups were studied under various temperature and current conditions to support broader model validation for battery aging and SOH estimation.

Appendix B.1. Group A

Four LIBs with IDs 5, 6, 7, and 18 were considered in group A, which underwent repeated charge–discharge cycling for the purpose of studying aging effects and predictive modeling development. The batteries were charged with 1.5 A constant current until they reached 4.2 V terminal voltage and maintained this voltage until the current dropped below 20 mA. The batteries were discharged under controlled current conditions at 2 A until voltage reached 2.7 V, 2.5 V, 2.2 V and then finished at 2.5 V. The experimental tests were carried out until the batteries reached the end of life, when they experienced a decrease in capacity to 1.4 Ah, starting from 2 Ah. In-depth cycle-level measurements of terminal voltage, along with output current measurements, temperature readings, and time registrations for capacity and both charging and discharging stages, were obtained. The established dataset provides an excellent foundation for creating and testing advanced SOC, SOH calculators, and RUL forecasting models. Figure A1 presents the cycle-level profiles of these LIBs (5, 6, 7, and 18) under the specified charge–discharge conditions.

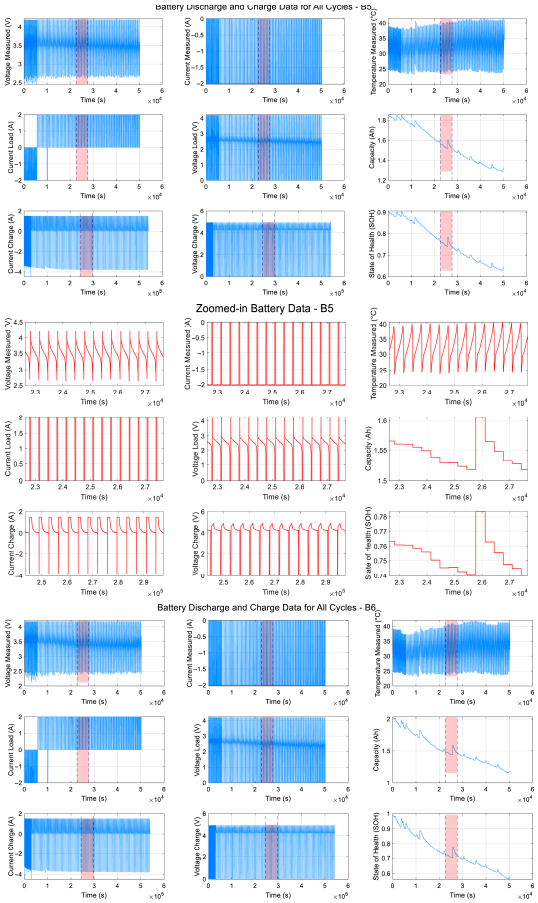


Figure A1. *Cont.*

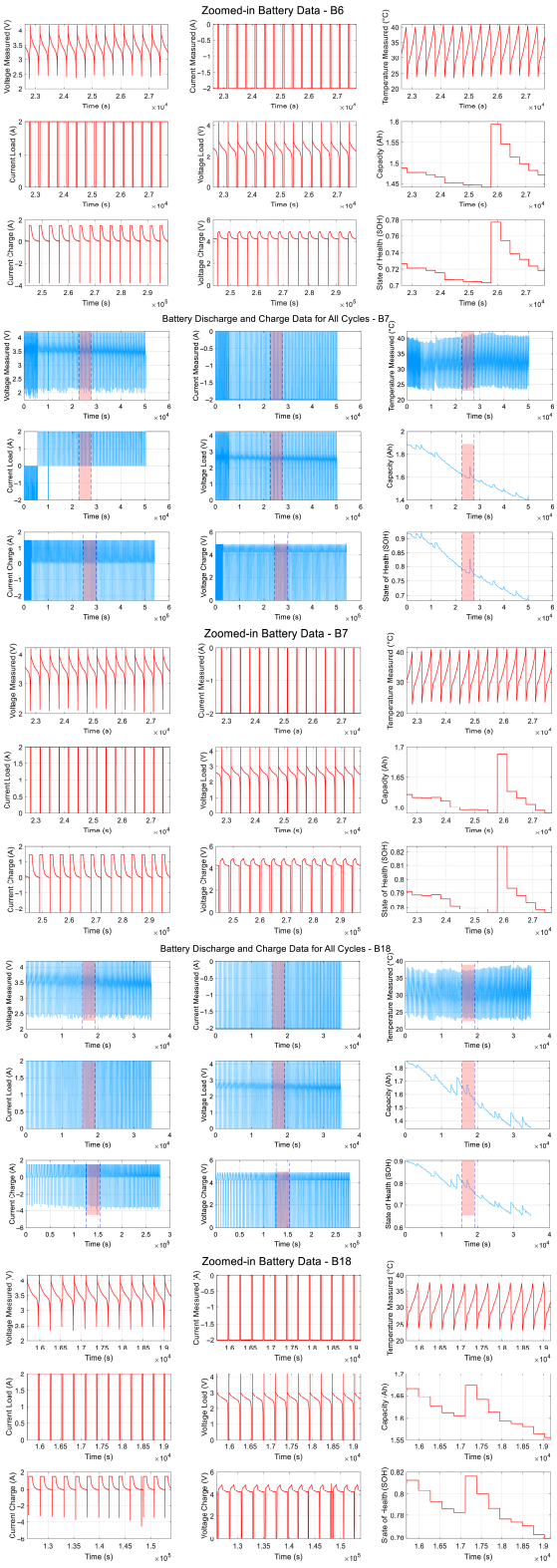


Figure A1. Cycle-level degradation profiles of LIBs: B5, B6, B7, and B18 under charge–discharge conditions. Figures B5, B6, B7, and B18 display a 3 × 3 grid of plots that track various battery performance metrics over time across all cycles. The top row includes voltage measurements (V), current measurements (A), and temperature measurements (°C), while the middle row shows current load (A), voltage measurements (V), and capacity (Ah). The bottom row presents the current charge (A) and the voltage charge (V) over time. These plots offer a comprehensive view of the battery's performance, illustrating the evolution of key parameters during both charging and discharging operations, highlighting the degradation behavior of the batteries across multiple cycles.

Appendix B.2. Group B

The study of LIB aging characteristics for IDs 25, 26, 27, and 28 occurred through repeated charge–discharge cycles at 24 °C in Group B. Each charging process began with a 1.5 A constant current phase that stopped when reaching 4.2 V terminal voltage, followed by a constant voltage phase until the current lowered past 20 mA. Pulses generated with 0.05 Hz frequency during discharging served to terminate the cells at individual voltage values of 2.0 V, 2.2 V, 2.5 V and 2.7 V. The test continued until the batteries faded to the end of life, at which point they displayed 30% less nominal capacity from 2 Ah to 1.4 Ah. High-resolution measurement tracked terminal voltage, current and temperature as well as time and capacity throughout charging and discharging sequences during each cycle. Figure A2 illustrates the cycle-level degradation profiles of LIBs (25, 26, 27, and 28) under charge–discharge conditions.

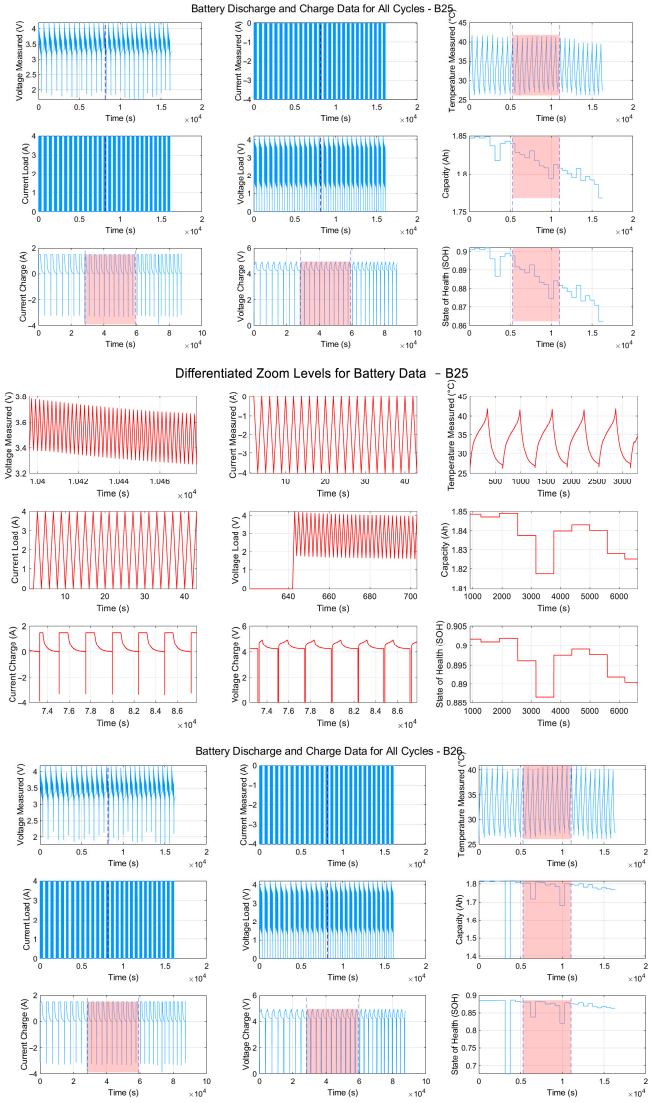


Figure A2. *Cont.*

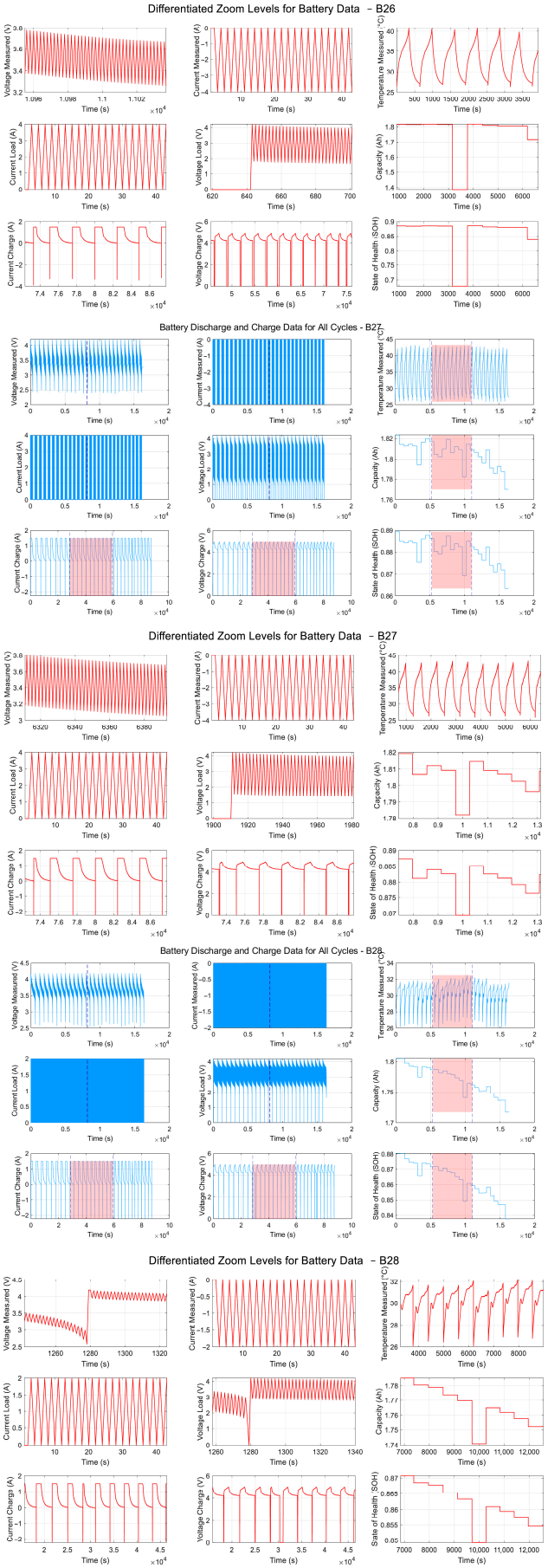


Figure A2. Cycle-Level profiles of LIBs (25, 26, 27, and 28) under charge–discharge conditions.

Appendix B.3. Group C

The aging characteristics of LIBs with ID numbers 29, 30, 31, and 32 were evaluated through repeated charge–discharge cycles carried out at 43 °C temperatures in Group C. The LIBs endured a 1.5 A constant current charge until achieving 4.2 V, while the charge phase continued until the current level descended below 20 mA. A constant current discharged the batteries while the cutoff voltages reached 2.0 V, 2.2 V, 2.5 V, and 2.7 V. The testing process continued until the batteries experienced end-of-life when their capacity reduction reached 30% from 2 Ah to 1.4 Ah. The experimental data acquisition included detailed measurement of terminal voltage, along with current measurements and temperature readings and capacity data from charging and discharging cycles. The gathered data enables us to understand battery deterioration under high temperatures for developing precise SOH estimation and RUL forecasting models. Figure A3 presents the cycle-level profiles of LIBs (29, 30, 31, and 32) under charge–discharge conditions.

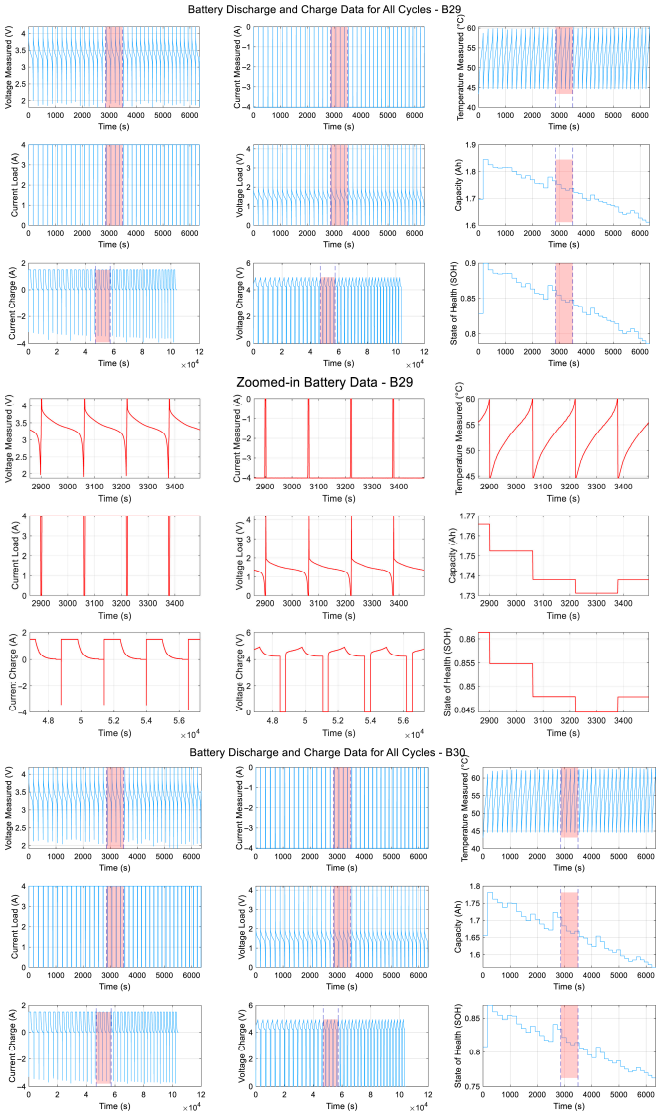


Figure A3. *Cont*.

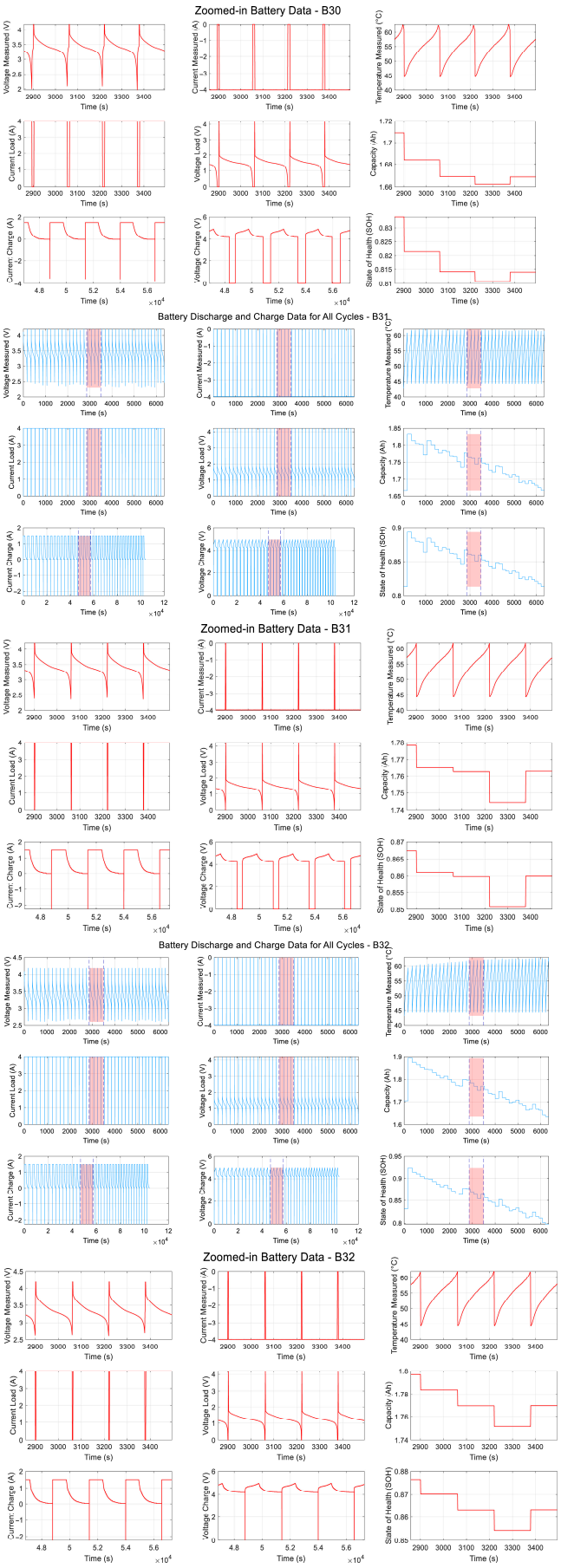


Figure A3. Cycle-Level profiles of LIBs (29, 30, 31, and 32) under charge–discharge conditions.

Appendix B.4. Group D

Three LIBs (IDs: 33, 34, and 36) were studied using repeated charge–discharge operations at 24 °C temperature within Group D. Throughout charging, the process maintained a standard protocol with 1.5 A constant current until it reached 4.2 V, then used a constant voltage phase until the current reached 20 mA. The discharging process of batteries 33 and 34 utilized a 4 A constant current until reaching 2.0 V and 2.2 V cut-off voltage, respectively, but battery 36 received a 2 A constant current discharge procedure, which stopped at 2.7 V cut-off. The experiment lasted until the cells reduced their capacity to 20% from 2 Ah to 1.6 Ah. The testing period produced precise measurements of voltage, together with current and temperature, until reaching capacity levels. These measurements included time and high-fidelity data. The data collection serves primary purposes for developing SOH estimation methods under different discharge load conditions. Figure A4 displays the cycle-level profiles of LIBs (33, 34, and 36) under charge–discharge conditions.

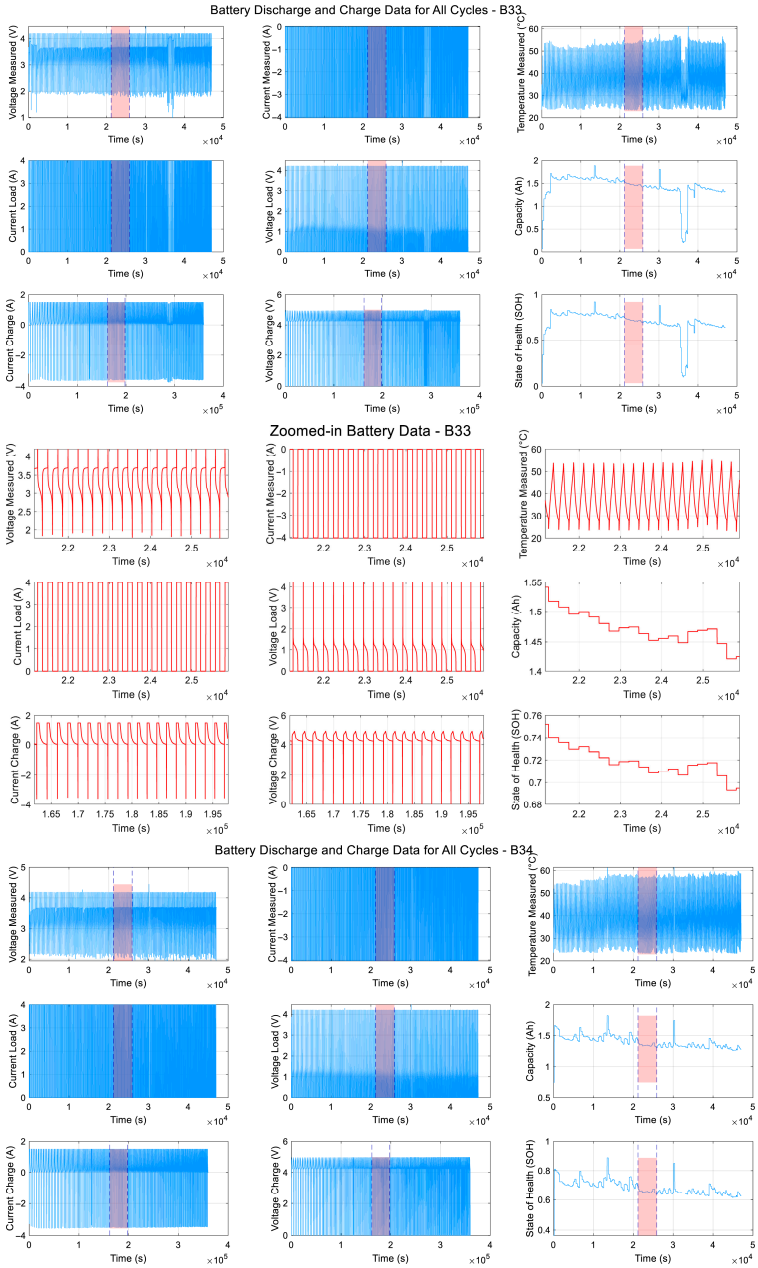


Figure A4. *Cont.*

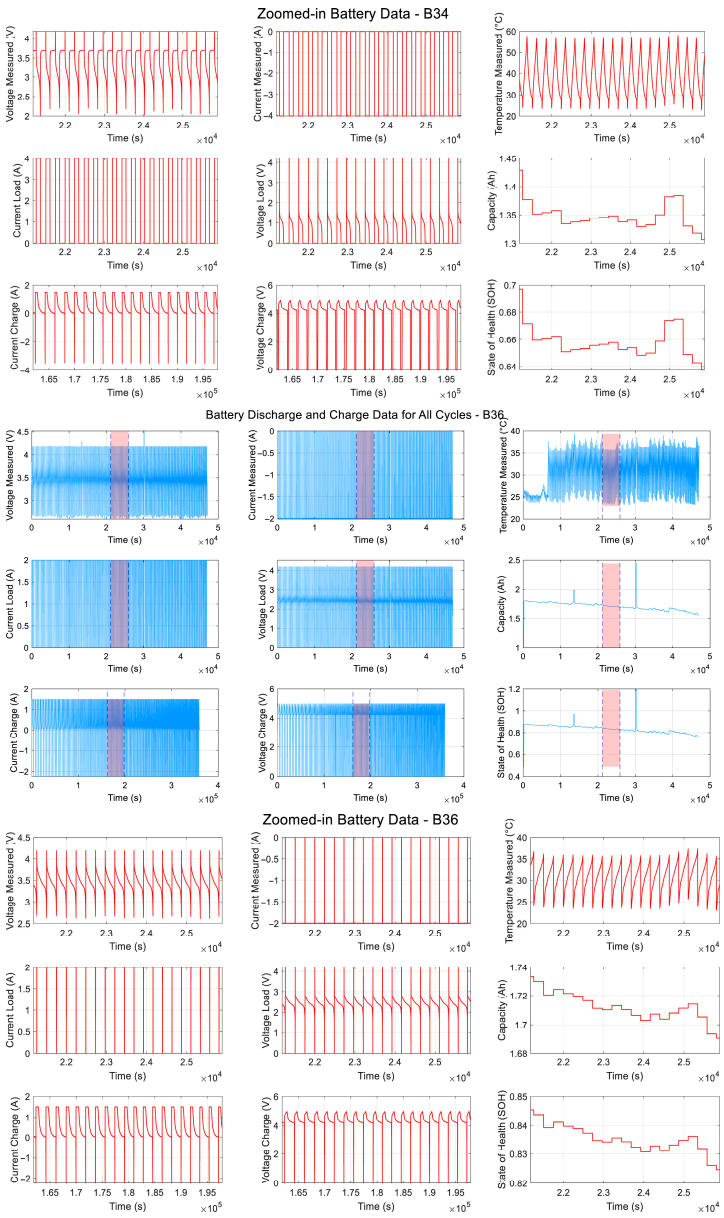


Figure A4. Cycle-Level profiles of LIBs (33, 34, and 36) under charge–discharge conditions.

Appendix B.5. Group E

The repeated charge–discharge cycling tests were performed on four LIBs with IDs 41 to 44 at a temperature of 4 °C through Group E. The battery charging proceeded with 1.5 A constant current up to 4.2 V until the current reached 20 mA. A fixed load current of 4 A and 1 A powered the discharge process for each battery while the tool cut off at specific voltages of 2.0 V, 2.2 V, 2.5 V, and 2.7 V. The tests were carried out multiple times until the nominal capacity declined by 30% while the capacity degraded from 2 Ah to 1.4 Ah. Some of the discharge cycles showed abnormally low-capacity values, while the majority experienced normal degradation patterns, but the specific causes remain unknown. The extensive quantitative dataset, which contains precise measurements of voltage, current, temperature, time, and capacity measurements across all cycles, functions as a vital tool for understanding battery degradation while predicting SOH and RUL under cold conditions. Figure A5 shows the cycle-level profiles of LIBs (labeled 41 to 44) under charge–discharge conditions.

Appendix B.6. Group F

The present study evaluated four LIBs (IDs: 45 to 48) at a temperature of 4 °C within Group F to determine their degradation pattern under such conditions. The batteries underwent 1.5 A constant current charging until achieving a 4.2 V terminal voltage and then operated at constant voltage until reaching 20 mA current point. The discharge procedure utilized a constant current of 1 A while batteries 45 through 48 reached termination voltages of 2.0 V, 2.2 V, 2.5 V, 2.7 V, respectively. The battery cells underwent continuous cycles until their capacity reached 1.4 Ah after starting from 2 Ah. The degradation patterns in most cycles remained consistent, but some discharge cycles showed unexpectedly low-capacity readings that scientists have yet to identify the root causes. Each cycling operation captured extensive time-series data about voltage, current, temperature, time and capacity, which led to valuable information to develop precise SOH assessment and RUL prediction models when operating at low temperatures. Figure A6 illustrates the cycle-level profiles of LIBs (labeled 45 to 48) under charge–discharge conditions.

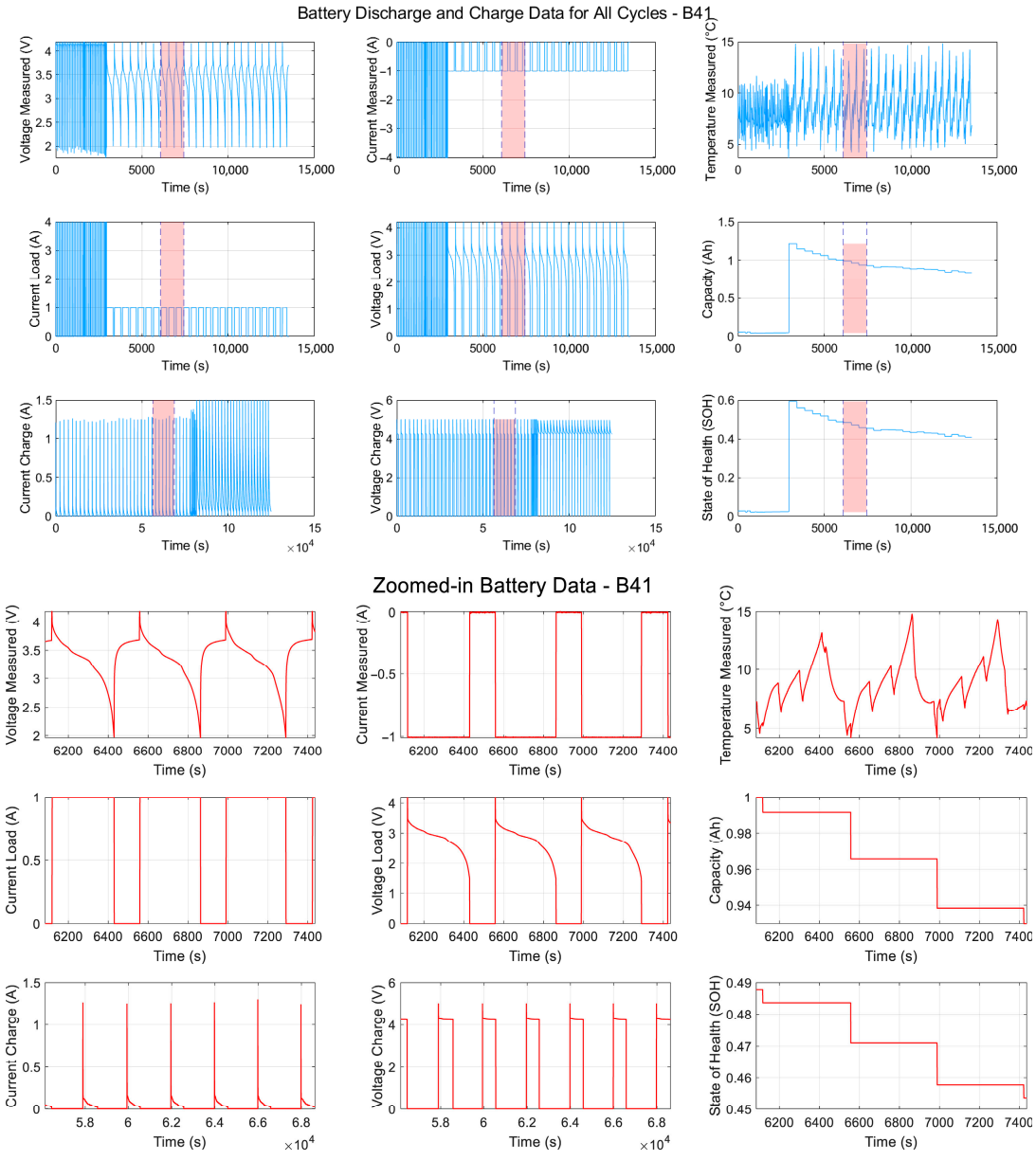


Figure A5. *Cont.*

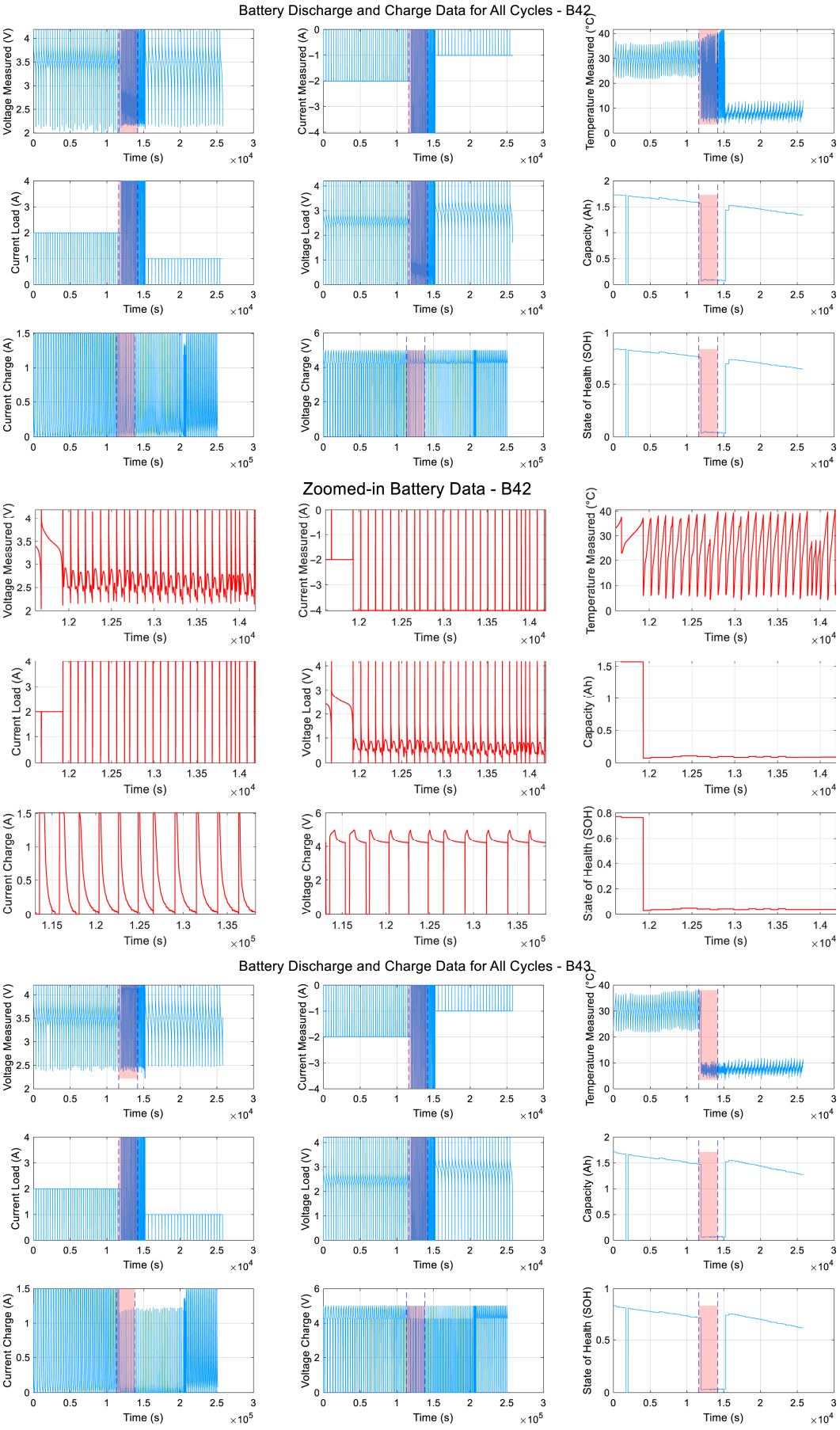


Figure A5. *Cont.*

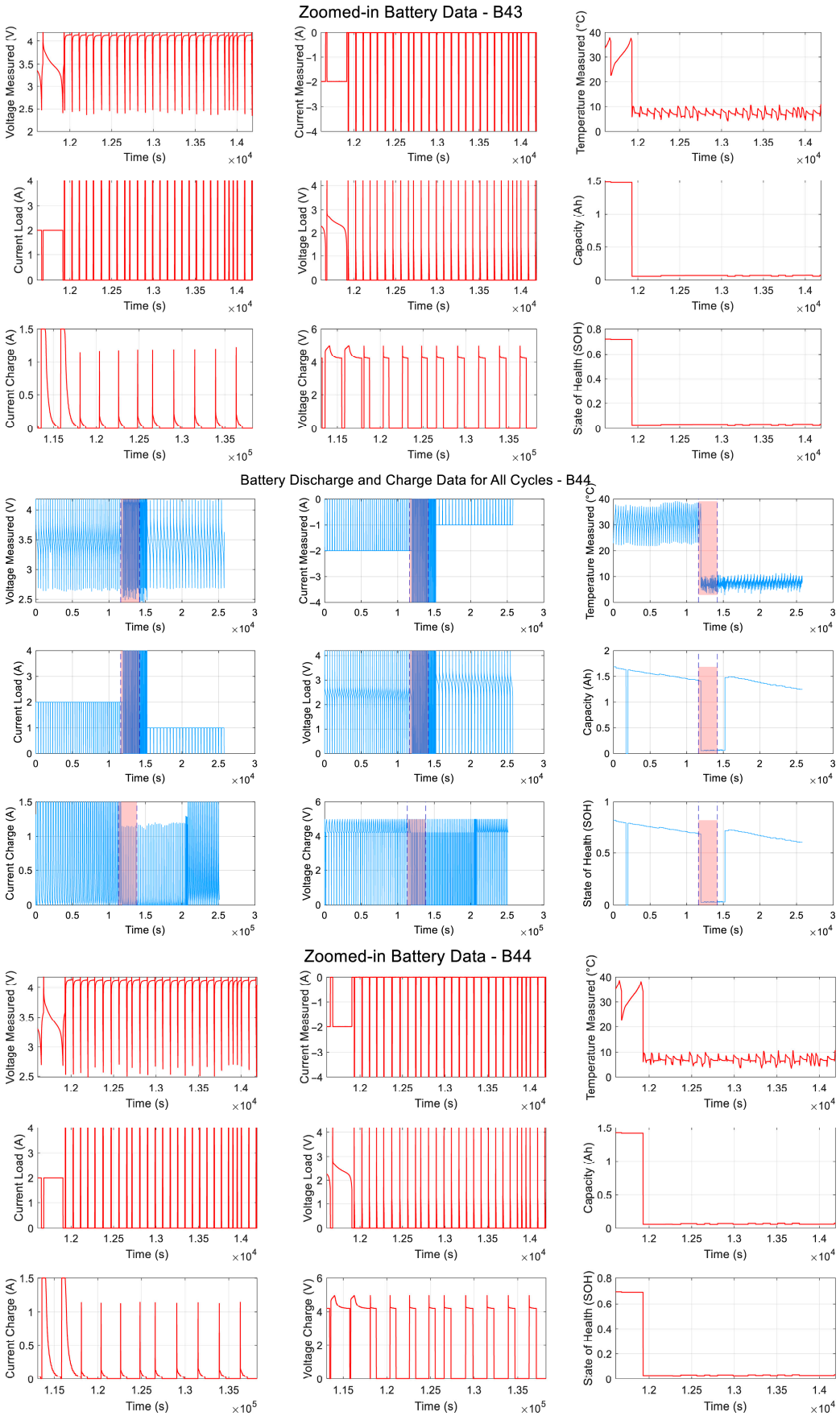


Figure A5. Cycle-Level profiles of LIBs (labeled 41 to 44) under charge–discharge conditions.

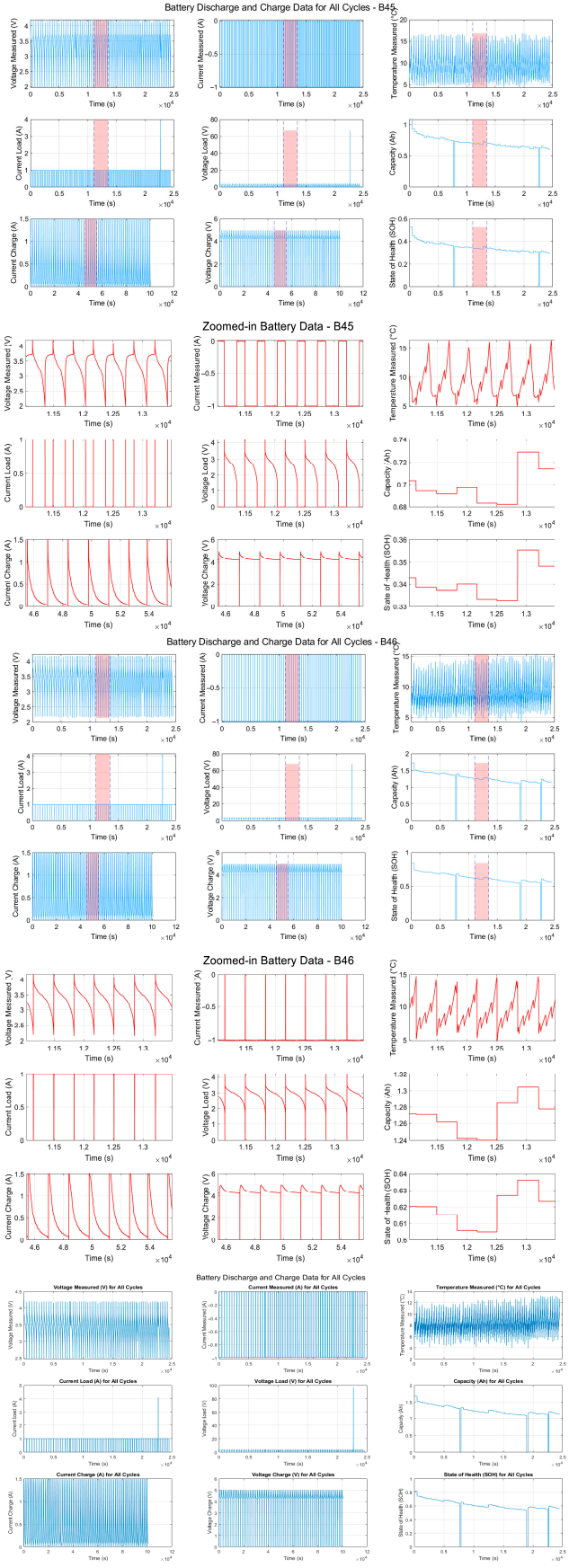


Figure A6. *Cont.*

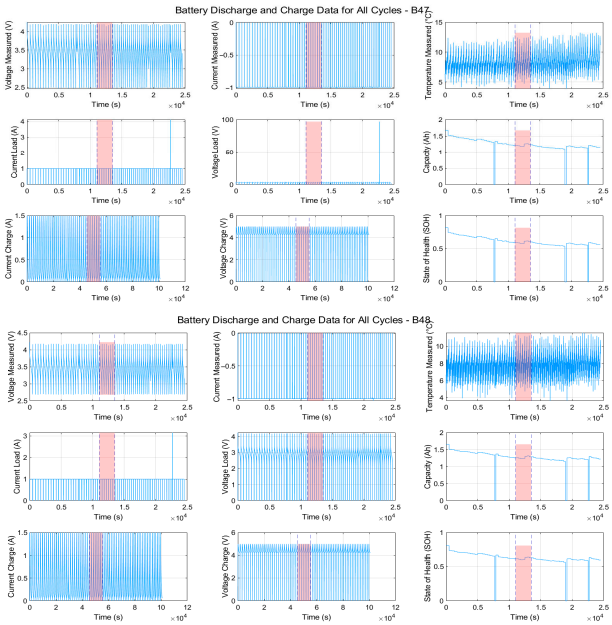


Figure A6. Cycle-Level profiles of LIBs (labeled 45 to 48) under charge–discharge conditions.

Appendix B.7. Group G

The aging process of four LIBs with IDs 49 to 52 was monitored through multiple charge–discharge cycles, which took place at 4 °C. A constant 1.5 A charge was implemented until it reached 4.2 V terminal voltage, then the procedure switched to constant voltage until the current dropped to 20 mA. The discharge process of batteries 49 through 52 was carried out under a 2 A current level with termination points of 2.0 V, 2.2 V, 2.5 V and 2.7 V, respectively. Under the experimental condition, the study had to be stopped early due to a program fault, but researchers successfully acquired data sets for all measurements at each cycle, including voltage, current, temperature, time and cumulative capacity. The research team needs to investigate the factors behind the discharge cycles that showed irregular voltage and capacity measurements. These anomalies, together with the terminated lifecycle, still make the dataset appropriate for developing SOH estimation and RUL prediction models in cold environments. Figure A7 presents the cycle-level profiles of LIBs (labeled 49 to 52) under charge–discharge conditions.

Appendix B.8. Group H

The degradation pattern of LIBs (IDs: 53 to 56) was studied under low temperature at 4 °C for repeated charge–discharge cycling in Group H. A 1.5 A constant current charge was used to elevate the battery voltage to 4.2 V, where it was maintained through a constant voltage phase until the current dropped to 20 mA. Discharge process occurred at 2 A throughout the test battery 53 through 56 terminated operations when their voltages reached 2.0 V, 2.2 V, 2.5 V, and 2.7 V. Battery cycling proceeded until all batteries met their end-of-life criteria, which required their capacity to diminish to 1.4 Ah from an initial 2 Ah. Several discharges featured abnormally low-capacity values, for which researchers were unable to identify the exact causes. Regarding cold-temperature applications, the gathered dataset provides adequate information for developing predictive models to forecast RUL and monitor SOH through its collection of key measurements, including voltage, current, temperature, time, and capacity. Figure A8 displays the cycle-level profiles of LIBs (labeled 53 to 56) under charge–discharge conditions.

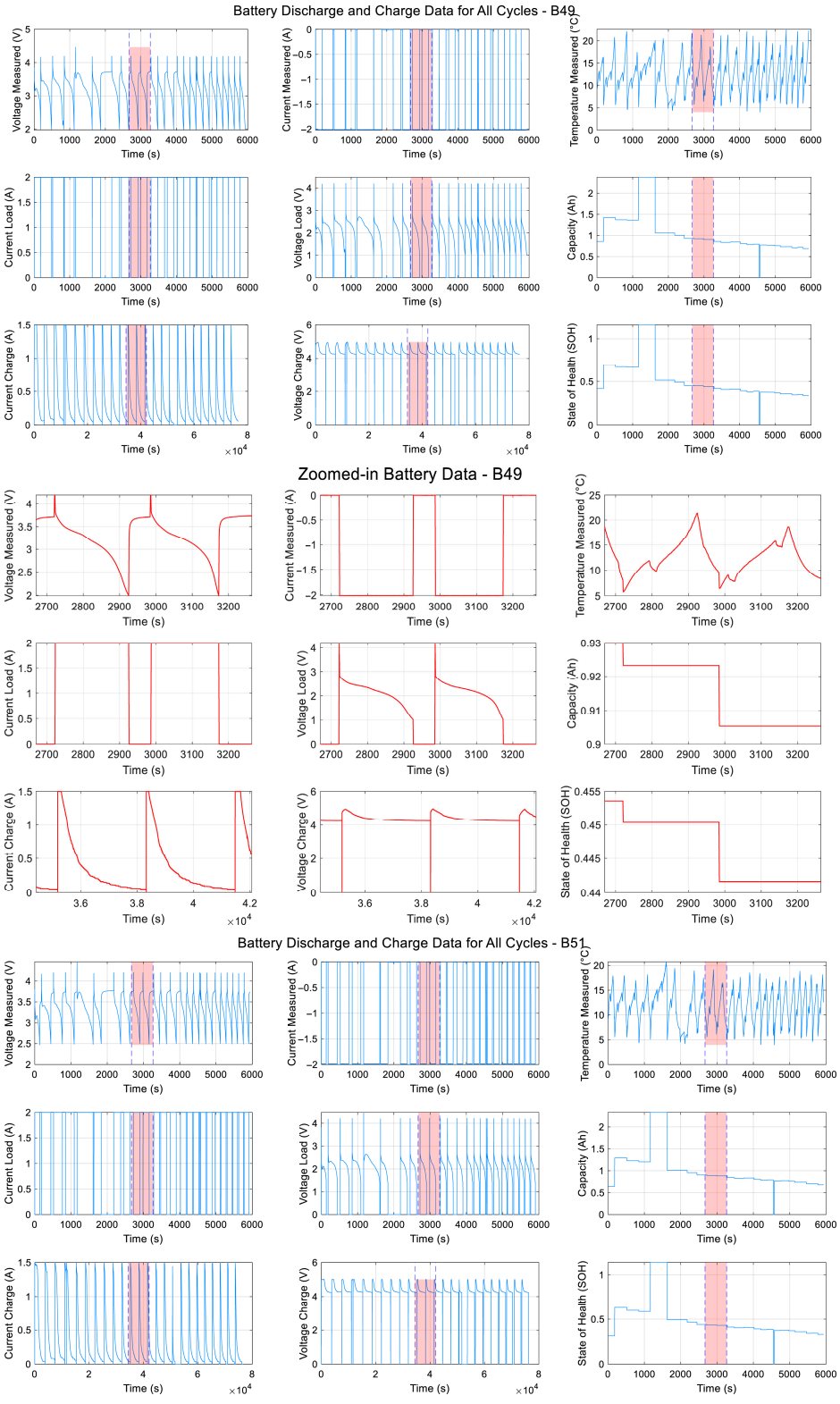


Figure A7. *Cont.*

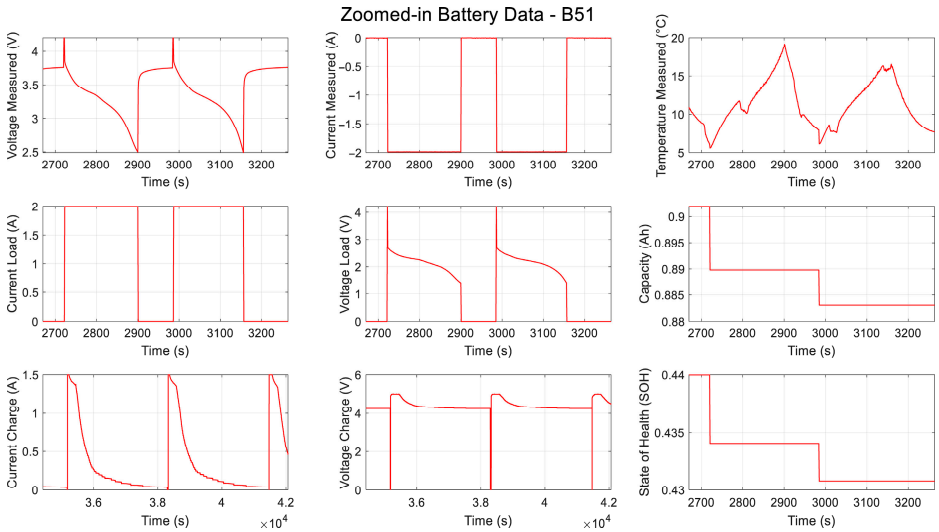


Figure A7. Cycle-Level profiles of LIBs (labeled 49 to 52) under charge–discharge conditions.

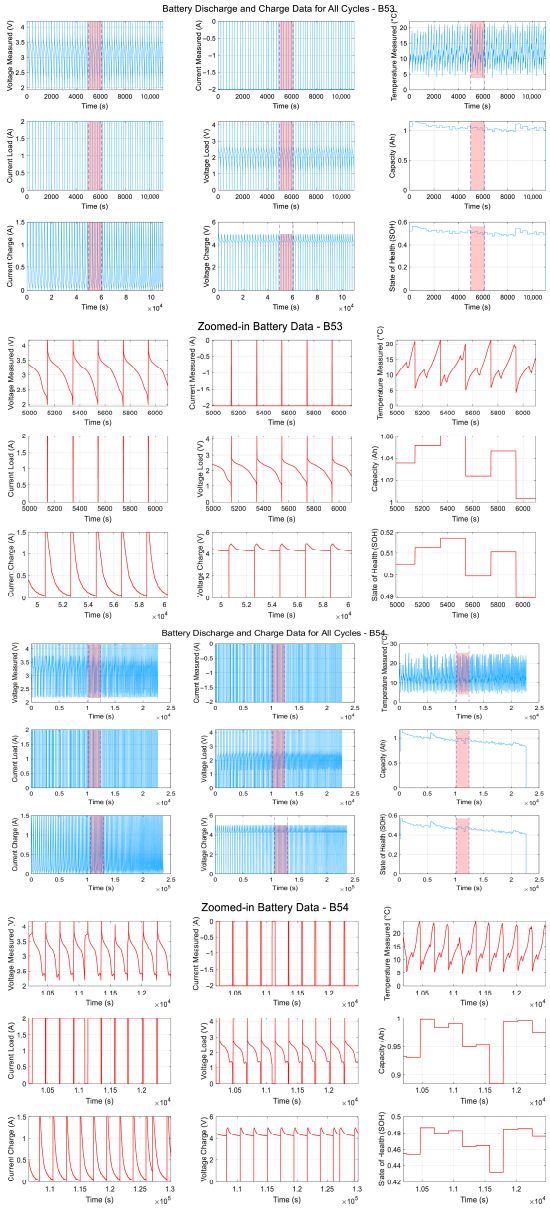


Figure A8. *Cont*.

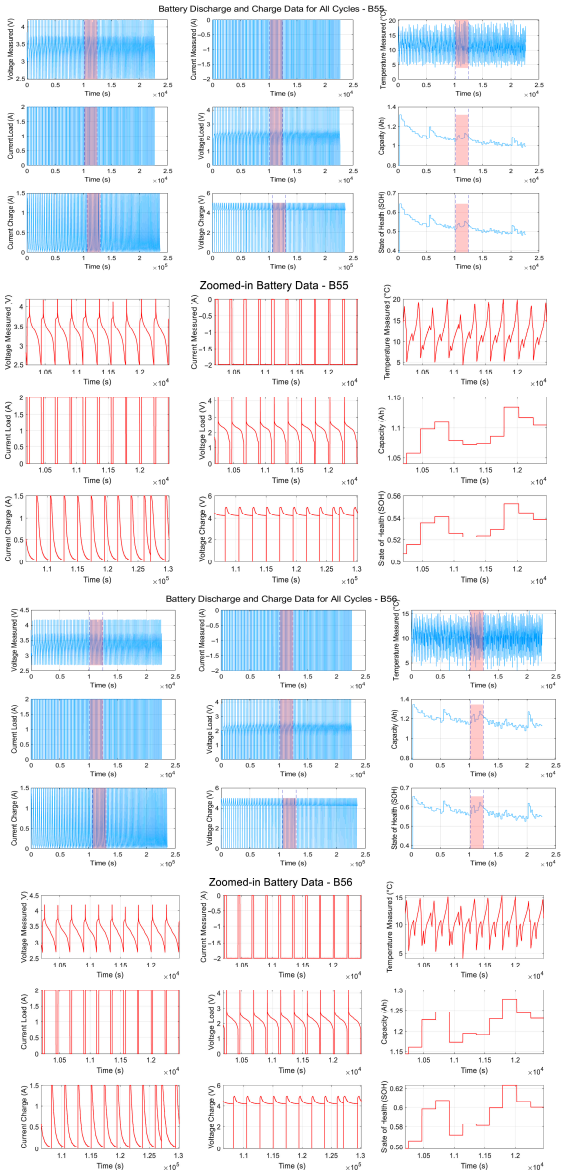


Figure A8. Cycle-Level profiles of LIBs (labeled 53 to 56) under charge–discharge conditions.

References

1. Madani, S.S. A Comprehensive Review on Lithium-Ion Battery Lifetime and Aging. *Batteries* **2025**, *11*, 127. [CrossRef]
2. Shen, W.; Cao, H.; Yuan, C. Heat Generation and Degradation Mechanism of Lithium-Ion Batteries. *ACS Omega* **2022**, *7*, 46847–46856. [CrossRef] [PubMed]
3. Rahman, T. Exploring Lithium-Ion Battery Degradation: A Concise Review. *Batteries* **2024**, *10*, 220. [CrossRef]
4. Lei, X.; Ouyang, Q.; Zhang, L.; Gao, Y. A Review of Li-Ion Battery State-of-Health and Remaining Useful Life Estimation (2010–2023). *eTransportation* **2024**, *20*, 100333. [CrossRef]
5. Zhang, M.; Liu, Y.; Sun, H. Electrochemical Impedance Spectroscopy: A New Chapter for Lithium-Ion Batteries. *Energies* **2023**, *16*, 1599. [CrossRef]
6. Liu, Y.; Li, H.; Xu, Y. State-of-Health Estimation of Lithium-Ion Batteries Based on Electrochemical Impedance Spectroscopy: A Review. *Prot. Control Mod. Power Syst.* **2023**, *8*, 34. [CrossRef]
7. Ma, C.; Wang, X. Review of Parameter Identification and State-of-Power Estimation for Lithium-Ion Batteries. *Processes* **2024**, *12*, 2166. [CrossRef]
8. Cheng, Y.S. Identification of parameters for equivalent circuit model of Li-ion battery cell with population based optimization algorithms. *Ain Shams Eng. J.* **2024**, *15*, 102481. [CrossRef]
9. Ho, K.C.; Chan, W.H. Deep Learning Approaches for Equivalent Circuit Model Parameter Identification. *Electronics* **2025**, *14*, 2201. [CrossRef]

10. Faraji Niri, M.; Chen, Y. Explainable Machine Learning for Li-Ion Batteries: From Production to State Estimation. *Energies* **2023**, *16*, 6360. [CrossRef]
11. Severson, K.A.; Attia, P.M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M.H.; Aykol, M.; Herring, P.K.; Fraggedakis, D.; et al. Data-Driven Prediction of Battery Cycle Life Before Capacity Degradation. *Nat. Energy* **2019**, *4*, 383–391. [CrossRef]
12. Xu, H.; Zhao, Y.; Zhou, X. Improved CNN–LSTM for State-of-Health Estimation of Lithium-Ion Batteries. *Energy* **2023**, *276*, 127998. [CrossRef]
13. Lu, J.; Zhang, Y.; Wang, L. Deep Learning State-of-Health Estimation without Target Labels Using Domain Adaptation. *Nat. Commun.* **2023**, *14*, 2912. [CrossRef]
14. Li, Z.; He, C.; Dong, L. BiLSTM–Transformer Hybrid Model for SOH Estimation under Fast Charging. *Energy* **2024**, *283*, 129196. [CrossRef]
15. Cai, X.; Liu, J. Transformer–LSTM Fusion for Robust SOH Estimation. *Appl. Sci.* **2025**, *15*, 3747. [CrossRef]
16. He, Y.; Deng, Z.; Chen, J.; Li, W.; Zhou, J.; Xiang, F.; Hu, X. State-of-health estimation for fast-charging lithium-ion batteries based on a short charge curve using graph convolutional and long short-term memory networks. *J. Energy Chem.* **2024**, *98*, 1–11. [CrossRef]
17. Wang, F.; Xu, C.; Offer, G. Physics-Informed Neural Networks for Battery SOH Estimation. *Nat. Commun.* **2024**, *15*, 7188. [CrossRef]
18. Ren, J.; Wang, Y.; Liu, S. Hybrid Deep Learning with Transfer Learning for Lithium-Ion Battery SOH Estimation. *Energies* **2025**, *18*, 1491. [CrossRef]
19. Sedlařík, M.; Vyroubal, P.; Capkova, D.; Omerdic, E.; Rae, M.; Mačák, M.; Šedina, M.; Kazda, T. Advanced machine learning techniques for State-of-Health estimation in lithium-ion batteries: A comparative study. *Electrochim. Acta* **2025**, *524*, 145988. [CrossRef]
20. Sylvestrin, G.R.; Silva, R.; Kowal, J. State-of-the-Art Machine Learning Approaches for Battery SOH Estimation: A Systematic Review. *Energies* **2025**, *18*, 746. [CrossRef]
21. Borah, M.; Wang, Q.; Moura, S.; Sauer, D.U.; Li, W. Synergizing physics and machine learning for advanced battery management. *Commun. Eng.* **2024**, *3*, 134. [CrossRef]
22. Pandit, R.; Ahlawat, N. A standardized comparative framework for machine learning techniques in lith-ium-ion battery state of health estimation. *Future Batter.* **2025**, *7*, 100099. [CrossRef]
23. Berger, F.; Joest, D.; Barbers, E.; Quade, K.; Wu, Z.; Sauer, D.U.; Dechent, P. Benchmarking battery management system algorithms-Requirements, scenarios and validation for automotive applications. *ETransportation* **2024**, *22*, 100355. [CrossRef]
24. Sun, B.; Zhao, H.; Wang, C. Low-Temperature Performance, Modeling, and Heating of Lithium-Ion Batteries. *Energies* **2023**, *16*, 7142. [CrossRef]
25. Shanbedi, M.; Shahali, H.; Polycarpou, A.A.; Amiri, A. Advances and future prospects of low-temperature electrolytes for lithium-ion batteries. *EES Batter.* **2025**. [CrossRef]
26. Li, S.; Zhang, C.; Zhao, Y.; Offer, G.J.; Marinescu, M. Effect of Thermal Gradients on Inhomogeneous Degradation in Lithium-Ion Batteries. *Commun. Eng.* **2023**, *2*, 124. [CrossRef]
27. Wang, Z.; Zhao, Q.; Yu, X.; An, W.; Shi, B. Impacts of vibration and cycling on electrochemical characteristics of batteries. *J. Power Sources* **2024**, *601*, 234274. [CrossRef]
28. Jeong, D.; Kim, H. Lithium-Ion Batteries for Extreme Conditions (−40 °C to 50 °C). *Commun. Chem.* **2025**, *8*, 215. [CrossRef]
29. Liu, K.; Hu, X.; Lucu, M.; Widanage, W.D. Transfer Learning for Battery State Estimation and Ageing Prognostics. *Energy AI* **2023**, *13*, 100247. [CrossRef]
30. Zhang, J.; Yang, Z.; Wang, C. Realistic Deep Learning for Lithium-Ion Battery Fault Detection Deployable on BMS Data. *Nat. Commun.* **2023**, *14*, 5421. [CrossRef]
31. NASA Prognostics Center of Excellence (PCoE). Battery Data Set Repository. NASA. Available online: https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository (accessed on 1 March 2023).
32. NASA Data.gov. Randomized Battery Usage Datasets (2022–2025). Available online: https://catalog.data.gov/dataset/randomized-battery-usage-2-room-temperature-random-walk (accessed on 24 July 2025).
33. Dos Reis, G.; Strange, C.; Yadav, M.; Li, S. Lithium-Ion Battery Data and Where to Find It. *Patterns* **2021**, *2*, 100247. [CrossRef]