Journal of Industrial and Management Optimization

JIMO, 22(12): 6840–6866.
DOI:10.3934/jimo.2026228
Received: 17 September 2026
Revised: 06 November 2026
Accepted: 16 November 2026
Published: 24 November 2026

https://www.aimspress.com/journal/jimo

Research article

Risk-aware maintenance optimization for lithium-ion batteries

M. Musab Bayat*

Electrical and Electronics Engineering Department, Sivas Cumhuriyet University, Sivas, Türkiye

* **Correspondence:** Email: muhammed@cumhuriyet.edu.tr.

Abstract: Accurate battery health prediction alone does not ensure effective maintenance decisions if prognostic uncertainty and intervention timing are not explicitly considered. This study proposes an explainable, risk-aware predictive maintenance framework integrating multi-horizon state-of-health (SOH) prognosis, battery-aware uncertainty quantification, and leakage-safe maintenance policy evaluation. The framework was evaluated on 519 characterization observations from eight cells in the Oxford Battery Degradation Dataset under a strict leave-one-battery-out validation protocol. Bagged Trees predicted SOH at exact 100-, 500-, and 1000-cycle horizons, with pooled RMSE values of 0.0194, 0.0201, and 0.0211, respectively. Comparison with a privileged persistence baseline showed that persistence was particularly strong at short horizons, whereas the learned model provided improved predictive performance at the longest horizon. Battery-balanced empirical uncertainty bounds achieved mean held-out-cell coverages of 91.5%, 91.8%, and 90.4% at a nominal 90% level. Prognostic outputs were subsequently converted into maintenance triggers using an 80% SOH reference threshold. The risk-aware rule provided median advance warnings of 1450, 2050, and 2400 aging cycles across the three horizons, whereas the privileged reactive comparator provided no advance warning by construction. Censoring-aware timing-loss analysis and leakage-safe fixed-interval optimization characterized the trade-off between early intervention and delayed maintenance. In a separate 500-cycle joint decision optimization experiment, the risk-aware policy achieved a mean normalized timing loss of 1.1125, representing reductions of 38.2% and 32.6% relative to the optimized point-based (1.80) and fixed-interval (1.65) policies, respectively. However, the improvement over the point-based policy was confined to two of the eight held-out batteries, indicating that the aggregate improvement did not extend uniformly across cells. Global permutation importance and local Shapley analysis provided complementary explanations of model behavior and maintenance decisions. Overall, the results show that predictive accuracy alone is insufficient to characterize maintenance utility and that prediction horizon, prognostic uncertainty, and intervention rules should be considered jointly in battery predictive maintenance.

Keywords: explainable artificial intelligence; lithium-ion battery; predictive maintenance; state of health; uncertainty quantification
Mathematics Subject Classification: 90B25, 68T05

1. Introduction

Lithium-ion batteries (LIBs) are central to electric vehicles and stationary energy-storage systems, where progressive degradation affects reliability, safety, and lifecycle management. Consequently, state-of-health (SOH) estimation and future health prognosis have become important components of battery prognostics and health management. Machine-learning approaches are increasingly used to capture nonlinear degradation patterns from operational measurements; however, data scarcity, cell-to-cell variability, domain shift, and limited generalization remain major barriers to reliable deployment [1,2]. Complementary battery condition monitoring approaches have also been investigated. For example, He et al. [3] explored current-pulse-excited active acoustic emission for detecting casing defects in cylindrical lithium-ion batteries, providing an alternative diagnostic perspective on battery condition assessment.

Recent research has also approached lithium-ion battery management through complementary operational strategies. Es-satte et al. [4] reviewed advances in battery thermal management involving phase change materials, bio-inspired structures, and artificial intelligence, emphasizing the importance of temperature regulation for safe operation. Siricharoenpanich et al. [5] proposed a self-recovery multi-stage battery system that combines electrochemical recovery with adaptive scheduling to improve energy utilization. Although these approaches address thermal and operational performance rather than maintenance timing, they illustrate the broader importance of integrating battery condition information with management decisions.

Data-driven battery prognosis has progressed from conventional regression and ensemble methods toward deep sequence–learning architectures. Comprehensive reviews have identified short-term SOH estimation, end-of-life prediction, and degradation trajectory forecasting as distinct but related prognostic objectives, emphasizing that the suitability of a prediction method depends on the intended application and prediction horizon [6,7]. Physically meaningful health indicators derived from measurable signals provide compact representations of degradation [1], while interpretable ensemble models have demonstrated the predictive value of short voltage profiles [8]. Recent Transformer- and recurrent-based approaches have further extended battery health assessment toward multi-step prediction [9,10]. Mao et al. [11] integrated Bayesian filtering with neural network–generated observations for long-term battery RUL prediction, explicitly addressing observation uncertainty and incomplete measurements. This illustrates the continuing importance of robust forecasting under imperfect operational information. Nevertheless, predictive-maintenance applications require not only accurate health estimates but also reliable forecasts over decision-relevant future horizons.

A critical challenge is whether such performance transfers to batteries not represented during model development. Degradation trajectories vary across cells, operating conditions, temperatures, and chemistries. Transfer learning and domain adaptation have therefore been investigated to mitigate distribution shifts [12,13], while domain-generalization and cross-domain methods increasingly target previously unseen batteries or operating domains [14–16]. A standardized cross-battery comparison

has likewise emphasized battery-level separation between model development and evaluation [17]. Using the Oxford Battery Degradation Dataset, Mousavi et al. [18] developed a Bayesian neural network framework for SOH and RUL estimation from single-cycle health indicators and evaluated its generalization using held-out cells. This provides a closely related example of battery-level prognostic validation with uncertainty quantification, although the present study focuses on exact-horizon forecasts and their implications for maintenance decisions. More recently, Rizanov et al. [19] critically examined the trustworthiness and transferability of battery SOH and remaining useful life estimation, emphasizing the need for validation protocols that account for differences between laboratory cells, operating conditions, and real-world battery packs. These findings motivate strict battery-level validation when prognosis is intended for previously unseen cells. Beyond battery applications, Ren et al. [20] investigated remaining useful life prediction under varying operating conditions in ship electric propulsion systems using a domain adaptation framework. Their work illustrates the broader importance of addressing distribution shifts in data-driven prognostics.

Another challenge is prognostic uncertainty. Deterministic predictions provide no explicit indication of forecast reliability, although battery-to-battery variability can substantially affect degradation behavior. Probabilistic approaches to battery prognostics have therefore attracted increasing attention [2]. Recent approaches include Transformer ensembles for quantifying RUL uncertainty [21], data-driven SOH prediction intervals [22], and real-world EV frameworks combining health estimation with predictive intervals [23]. Rasul et al. [24] combined ensemble learning with conformal prediction to construct uncertainty intervals for lithium-ion battery SOH estimation across multiple benchmark datasets, including Oxford. Their findings further motivate evaluating prediction interval reliability alongside point prediction accuracy. Zhao et al. [25] further developed an uncertainty-aware remaining useful life prediction framework combining multi-stage degradation modeling with evidential regression, demonstrating the importance of distinguishing uncertainty sources and evaluating predictive interval calibration. However, uncertainty must also be assessed across held-out batteries, as satisfactory average coverage may conceal battery-specific calibration failures under distribution shift. Recent uncertainty quantification benchmarks in prognostics have also highlighted the limited robustness of several predictive uncertainty methods under out-of-distribution conditions [26]. This raises a decision-relevant concern: uncertainty estimates that appear satisfactory on average may still produce unreliable intervention triggers for individual batteries.

Interpretability is similarly important for data-driven health management. Shapley-based methods have been used to identify degradation indicators influencing SOH estimates in interpretable ensemble and explainability-driven frameworks [6,27], while recent studies have combined physically meaningful feature engineering with explainable artificial intelligence [28] and integrated uncertainty estimation with SHAP-based interpretation using real-world battery data [23]. These approaches improve transparency at the prediction level. Predictive maintenance, however, introduces an additional question: How do health forecasts and their associated uncertainty translate into maintenance intervention decisions?

This question highlights the distinction between accurate prognosis and effective maintenance decision-making. Low SOH prediction error does not necessarily provide sufficient advance warning, whereas conservative uncertainty-aware decisions may trigger earlier intervention at the expense of usable battery life. Recent studies have begun connecting battery prognosis with maintenance optimization. Zhen et al. integrated degradation updating with condition-based maintenance decision-making [29], while predictive opportunistic maintenance has combined point and probabilistic RUL

prediction with multi-objective maintenance optimization [30]. Valizadeh et al. [31] connected physics-informed battery health prediction with Markov-based reliability, availability modeling, and replacement-time optimization, demonstrating how prognostic information can be incorporated into maintenance planning and economic decision-making. Patrizi et al. [32] reviewed battery degradation modeling and remaining useful life prediction approaches in relation to testing strategies and predictive maintenance, highlighting the importance of connecting prognostic information with maintenance planning. Che et al. [33] incorporated probabilistic degradation predictions into early detection of accelerating battery aging, illustrating how prognostic uncertainty can support warning decisions. Together, these studies demonstrate growing interest in integrating battery prognosis with maintenance decisions, although their objectives and evaluation protocols differ. Despite these advances, an important evaluation problem remains: accurate SOH forecasts and informative uncertainty estimates do not necessarily lead to better maintenance timing. Existing approaches employ different prognostic objectives, uncertainty treatments, and maintenance optimization criteria, making it difficult to determine whether uncertainty-aware intervention rules provide a consistent decision-level benefit across previously unseen batteries. In particular, earlier warnings may lead to premature intervention, while aggregate improvements may be driven by outcomes from only a few batteries. This motivates a battery-disjoint evaluation that explicitly separates predictive accuracy, uncertainty reliability, and maintenance decision quality, while examining both predefined intervention rules and jointly optimized maintenance policies.

To address this gap, this study develops an explainable, risk-aware predictive-maintenance framework for LIBs. SOH forecasts are generated at exact 100-, 500-, and 1000-cycle horizons under strict leave-one-battery-out evaluation and combined with battery-aware empirical uncertainty bounds for maintenance triggering. Prognosis-driven decisions are evaluated against privileged persistence and reactive reference strategies, as well as a leakage-safe fixed-interval strategy optimized exclusively using development batteries. To extend the evaluation beyond predefined maintenance rules, a separate 500-cycle decision experiment jointly optimizes the prognostic alarm threshold and scheduled maintenance deadline by minimizing asymmetric intervention-timing loss using development batteries, followed by held-out battery evaluation. Right-censored no-trigger cases are distinguished from observable missed triggers, and normalized scenario-based timing loss is used instead of unsupported monetary assumptions. Finally, feature-level model reliance and local Shapley attribution are connected with decision-level decomposition to trace how degradation indicators, prognosis, and uncertainty contribute to maintenance decisions.

The originality of this study lies in its decision-oriented evaluation of battery health prognosis under a unified, battery-disjoint validation framework, rather than in proposing a new forecasting algorithm. Specifically, the study distinguishes the predictive accuracy of multi-horizon SOH forecasts from the reliability of their uncertainty estimates and the quality of the maintenance decisions they support. Beyond evaluating predefined intervention rules, it jointly optimizes prognostic alarm thresholds and scheduled maintenance deadlines under a common asymmetric timing-loss objective, using development batteries only. The framework further accounts for right-censored warning outcomes and examines the battery-level sensitivity of the resulting policy comparisons. This design enables a clearer assessment of when uncertainty-aware prognosis improves maintenance decisions and when earlier warnings do not translate into better intervention timing.

Related methodological developments in other signal-based monitoring applications further illustrate the importance of feature representation, model selection, and reliable performance

evaluation. For example, EEG-based driver-state studies have investigated nonlinear feature extraction combined with residual attention networks [34], multi-band feature optimization with conventional machine learning classifiers [35], and subject-independent evaluation using leave-one-subject-out cross-validation [36]. A recent review has also highlighted preprocessing, feature extraction, evaluation metrics, and generalization as important considerations in data-driven monitoring systems [37]. Although these studies address a different application domain, they provide complementary methodological perspectives on the design and evaluation of learning-based monitoring frameworks.

The main contributions are:

- Leakage-safe multi-horizon prognosis: Exact 100-, 500-, and 1000-cycle SOH prediction under strict battery-level holdout with development-only preprocessing and model selection.
- Uncertainty-aware maintenance triggering: Battery-aware empirical uncertainty is propagated into the intervention rule to distinguish predictive accuracy from maintenance-warning capability.
- Maintenance decision optimization: In addition to evaluating predefined point-based and risk-aware warning rules and a development-optimized fixed-interval strategy, a separate 500-cycle experiment jointly optimizes the prognostic alarm threshold and scheduled maintenance deadline for both point-based and risk-aware policies. Battery-disjoint nested validation is used to compare the optimized policies against an optimized fixed-interval reference under a common asymmetric intervention-timing objective. Battery-level results and a descriptive influence analysis are presented to examine how performance differences vary across cells and how sensitive the aggregate findings are to individual battery outcomes.
- Decision-oriented explainability: Global feature reliance and local Shapley attribution are linked with decision-level uncertainty decomposition to trace degradation information from prognosis to maintenance action.

The remainder of this paper is organized as follows: Section 2 describes the battery dataset, health-indicator extraction, multi-horizon prognosis models, battery-disjoint validation, uncertainty quantification, maintenance decision policies, and explainability methods. Section 3 presents the experimental results, including cross-battery prediction performance, uncertainty calibration, maintenance warning and timing analyses, joint policy optimization, and model interpretability. Finally, Section 4 summarizes the main findings and conclusions of the study.

2. Materials and methods

2.1. Overall framework

The proposed framework integrates multi-horizon SOH prognosis with uncertainty-aware predictive-maintenance decision-making. As illustrated in Figure 1, discharge measurements are first processed to derive health indicators, followed by SOH prognosis at exact 100-, 500-, and 1000-cycle horizons. A strict leave-one-battery-out framework separates the development batteries from the held-out battery throughout model development and evaluation. Residuals from the development batteries are then used to quantify prognostic uncertainty, which informs point-based and risk-aware maintenance triggers and their subsequent censoring-aware evaluation. Finally, feature- and decision-level explainability is used to trace the pathway from degradation indicators to maintenance actions.

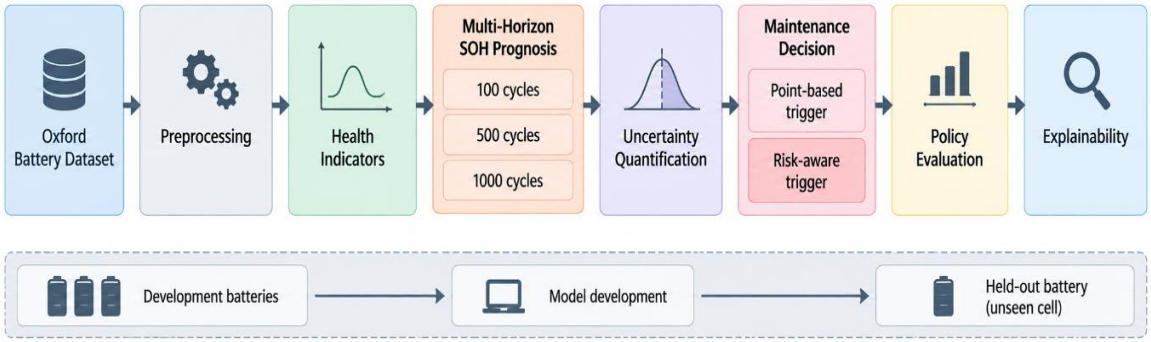


Figure 1. Overall workflow of the proposed multi-horizon, uncertainty-aware predictive-maintenance framework.

2.2. Dataset and SOH definition

The Oxford Battery Degradation Dataset was used to evaluate the proposed framework [38]. The dataset contains eight lithium-ion pouch cells with a rated capacity of 740 mAh subjected to long-term aging and periodic characterization at 40 °C. The present study used the C1 discharge records, comprising 519 characterization observations across the eight cells. Discharge time was converted from MATLAB serial-date format to elapsed seconds, and discharge capacity was calculated as the absolute difference between the initial and final values of the recorded charge-throughput signal. All 519 records passed the basic time-series integrity checks and were retained without smoothing, interpolation, or manual correction. Four observations with relatively large capacity changes between consecutive characterizations were retained and flagged for subsequent sensitivity analysis.

SOH was defined relative to the rated capacity as

$$SOH_t = \frac{C_t}{C_{rated}} = \frac{C_t}{740} \tag{1}$$

where C_t is the measured discharge capacity at characterization t, and C_{rated} = 740 mAh is the rated capacity of the Oxford cells.

2.3. Health indicators and multi-horizon prognosis dataset

Health indicators were extracted from each C1 discharge record without using capacity, SOH, aging cycle, or discharge duration as predictors. The final electrical feature set comprised voltage range (V_{range}) and normalized voltage slope (V_{slope}). V_{range} was calculated as the difference between the maximum and minimum discharge voltage, while V_{slope} was obtained from the voltage trajectory with respect to discharge time normalized to the interval [0,1]. A multidomain variant additionally included mean temperature (T_{mean}) and normalized temperature slope (T_{slope}), with the latter calculated analogously on the normalized elapsed-time axis. This variant was used to assess whether temperature information improved cross-battery generalization. Missing temperature features were imputed using development-data medians only.

For each characterization record at aging cycle c, future SOH targets were matched within the same battery using the original aging-cycle labels at exact horizons of 100, 500, and 1000 cycles. A

source–target pair was constructed only if an SOH measurement was available at the exact target cycle c + h, where h $\in$ {100, 500, 1000} cycles. If the required future measurement was unavailable, the corresponding source record was excluded from the dataset for that specific prediction horizon. No interpolation, extrapolation, nearest-cycle substitution, or imputation was applied to missing future SOH targets. This procedure yielded 470, 444, and 404 valid source–target pairs at the 100-, 500-, and 1000-cycle horizons, respectively. Consequently, source observations near the end of a battery trajectory were excluded from training, validation, and testing for a given prediction horizon whenever the corresponding future SOH measurement was unavailable. These observations were not assigned artificial target values or treated as prediction errors.

2.4. Prognostic models and validation

Four regression models were evaluated: linear regression, support vector regression (SVR), Bagged Trees, and LSBoost. These models were selected to represent complementary learning approaches within a controlled comparison: linear regression provides an interpretable linear baseline, SVR captures nonlinear relationships through kernel-based regression, Bagged Trees represents bootstrap-aggregated tree ensembles, and LSBoost represents sequential gradient-boosted ensembles. This selection enables the comparison of distinct regression mechanisms using the same compact health indicators and battery-disjoint validation protocol. Given the limited number of independent battery cells, the emphasis was placed on a consistent and reproducible evaluation of these model families rather than an exhaustive search across prognostic architectures. Model selection and hyperparameter tuning were performed using nested leave-one-battery-out (LOBO) validation. In each outer fold, one battery was reserved exclusively for testing, while the remaining seven batteries formed the development set. Hyperparameter selection was performed by inner LOBO validation using only these development batteries, after which the selected model configuration was fitted to the complete seven-battery development set and evaluated on the held-out battery. Hyperparameters were selected independently for each prediction horizon, feature set, and outer held-out battery by minimizing the mean battery-level RMSE across the seven inner LOBO validation folds. Linear regression was used without hyperparameter tuning. For SVR with a Gaussian kernel, the BoxConstraint and KernelScale parameters were each searched over {0.1, 1, 10}. For Bagged Trees, the numbers of trees were {100, 300}, and the minimum leaf sizes were {1, 5, 10}. For LSBoost, the numbers of boosting iterations were {100, 300}, the learning rates were {0.03, 0.1}, and the minimum leaf sizes were {5, 10}. Accordingly, the selected hyperparameter configurations could differ between outer folds rather than being fixed across all batteries. The complete selected hyperparameter configurations and inner validation results were retained for reproducibility. The selected hyperparameters varied across prediction horizons, feature sets, and held-out batteries. Across the 48 outer configurations per model, the most frequently selected settings were C = 1 and KernelScale = 10 for SVR (18/48), 100 trees and a minimum leaf size of 5 for Bagged Trees (15/48), and 300 boosting iterations, a learning rate of 0.03, and a minimum leaf size of 10 for LSBoost (23/48). No single hyperparameter combination was selected consistently across all outer configurations for any of the tuned models.

As an additional benchmark, two multilayer perceptron (MLP) regressors were evaluated, comprising a single-hidden-layer network (MLP-1) and a two-hidden-layer network (MLP-2). The candidate hidden-layer sizes were 8 and 16 neurons for MLP-1, and [8, 4] and [16, 8] for MLP-2, with regularization parameters of 0.001 and 0.01. Both architectures used rectified linear unit activation and

a maximum of 300 training iterations. Hyperparameters were selected independently for each prediction horizon, feature set, and outer fold using the seven development batteries under nested LOBO validation. Training-only imputation and standardization were applied throughout. Sensitivity to random initialization was subsequently assessed across 10 training seeds, with the selected hyperparameters held fixed. This supplementary benchmark did not alter the original model selection, uncertainty quantification, or maintenance optimization experiments. Both MLP architectures were implemented in MATLAB R2023b using the fitrnet function from the Statistics and Machine Learning Toolbox, with the solver and loss settings left at their MATLAB defaults.

A persistence baseline was also evaluated by assigning the observed SOH at the source cycle c as the prediction for cycle c + h, for each horizon $h \in \{100, 500, 1000\}$. Unlike the learned regression models, which used discharge-derived health indicators without measured SOH as an input, this baseline directly accessed the measured source-cycle SOH of the held-out battery. It was therefore designated as privileged. Importantly, the baseline did not use the future SOH at c + h or any future observations to construct its prediction.

Missing values were imputed using training-fold medians, followed by training-fold standardization. The resulting preprocessing parameters were applied to validation and test data without recalculation. Model performance was evaluated using RMSE, MAE, R^2, and prediction bias. Thus, in every outer fold, the held-out battery was excluded from all preprocessing parameter estimation, hyperparameter tuning, model fitting, and uncertainty calibration. Preprocessing parameters were estimated separately within the relevant training folds, and uncertainty calibration was performed exclusively using development-battery residuals, as detailed in Section 2.5.

Although the dataset comprises 519 characterization observations, these measurements originate from only eight battery cells and are therefore not statistically independent. The nested battery-disjoint validation procedure described above provides a controlled internal evaluation while preventing information leakage from held-out cells. Nevertheless, the limited number of independent batteries constrains the precision and generalizability of the resulting performance estimates, motivating further validation on additional datasets.

2.5. Battery-aware uncertainty quantification

Prediction uncertainty was estimated from absolute prognostic residuals obtained exclusively from the development batteries. For each outer LOBO fold and prediction horizon, each development battery was internally held out once, producing battery-specific empirical residual distributions. These distributions were combined with equal battery weighting to prevent cells with more observations from dominating calibration.

The resulting uncertainty bounds represent the combined effects of prediction errors observed across development batteries rather than a decomposition into separate uncertainty components. Measurement noise may contribute through variability in the recorded signals and SOH measurements, while model-related uncertainty is reflected indirectly in the residuals generated by models trained on different battery subsets. Battery-to-battery variability is incorporated through the battery-disjoint calibration procedure and equal weighting of the development cells. Because these sources are not estimated independently, the resulting empirical bounds reflect their combined contribution to predictive uncertainty rather than providing separate estimates of measurement, model, and inter-battery uncertainty.

For a nominal 90% level, the battery-balanced empirical residual quantile was used to construct the prediction interval defined in Eq. (2):

$$[\widehat{SOH}_{t+h} - q_{h,c}^{bal}, \widehat{SOH}_{t+h} + q_{h,c}^{bal}] \tag{2}$$

where $\widehat{SOH}_{t+h}$ denotes the predicted SOH at future cycle $t+h$, h is the prediction horizon (100, 500, or 1000 cycles), c denotes the held-out battery in the outer LOBO fold, and $q_{h,c}^{bal}$ is the battery-balanced empirical quantile estimated exclusively from development-battery residuals. The outer test battery was not used for uncertainty calibration. Empirical interval coverage and width were subsequently evaluated on the held-out battery in each outer fold.

2.6. Maintenance decision and policy evaluation

Future SOH predictions were converted into maintenance decisions using a reference threshold of SOHcrit = 0.80. Two prognosis-driven decision rules were considered. The point-based policy triggers an intervention when the predicted future SOH reaches the critical threshold, whereas the risk-aware policy triggers when the lower battery-balanced prediction bound reaches the same threshold. The corresponding trigger cycles are defined in Eq. (3):

$$t_{P,h} = min\{t : \widehat{SOH}_{t+h} \leq SOH_{crit}\} \qquad t_{R,A,h} = min\{t : \widehat{SOH}_{t+h} - q_{h,c}^{bal} \leq SOH_{crit}\} \tag{3}$$

Here, $t_{P,h}$ and $t_{R,A,h}$ denote the first intervention cycles identified by the point-based and risk-aware policies, respectively; t is the source characterization cycle; h is the prediction horizon (100, 500, or 1000 cycles); $\widehat{SOH}_{t+h}$ is the predicted SOH at the future cycle $t+h$; and $q_{h,c}^{bal}$ is the battery-balanced uncertainty quantile defined in Section 2.5.

For each battery, the first observed cycle at which the measured SOH reached or fell below SOH_{crit} was denoted by t_{crit}. This observed crossing was used as the reference for evaluating intervention timing. A privileged reactive comparator, which has access to the observed current SOH, therefore intervenes at t_{crit} and provides no advance warning. When a policy did not trigger because its evaluable prediction window ended before t_{crit}, the case was treated as right-censored rather than as a missed maintenance event.

To quantify the trade-off between premature and delayed intervention, the normalized timing loss in Eq. (4) was used:

$$J(t_\pi, t_{crit};\, \lambda_L) = \frac{(t_{crit} - t_\pi)_+}{1000} + \lambda_L \frac{(t_\pi - t_{crit})_+}{1000} \tag{4}$$

In Eq. (4), t_π denotes the intervention cycle generated by policy π, t_{crit} is the observed SOH-threshold crossing, and $(x)_+ = max(x, 0)$. The first term represents the normalized loss associated with early intervention and unused battery life, whereas the second represents delayed intervention beyond the observed threshold crossing. The parameter λ_L controls the relative penalty assigned to delayed intervention and was evaluated at $\lambda_L = \{1,2,5,10\}$. These values represent dimensionless sensitivity scenarios and should not be interpreted as monetary maintenance costs.

For the original three-horizon policy evaluation, a fixed-interval preventive strategy was optimized over candidate intervention cycles from 500 to 5000 cycles in 500-cycle increments. For each held-out battery, the fixed interval was selected by minimizing the mean timing loss in Eq. (4) using only the other seven development batteries and was subsequently applied unchanged to the held-

out battery. In this evaluation, the point-based and risk-aware policies defined in Eq. (3) used prespecified decision thresholds, whereas the fixed-interval strategy involved development-side optimization.

2.7. Joint maintenance decision optimization

To extend the predefined maintenance rules described in Section 2.6, a separate joint optimization experiment was performed at the 500-cycle prediction horizon. This experiment used a Bagged Trees regression model with the ElectricalOnly feature set, fixing the number of trees at 100 and the minimum leaf size at 5 for all model fits. These hyperparameters were specified independently of the nested hyperparameter search used for the main multi-horizon benchmark. All prediction and calibration models were fitted using training-only median imputation and standardization. Two decision variables were considered: the prognostic alarm threshold, θ, expressed in normalized SOH, and the scheduled maintenance deadline, d, expressed in aging cycles. These parameters were selected using development batteries and subsequently evaluated on an excluded test battery under the asymmetric intervention-timing loss defined in Eq. (4).

For each held-out battery b, let $\mathcal{O}_b$ denote the set of source characterization cycles with valid 500-cycle-ahead SOH predictions. The alarm cycle is defined as the earliest cycle in this set at which the predicted SOH, adjusted downward by the age-dependent empirical calibration radius, reaches or falls below θ. The intervention cycle is defined as the earlier of the prognostic alarm cycle and the scheduled deadline d.

$$a_b(\theta) = inf\left\{t \in \mathcal{O}_b : \widehat{SOH}_{b,t+500} - q_b^{outer}(t) \le \theta\right\} \tag{5}$$

$$\tau_b(\theta, d) = min\{a_b(\theta), d\}, \qquad inf\emptyset = +\infty \tag{6}$$

The term $q_b^{outer}(t)$ in Eq. (5) denotes the age-dependent empirical calibration radius for held-out battery b, estimated exclusively from the seven development batteries. It is determined across three predefined aging-cycle bins: below 3000, from 3000 to below 5000, and 5000 or above. Within each bin, the radius is the maximum of the battery-specific 90th percentiles of absolute prediction residuals. If fewer than three calibration batteries contribute observations to a bin, or fewer than 15 residuals are available in total, a global calibration radius is used instead. This global radius is defined as the maximum of the battery-specific 90th percentiles calculated over all available calibration residuals. The calibration radius is an empirical uncertainty adjustment and does not imply a distribution-free coverage guarantee. If no alarm condition is satisfied, the alarm cycle is assigned positive infinity, allowing the scheduled deadline to determine the intervention cycle in Eq. (6).

For the risk-aware policy, the optimal alarm threshold and scheduled maintenance deadline were jointly selected for each outer held-out battery by minimizing the mean intervention-timing loss across the seven development batteries, as formulated in Eqs. (7) and (8). Each development battery was evaluated using its own held-out predictions and an independently constructed inner calibration radius. The outer test battery was excluded from both parameter selection and calibration.

$$J_b(\theta, d) = \frac{1}{7}\sum\nolimits_{c \ne b} L\left(\tau_c^{(b)}(\theta, d), t_{cirit,c}; 5\right) \tag{7}$$

$$J_b(\theta_b^*, d_b^*) \le J_b(\theta, d), \quad \forall (\theta, d) \in \Theta \times \mathcal{D} \tag{8}$$

Here, J denotes the mean intervention-timing loss across the development batteries, while Θ and $\mathcal{D}$ represent the candidate sets for the alarm threshold and scheduled deadline, respectively. The superscript (b) indicates that the inner validation procedure is performed within the development set associated with outer held-out battery b. The optimal parameter pair is denoted by an asterisk. The asymmetric loss in Eq. (4) is evaluated with a late-intervention penalty of 5.

The same nested development-set evaluation was also used to optimize the comparison policies. For the fixed-interval policy, only the scheduled intervention deadline was optimized. For the point-based policy, the alarm threshold and scheduled deadline were jointly optimized using the unadjusted SOH predictions. The risk-aware policy used the same candidate search and timing-loss objective as the point-based policy, but its alarm condition incorporated the age-dependent empirical calibration radius. This comparison therefore examines the contribution of uncertainty-aware alarm construction and policy selection within a common evaluation framework, rather than comparing an optimized policy with predefined alternatives.

The candidate search included alarm thresholds from 0.68 to 0.96 in increments of 0.01 and scheduled maintenance deadlines from 500 to 8000 cycles in increments of 500 cycles. For each outer held-out battery b, policy selection was performed using the remaining seven development batteries. Within this development set, each battery c served in turn as an inner validation battery. Its 500-cycle-ahead predictions were obtained from a model trained on the other six development batteries. Separately, the inner calibration radius was constructed from residuals of these six batteries, with each calibration battery predicted by a model trained on the other five. After selecting the optimal parameter pair, the outer test battery was evaluated using predictions from a model trained on all seven development batteries and a calibration radius derived from leave-one-battery-out residuals within that development set. Candidate pairs were examined in ascending order of threshold and then deadline. A candidate replaced the current best only when its mean development loss was lower by more than 10^{-12}. Intervention times were not truncated at the last available prediction source cycle; deadlines beyond the observation window therefore represent planned interventions rather than observed maintenance events.

All computational experiments were implemented in MATLAB R2023b (Update 7, version 23.2.0.2515942) using the Statistics and Machine Learning Toolbox (version 23.2) under Windows 11 Pro. The main nested LOBO benchmark was initialized using the Mersenne Twister random number generator with seed 20260916. For the separate joint maintenance optimization experiment, deterministic seeds were assigned to individual model fits: 100000 + i for outer test predictions, 200000 + 100i + j for development-battery predictions, and 300000 + 1000i + 100j + k for inner calibration models, where i, j, and k denote the corresponding loop indices. This seed specification permits the model-fitting sequence to be reproduced independently of the main benchmark.

The statistical reliability of maintenance policy comparisons was assessed using paired battery-level intervention-timing losses from the eight held-out batteries. Two-sided exact paired sign-flip tests were applied to the nonzero differences between the optimized risk-aware policy and each comparator (point-based and fixed-interval). Holm correction was used to account for the two comparisons, with statistical significance evaluated at the 0.05 level.

2.8. Explainability analysis

Explainability was assessed at both feature and decision levels. Global feature reliance was evaluated using permutation importance, quantified as the increase in RMSE after permuting each feature on development-battery data. Local Shapley analysis was then used to quantify the contributions of V_{range} and V_{slope} to a representative 500-cycle SOH prediction for held-out Cell 1.

Because the original stochastic Bagged Trees ensembles were not stored, the outer-fold models were reconstructed using the frozen data splits, preprocessing parameters, and selected hyperparameters solely for post hoc feature attribution. Reconstruction stability was verified against the frozen predictions, while all reported prognostic, uncertainty, and maintenance results remained based on the original frozen outer-fold predictions. At the decision level, risk-aware triggers were classified according to whether the 0.80 SOH threshold was crossed by the point prediction itself or only after incorporating the battery-balanced uncertainty bound. This analysis linked the health indicators and prognostic uncertainty to the resulting maintenance decision.

3. Results

3.1. Battery degradation characteristics and prognosis dataset

The eight cells exhibited different degradation trajectories, providing a heterogeneous dataset for evaluating cross-battery prognostic performance. As shown in Figure 2, SOH generally decreased from near-rated initial values toward or below the 0.80 reference threshold, although the degradation rate and trajectory length varied across cells. Noticeable local fluctuations were observed in Cell 2, while Cell 5 exhibited a sharp terminal SOH decline. These variations originated from the original characterization records and were retained without smoothing, interpolation, or manual correction; the corresponding observations were flagged for subsequent sensitivity analysis.

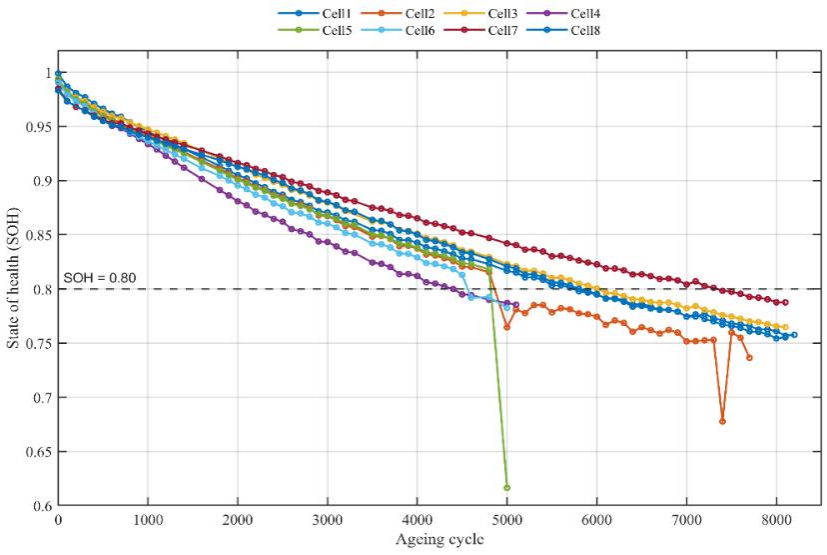


Figure 2. SOH degradation trajectories of the Oxford battery cells with the 0.80 SOH reference threshold.

The final dataset contained 519 characterization observations. Exact within-cell target matching yielded 470, 444, and 404 source–target pairs for the 100-, 500-, and 1000-cycle prediction horizons, respectively. The number of available pairs decreased with increasing prediction horizon because each battery trajectory contained a finite number of observations.

3.2. Cross-battery multi-horizon prognosis performance

The four regression models differed in their ability to generalize to held-out batteries. As shown in Figure 3, Bagged Trees achieved the lowest mean held-out-cell RMSE among the four evaluated regression models at all three prediction horizons. It was therefore selected as the prognostic model for the subsequent uncertainty quantification and maintenance analyses, based on its consistent comparative performance under the battery-disjoint evaluation protocol. The pooled predictive performance of the selected Bagged Trees model is reported separately in Table 1. These results characterize generalization to held-out cells within the Oxford dataset and should not be interpreted as evidence of predictive performance on batteries drawn from independent datasets or different operating domains.

To assess the robustness of the original model choice, Bagged Trees was additionally compared with two MLP architectures under the same battery-disjoint evaluation protocol. Table 2 reports the mean held-out-cell RMSE for both feature sets across the three prediction horizons. Bagged Trees achieved lower mean RMSE than both MLP architectures at all three horizons with multidomain features. With electrical-only features, MLP-1 achieved lower mean RMSE at the 500- and 1000-cycle horizons, whereas Bagged Trees performed better at 100 cycles. MLP-1 achieved lower mean RMSE than MLP-2 in all six feature-set and prediction-horizon combinations. However, exact paired Wilcoxon signed-rank tests at the battery level revealed no statistically significant differences between Bagged Trees and MLP-1, either before or after Holm adjustment for multiple comparisons. The ten-seed analysis further indicated that the observed performance patterns were generally stable with respect to neural network initialization.

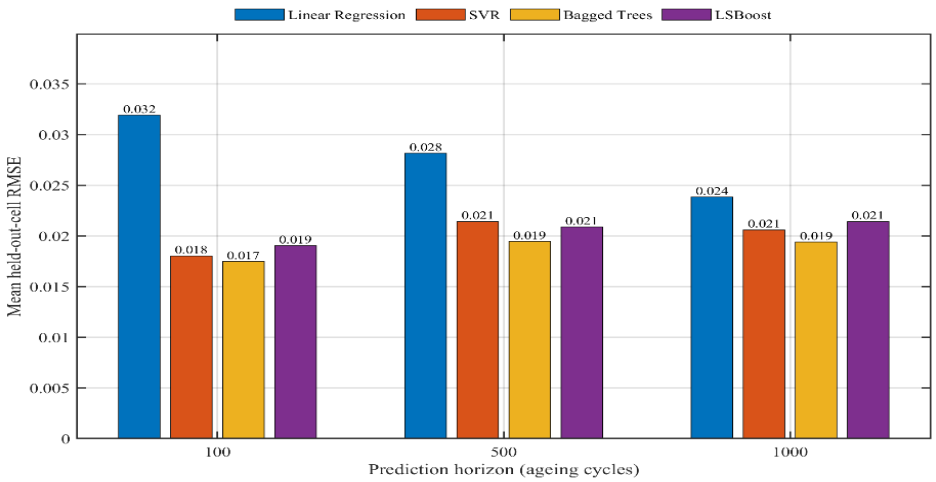


Figure 3. Mean held-out-cell RMSE of the prognostic models across prediction horizons.

Table 1. Pooled cross-battery performance of Bagged Trees.

Horizon (cycles)	RMSE	MAE	R^2	Bias
100	0.0194	0.0127	0.912	0.0027
500	0.0201	0.0123	0.893	0.0013
1000	0.0211	0.0121	0.857	0.0019

Table 2. Mean held-out-cell RMSE of Bagged Trees and MLP regressors across feature sets and prediction horizons.

Features	Horizon	Bagged Trees	MLP-1	MLP-2
ElectricalOnly	100	0.01746	0.01855	0.02070
ElectricalOnly	500	0.01945	0.01842	0.02077
ElectricalOnly	1000	0.01940	0.01743	0.01938
MultiDomain	100	0.02086	0.02425	0.02497
MultiDomain	500	0.02093	0.02368	0.02450
MultiDomain	1000	0.02082	0.02114	0.02201

Bagged Trees was retained as the primary prognostic model for three reasons. First, it had already demonstrated the lowest mean held-out-cell RMSE among the four conventional regression models at all three prediction horizons under the original nested LOBO benchmark. Second, the supplementary neural network experiments showed that its predictive accuracy remained competitive, particularly for the multidomain feature set, without establishing a statistically significant advantage for either model family. Third, Bagged Trees provides a practical ensemble regression approach for the compact, engineered health indicators used in this study, without requiring the additional architectural and initialization choices associated with neural networks. Importantly, the objective of the proposed framework is not to establish a universally optimal regression architecture but to investigate how battery-disjoint prognosis, calibrated uncertainty, and maintenance decision optimization interact.

Table 3. Cell-wise RMSE and MAE of Bagged Trees across three prediction horizons.

Held-out cell	100 RMSE	100 MAE	500 RMSE	500 MAE	1000 RMSE	1000 MAE
Cell 1	0.0173	0.0133	0.0146	0.0112	0.0139	0.0111
Cell 2	0.0344	0.0219	0.0299	0.0206	0.0367	0.0250
Cell 3	0.0111	0.0079	0.0118	0.0086	0.0108	0.0078
Cell 4	0.0222	0.0166	0.0226	0.0143	0.0150	0.0108
Cell 5	0.0108	0.0081	0.0325	0.0124	0.0360	0.0122
Cell 6	0.0117	0.0092	0.0124	0.0090	0.0145	0.0092
Cell 7	0.0175	0.0125	0.0162	0.0112	0.0140	0.0090
Cell 8	0.0147	0.0103	0.0155	0.0112	0.0142	0.0105

The cell-wise results in Table 3 reveal substantial heterogeneity that is not apparent from the pooled metrics. Cell 2 exhibited the highest RMSE at the 100- and 1000-cycle horizons (0.0344 and 0.0367, respectively), while Cell 5 had the highest RMSE at the 500-cycle horizon (0.0325). Notably,

the RMSE for Cell 5 increased from 0.0108 at 100 cycles to 0.0360 at 1000 cycles. These findings indicate that favorable aggregate performance does not imply uniformly accurate predictions across held-out batteries. The reported cell-wise errors therefore complement the pooled results by making battery-specific limitations explicit.

The effect of prediction horizon can be further examined by comparing Bagged Trees with the privileged persistence baseline. As illustrated in Figure 4, persistence was particularly strong at the 100-cycle horizon, while the two approaches were nearly equivalent in mean held-out-cell RMSE at 500 cycles. At 1000 cycles, Bagged Trees achieved lower RMSE than persistence for seven of the eight held-out cells. These horizon-dependent results suggest that persistence provides a strong short-term reference, whereas the learned model offers greater predictive benefits at longer horizons. Persistence should nevertheless be interpreted as a privileged comparator because it directly uses the observed source SOH, which was not provided to the learned model.

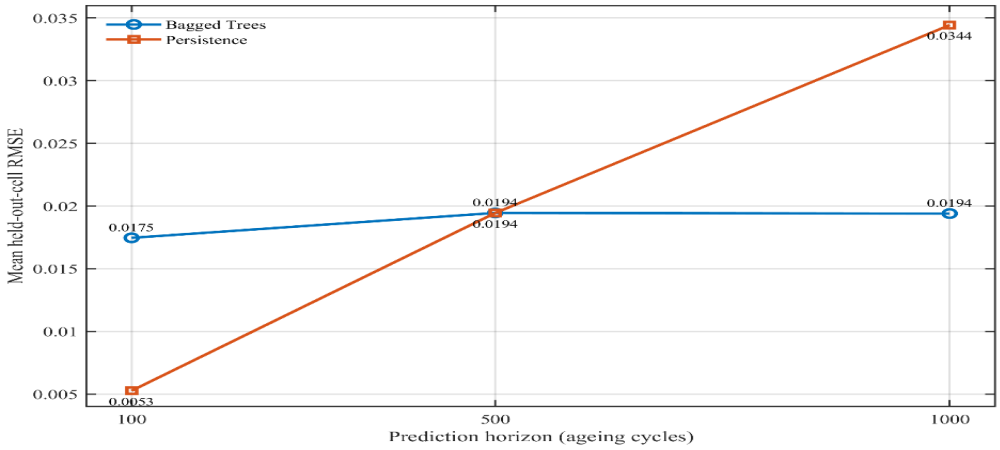


Figure 4. Mean held-out-cell RMSE of Bagged Trees and the privileged persistence baseline across prediction horizons.

At the 1000-cycle horizon, Bagged Trees outperformed the privileged persistence baseline in seven of the eight held-out batteries. A paired comparison of cell-wise RMSE values confirmed a statistically significant advantage for Bagged Trees (two-sided exact Wilcoxon signed-rank test, $p = 0.0156$). This advantage remained statistically significant after Holm correction across the three prediction horizons (adjusted $p = 0.03125$). The mean reduction in battery-level RMSE was 0.0150 SOH units, indicating that the improvement was also evident in the battery-level comparisons rather than only in the pooled metrics. A complementary exact sign test yielded $p = 0.0703$, reflecting the limited resolution of a direction-only test with eight paired observations.

Recent studies provide useful reference points for interpreting these results. Ali et al. [39] evaluated LSTM, TCN, and Temporal Fusion Transformer models on the NASA and Oxford battery datasets using leave-one-cell-out validation. On the Oxford dataset, their best-performing model achieved an RUL prediction RMSE of 4.27 ± 0.15 h. Although both studies evaluate generalization to unseen Oxford cells, the RUL prediction task considered by Ali et al. [39] differs from the exact-horizon SOH forecasting task examined here. Consequently, their prediction errors cannot be directly compared numerically. Wu et al. [40] investigated uncertainty-aware Transformer-based SOH estimation on NASA cells and reported a prediction interval coverage of 0.965 for cells within the

training degradation range, compared with 0.697 for cells extending beyond that range. These findings emphasize the importance of evaluating predictive uncertainty at the cell level. In contrast to these studies, the present framework evaluates exact-horizon SOH prediction, battery-balanced uncertainty intervals, and the resulting maintenance intervention decisions within a unified battery-disjoint validation setting. Thus, the principal contribution is not a claim of superior prediction accuracy over models evaluated under different protocols but an integrated assessment of how prognostic performance and uncertainty affect maintenance decisions.

3.3. Uncertainty quantification

The battery-balanced empirical prediction intervals were evaluated at a nominal 90% coverage level. As summarized in Table 4, mean held-out-cell coverage remained close to the nominal level across all three prediction horizons. Mean interval width varied only slightly across prediction horizons. However, the increasing variability in coverage between batteries indicates that average coverage alone does not adequately characterize uncertainty performance.

Table 4. Battery-balanced uncertainty performance across prediction horizons.

Horizon (cycles)	Mean coverage (%)	SD (percentage points)	Mean interval width
100	91.5	8.0	0.0504
500	91.8	8.3	0.0497
1000	90.4	10.3	0.0485

Despite near-nominal mean coverage across held-out batteries, calibration reliability varied substantially at the cell level. In particular, Cell 2 achieved only 65.5% empirical coverage at the 1000-cycle horizon, representing a 24.5-percentage-point shortfall relative to the nominal 90% target, despite a mean held-out-cell coverage of 90.4%. This discrepancy demonstrates that satisfactory aggregate coverage can conceal substantial battery-specific uncertainty failures and highlights the need for cell-level calibration assessment when prediction intervals inform maintenance decisions.

This limitation is particularly important for maintenance decisions because underestimating predictive uncertainty may influence intervention timing. Accordingly, the battery-balanced intervals are interpreted as empirical decision-support bounds rather than universal probabilistic guarantees. The following section therefore examines how the lower prediction bound influences maintenance warnings through the risk-aware triggering rule.

3.4. Maintenance warning performance

The multi-horizon SOH predictions were translated into maintenance warnings using the 0.80 SOH reference threshold. The risk-aware rule generated a trigger for all eight held-out batteries at each prediction horizon, whereas the point-based rule triggered for six, seven, and eight batteries at the 100-, 500-, and 1000-cycle horizons, respectively. All three cases in which the point-based policy did not generate a trigger were right-censored, as the evaluable prediction window ended before the first observed threshold crossing. Therefore, none could be classified as an observed missed trigger.

To quantify the resulting warning time, advance warning was defined as the number of aging cycles between the policy trigger cycle and the first observed cycle at which SOH reached or fell below SOHcrit. The median advance-warning values across the held-out batteries are reported in Table 5.

Table 5. Median advance warning before the first observed SOH threshold crossing.

Prediction horizon (cycles)	Point-based warning (cycles)	Risk-aware warning (cycles)
100	350	1450
500	1200	2050
1000	1600	2400

Advance warning increased with prediction horizon for both prognosis-driven policies. The risk-aware rule consistently produced earlier warnings because its trigger incorporated the lower battery-balanced prediction bound, whereas the point-based rule relied only on the predicted future SOH. The privileged reactive comparator, in contrast, provides zero advance warning by construction because intervention occurs at the observed SOH threshold crossing.

Importantly, a longer warning time does not by itself imply a more favorable maintenance outcome. Earlier intervention provides additional planning time but can also remove a battery from service while usable life remains. The warning results therefore expose a trade-off between anticipatory maintenance and premature intervention rather than establishing a preferred policy. This trade-off is evaluated explicitly in the following section using censoring-aware timing loss and leakage-safe fixed-interval optimization.

3.5. Maintenance policy evaluation and optimization

Advance warning alone does not account for the trade-off between premature intervention and maintenance performed after the observed SOH threshold crossing. The prognosis-driven policies were therefore evaluated using the normalized timing loss in Eq. (4). At the 100-, 500-, and 1000-cycle prediction horizons, the point-based policy produced evaluable intervention times for six, seven, and eight batteries, respectively. Right-censored no-trigger cases were not assigned artificial intervention times or treated as observed maintenance failures. Although comparisons within each horizon used the same evaluable batteries for all policies, the battery composition differed across horizons. To ensure consistent comparisons across prediction horizons, the primary policy evaluation uses the same six commonly evaluable batteries (cells 1, 2, 3, 4, 7, and 8) at all three horizons.

Figure 5 presents the common-support policy comparison under the illustrative $\lambda_L = 5$ scenario. Using the same six batteries at every prediction horizon, the mean normalized timing losses of the point-based policy were 0.7167, 1.25, and 1.6833 at the 100-, 500-, and 1000-cycle horizons, respectively. The corresponding losses of the risk-aware policy were 1.55, 2.15, and 2.4333, while the fixed-interval policy yielded 1.7833 at each horizon. The fixed-interval result remained constant across horizons because its intervention schedule was independent of the prognosis horizon, and the same batteries were evaluated. The risk-aware policy incurred greater timing loss than the point-based policy at all three horizons, reflecting the cost of earlier intervention under the specified loss function. Thus, the additional warning time provided by uncertainty-aware triggering was accompanied by a greater penalty for unused battery life in this evaluation.

To examine the sensitivity to the evaluation-set composition, the original horizon-specific comparisons were also retained as a secondary analysis. These included six, seven, and eight evaluable batteries at the 100-, 500-, and 1000-cycle horizons, respectively. Under $\lambda_L = 5$, the corresponding mean timing losses were 0.7167, 1.0857, and 1.6625 for the point-based policy; 1.55, 2.0429, and 2.2250 for the risk-aware policy; and 1.7833, 1.8143, and 1.65 for the fixed-interval policy. Although the magnitude of the mean losses changed with battery composition, the risk-aware policy continued to exhibit higher timing loss than the point-based policy. The same ordering was also observed in the common-support analysis across all four predefined late-intervention penalty scenarios ($\lambda_L = 1$, 2, 5, and 10). These findings indicate that restricting the evaluation to a fixed set of batteries does not reverse the relative ordering of the point-based and risk-aware policies. Nevertheless, the common-support approach does not estimate counterfactual intervention outcomes for right-censored no-trigger cases; its purpose is to ensure consistent sample composition across prediction horizons.

To provide a scheduled-maintenance reference, the fixed intervention interval was optimized independently within each outer fold using only the seven development batteries. The selected interval was then applied unchanged to the held-out battery. Unlike the common-support policy comparison in Figure 5, the held-out evaluation of the optimized fixed-interval strategy included all eight batteries, because its intervention schedule did not depend on the availability of a prognosis-based trigger. The resulting development-selected intervals and held-out timing losses across the predefined late-intervention penalty scenarios are summarized in Table 6.

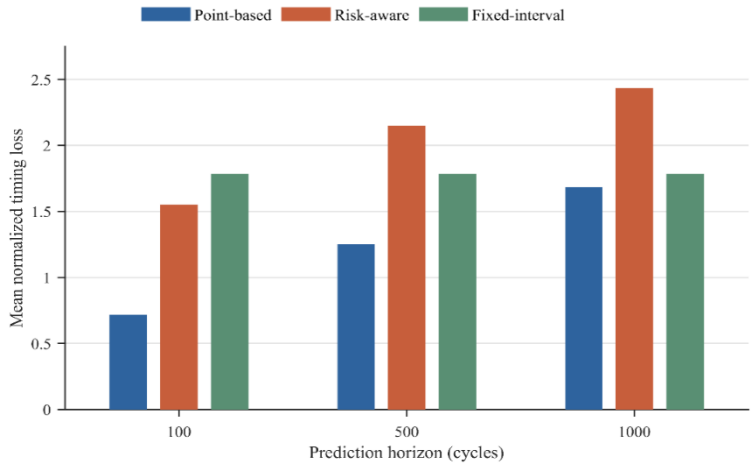


Figure 5. Mean normalized timing loss of the maintenance policies under the illustrative $\lambda_L = 5$ scenario across prediction horizons, evaluated on the same six commonly evaluable batteries (cells 1, 2, 3, 4, 7, and 8; n = 6 at each horizon).

Table 6. Development-only fixed-interval maintenance optimization across late-intervention penalty scenarios

λ_L	Development-selected interval (cycles)	Mean held-out timing loss	Median held-out timing loss
1	5000 in all folds	0.775	0.700
2	5000 in all folds	0.900	0.850
5	4500–5000 across folds	1.650	1.500
10	4500 in all folds	1.163	1.150

The development-selected interval shifted toward earlier intervention as greater weight was assigned to late maintenance. For $\lambda_L = 1$, however, the selected 5000-cycle interval coincided with the upper limit of the main search grid in all folds. An extended-grid sensitivity analysis showed that four of the eight folds shifted to 5500 cycles, indicating that this scenario was partially constrained by the original grid boundary. No corresponding boundary evidence was observed for the other penalty settings.

The policy results should be interpreted as scenario-based timing trade-offs rather than economic cost estimates. The Oxford dataset does not contain observed replacement, downtime, or monetary maintenance costs; consequently, λ_L represents a dimensionless relative penalty rather than a measured economic quantity. Overall, the analysis shows that maintenance decisions depend not only on prognostic accuracy but also on prediction horizon, uncertainty treatment, and the assumed relative consequences of premature and delayed intervention.

3.6. Joint optimization of maintenance decisions

To evaluate the contribution of uncertainty-aware maintenance decisions under a comparable optimization protocol, a separate experiment was conducted at the 500-cycle prediction horizon. Three policies were considered. The fixed-interval policy selected a scheduled intervention deadline by minimizing the mean development-set timing loss. The point-based policy jointly optimized a prognostic alarm threshold and a scheduled deadline, triggering intervention at the earlier of the point-prediction alarm and the deadline. The risk-aware policy used the same joint optimization procedure but defined its prognostic alarm using predictions adjusted by an age-dependent empirical calibration radius. For each held-out battery, policy parameters were selected exclusively from the seven development batteries, and performance was assessed using the asymmetric intervention-timing loss with a late-intervention penalty of 5. This experiment used a separately specified prediction model with fixed hyperparameters, rather than the tuned models used in the original multi-horizon comparison in Section 3.5. Battery-level results are reported in Table 7.

Table 7. Optimized maintenance parameters and battery-level intervention-timing losses at the 500-cycle prediction horizon (late-intervention penalty = 5)

Battery	Fixed d^*	Fixed loss	Point (θ^*, d^*)	Point loss	Risk-aware (θ^*, d^*)	Risk-aware loss
Cell 1	4500	1.30	(0.79, 5000)	0.80	(0.74, 5000)	0.80
Cell 2	4500	0.50	(0.80, 5000)	0.50	(0.74, 5000)	0.50
Cell 3	4500	1.60	(0.79, 5000)	1.10	(0.74, 5000)	1.10
Cell 4	5000	3.00	(0.79, 5000)	1.20	(0.73, 5000)	1.20
Cell 5	4500	0.50	(0.80, 6000)	5.00	(0.75, 5000)	0.00
Cell 6	5000	2.00	(0.78, 5000)	2.00	(0.72, 5000)	2.00
Cell 7	4500	2.90	(0.80, 5000)	2.40	(0.74, 5000)	2.40
Cell 8	4500	1.40	(0.80, 5000)	1.40	(0.74, 5000)	0.90
Mean	-	1.65	-	1.80	-	1.1125

The jointly optimized risk-aware policy achieved a mean intervention-timing loss of 1.1125, compared with 1.65 for the optimized fixed-interval policy and 1.80 for the optimized point-based policy, corresponding to reductions of 32.6% and 38.2%, respectively. However, the improvement over the point-based policy occurred only in Cell 5 and Cell 8; the two policies produced identical timing

losses for the remaining six batteries. In Cell 5, neither prognostic policy generated an alarm within the available prediction window; the difference arose from the selected maintenance deadlines of 6000 and 5000 cycles, respectively. Thus, the lower average loss reflects the combined effects of uncertainty-aware policy selection and the scheduled deadline, rather than uniformly earlier uncertainty–adjusted alarms. The observed advantage should therefore be interpreted as sample-specific.

Paired battery-level statistical analysis further characterized the performance of the optimized maintenance policies. The comparison between the risk-aware and fixed-interval policies yielded an unadjusted p-value of 0.03125, while the comparison with the point-based policy yielded $p = 0.5000$. After Holm correction for the two comparisons, the corresponding adjusted p-values were 0.0625 and 0.5000, respectively. Together with the battery-level results in Table 7, these statistical findings provide a more comprehensive assessment of maintenance policy performance and highlight the importance of evaluating uncertainty-aware optimization at both aggregate and individual battery levels.

To examine the contribution of prognostic alarm triggering beyond the scheduled intervention deadline, an additional retrospective comparison was conducted using the deadlines selected by the jointly optimized risk-aware policy. When the alarm mechanism was disabled and intervention was determined solely by these selected fixed deadlines, the mean timing loss was 1.2750, compared with 1.1125 for the complete risk-aware policy. The observed difference of 0.1625 reflects the reduction in aggregate timing loss associated with activating prognostic alarms while holding the jointly selected maintenance deadlines fixed. Because the deadlines were taken from the jointly optimized policy and were not reoptimized for the deadline-only comparator, this analysis evaluates the effect of activating prognostic alarms while keeping the selected deadlines fixed. It should therefore not be interpreted as a comparison of independently optimized policies.

A descriptive sensitivity analysis was performed to assess the influence of cell 5 on the aggregate results. Excluding this battery from the reported losses, without repeating policy optimization, increased the mean risk-aware loss to 1.2714, compared with 1.3429 for the point-based policy and 1.8143 for the fixed-interval policy. The relative reduction in mean loss for the risk-aware policy compared with the point-based policy consequently decreased from 38.2% to approximately 5.3%. This indicates that the observed average advantage over the point-based policy was strongly influenced by cell 5, although a smaller advantage remained among the other seven batteries. These results should be interpreted as a descriptive influence analysis rather than an independently reoptimized evaluation.

The comparison between predefined and jointly optimized maintenance policies provides an important decision-level insight. Incorporating prognostic uncertainty into a fixed triggering rule increased advance warning but also increased intervention-timing loss, demonstrating that earlier warnings and improved maintenance timing are distinct objectives. In contrast, the jointly optimized risk-aware policy achieved the lowest mean timing loss among the three optimized policies in the 500-cycle experiment. This finding highlights the potential value of incorporating uncertainty information within a jointly optimized maintenance decision framework, rather than using uncertainty solely to advance intervention triggers. The observed improvement was not uniform across batteries and reflects the combined effects of the selected alarm thresholds, maintenance deadlines, and the separately specified prognostic model. Accordingly, the results support uncertainty-aware decision optimization as a promising approach within the evaluated setting, while motivating further validation of its performance across broader battery populations.

3.7. Explainability analysis

The global permutation analysis showed that the prognostic models relied on both electrical health indicators across all prediction horizons. As shown in Figure 6, V_{range} produced the larger increase in RMSE when permuted at the 100- and 500-cycle horizons, whereas the permutation importance of V_{range} and V_{slope} became nearly equal at 1000 cycles. These values quantify model reliance under permutation and should not be interpreted as causal feature effects.

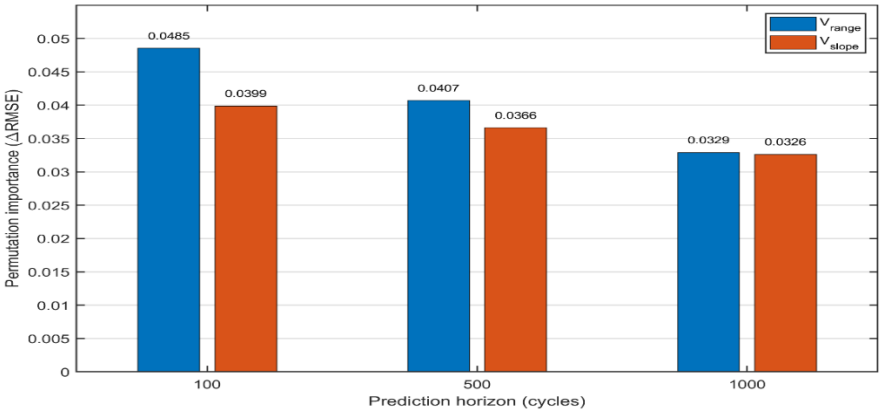


Figure 6. Global permutation importance of the electrical health indicators across prediction horizons.

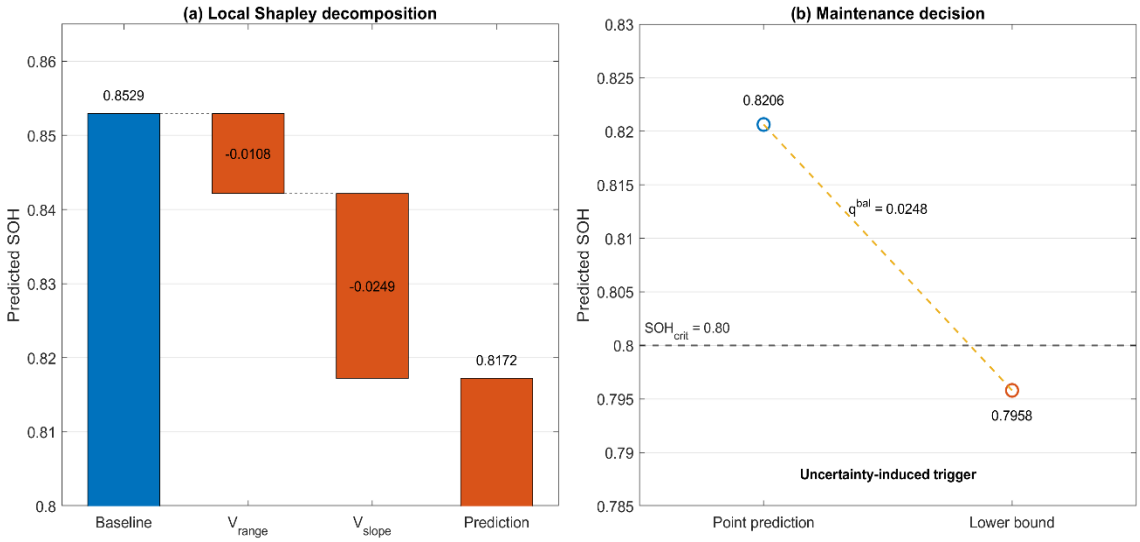


Figure 7. (a) Local Shapley decomposition; (b) uncertainty-induced maintenance decision for the representative 500-cycle prediction of cell 1.

A representative 500-cycle prediction for held-out cell 1 was further examined to connect feature-level attribution with the resulting maintenance decision. As illustrated in Figure 7(a), both V_{range} and V_{slope} reduced the reconstructed SOH prediction relative to the Shapley baseline, with V_{slope} providing the larger local contribution. The reconstructed prediction was 0.8172, while the corresponding frozen prediction used in the maintenance analysis was 0.8206. This small difference reflects the post hoc

reconstruction of the stochastic Bagged Trees model. The reconstructed models were used solely for feature attribution, whereas all reported prognostic, uncertainty, and maintenance results remained based on the frozen outer-fold predictions. Figure 7(b) shows how prognostic uncertainty altered the maintenance decision for the same case. Although the frozen point prediction of 0.8206 remained above the 0.80 SOH threshold, incorporating the battery-balanced uncertainty penalty reduced the lower prediction bound to 0.7958. The risk-aware policy was therefore triggered at this characterization even though the point-based policy was not, resulting in an uncertainty-induced trigger. This example illustrates the complete decision pathway from degradation-sensitive voltage indicators through SOH prognosis and uncertainty quantification to the resulting maintenance action.

4. Conclusions

This study developed an explainable, risk-aware predictive-maintenance framework that connects multi-horizon battery health prognosis with uncertainty-aware intervention decisions. Under strict leave-one-battery-out evaluation, Bagged Trees achieved pooled SOH prediction RMSE values of 0.0194, 0.0201, and 0.0211 at exact 100-, 500-, and 1000-cycle horizons, respectively. Comparison with the privileged persistence baseline showed that persistence remained a strong short-horizon reference, whereas the learned model offered greater predictive benefits at the longest horizon.

Battery-balanced empirical uncertainty bounds achieved mean held-out-cell coverage rates of 91.5%, 91.8%, and 90.4% at the 100-, 500-, and 1000-cycle prediction horizons, respectively, against a nominal 90% target. However, coverage for cell 2 was only 65.5% at the 1000-cycle horizon, corresponding to a 24.5-percentage-point shortfall and demonstrating that satisfactory average calibration may conceal substantial battery-specific uncertainty failures. When these uncertainty bounds were incorporated into maintenance decisions, the predefined risk-aware rule provided median advance warnings of 1450, 2050, and 2400 aging cycles at the three prediction horizons. Nevertheless, this earlier warning did not consistently improve intervention timing: under the common-support evaluation with a late-intervention penalty of 5, the risk-aware policy incurred higher mean timing losses than the point-based policy at all three horizons, illustrating the trade-off between conservative triggering and unused battery life. Leakage-safe fixed-interval optimization further showed that maintenance timing depends on the assumed relative penalty assigned to early and delayed intervention.

The joint maintenance optimization experiment illustrated the potential benefits of incorporating prognostic information into an explicit decision optimization framework. At the 500-cycle prediction horizon, the optimized risk-aware policy achieved a mean normalized timing loss of 1.1125, compared with 1.8000 for the optimized point-based policy and 1.6500 for the optimized fixed-interval policy, corresponding to relative reductions of 38.2% and 32.6%, respectively. This result highlights the potential of jointly selecting prognostic alarm thresholds and scheduled maintenance deadlines under a common asymmetric timing-loss objective, rather than relying solely on predefined intervention rules. The comparison also revealed that the observed performance advantage depended on the selected decision parameters and varied across individual batteries. Specifically, the improvement over the optimized point-based policy was confined to cells 5 and 8. In a descriptive sensitivity analysis excluding cell 5, without repeating policy optimization, the relative reduction in mean timing loss decreased from 38.2% to approximately 5.3%. The benefit observed for cell 5 was associated with the selected maintenance deadline rather than an earlier uncertainty–induced alarm. Furthermore, activating prognostic alarms while keeping the optimized risk-aware deadlines fixed reduced the mean

timing loss from 1.2750 to 1.1125. However, this conditional comparison does not establish superiority over an independently reoptimized deadline-only policy. Because the predefined and jointly optimized evaluations used different prognostic model configurations, their performance differences should not be attributed exclusively to decision optimization. Together, these findings provide a leakage-safe framework for examining how prognostic uncertainty and maintenance decision parameters influence intervention outcomes.

The explainability analysis connected this decision process to the underlying voltage-based health indicators. Global permutation importance showed that the prognostic models relied on both voltage range and normalized voltage slope, while local Shapley and decision-level analyses illustrated how feature-driven SOH prognosis and prognostic uncertainty jointly influenced the resulting maintenance trigger.

The study is limited to eight laboratory-aged cells from a single dataset, and the leave-one-battery-out validation assesses performance only across held-out cells within that dataset. Independent external validation was not performed; consequently, the observed prognostic and maintenance-decision performance cannot be assumed to transfer to other battery chemistries, operating conditions, or experimental protocols. Furthermore, the dataset does not contain observed maintenance actions, downtime, or economic cost records. Consequently, the maintenance analysis represents normalized scenario-based decision evaluation rather than direct lifecycle-cost optimization. The joint maintenance optimization experiment was limited to eight batteries, and its aggregate advantage over the optimized point-based policy was sensitive to the outcome for an individual battery. Furthermore, the age-dependent empirical calibration procedure does not provide a distribution-free coverage guarantee under battery-level distribution shifts. Consequently, the observed decision benefits should be regarded as evidence within the evaluated experimental setting rather than as established improvements in operational maintenance performance.

Future work should prioritize independent external validation using datasets with compatible discharge measurements, sufficiently long degradation trajectories, and observable SOH threshold crossings. Such validation should examine both prognostic transferability and the stability of maintenance-policy comparisons across different cells, chemistries, and operating conditions. Incorporating observed maintenance actions and operational cost data would further support deployment-oriented evaluation.

Data Availability

The Oxford Battery Degradation Dataset used in this study is publicly available at the source cited in Reference [38]. The MATLAB codes developed for the analyses presented in this study are available from the corresponding author upon reasonable request.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgements

The author has no acknowledgements to declare.

Use of Generative-AI tools declaration

During the preparation of this work, the author used Gemini in order to improve language, readability, and grammar. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Conflict of interest

The author declares no conflict of interest.

References

1. Y. Wang, S. Guo, Y. Cui, L. Deng, L. Zhao, J. Li, et al., A comprehensive review of machine learning-based state of health estimation for lithium-ion batteries: Data, features, algorithms, and future challenges, *Renew. and Sustain. Energy Rev.*, **224** (2025), 116125. https://doi.org/10.1016/j.rser.2025.116125
2. A. Thelen, X. Huan, N. Paulson, S. Onori, Z. Hu, C. Hu, Probabilistic machine learning for battery health diagnostics and prognostics—review and perspectives, *npj Mater. Sustain.*, **2** (2024), 14. https://doi.org/10.1038/s44296-024-00011-1
3. Y. He, Y. Ye, Q. Zeng, F. Liu, M. Cheng, H. Zhang, Defect detection of 18650 lithium-ion battery casing based on current-pulse-excited active acoustic emission, *IEEE Sens. J.,* 2026. https://doi.org/10.1109/JSEN.2026.3740910
4. M. Es-satte, I. Afaynou, H. Faraji, K. Choukairy, I. Regragui, K. Amghar, et al., Phase change materials, bionic structures and artificial intelligence for lithium-ion battery thermal management: A comprehensive review, *Int. Commun. Heat Mass Transf.,* **173** (2026), 110870. https://doi.org/10.1016/j.icheatmasstransfer.2026.110870
5. A. Siricharoenpanich, H. Faraji, P. Naphon, Self-recovery multi-stage battery system for enhanced energy utilization via electrochemical recovery and adaptive scheduling, *J. Energy Storage*, **176** (2026), 123515. https://doi.org/10.1016/j.est.2026.123515
6. Y. Che, X. Hu, X. Lin, J. Guo, R. Teodorescu, Health prognostics for lithium-ion batteries: mechanisms, methods, and prospects, *Energy Environ. Sci.,* **16** (2023), 338–371. https://doi.org/10.1039/D2EE03019E
7. S. Khaleghi, M. S. Hosen, J. Van Mierlo, M. Berecibar, Towards machine-learning driven prognostics and health management of Li-ion batteries. A comprehensive review, *Renew. Sustain. Energy Rev.*, **192** (2024), 114224. https://doi.org/10.1016/j.rser.2023.114224
8. G. Li, B. Li, C. Li, S. Wang, State-of-health rapid estimation for lithium-ion battery based on an interpretable stacking ensemble model with short-term voltage profiles, *Energy*, **263** (2023), 126064. https://doi.org/10.1016/j.energy.2022.126064
9. Y. Zhang, Z. Li, L. Kong, H. Xu, H. Shen, M. Chen, Multi-step state of health prediction of lithium-ion batteries based on multi-feature extraction and improved Transformer, *J. Energy Storage*, **105** (2025), 114538. https://doi.org/10.1016/j.est.2024.114538
10. D. Wang, C. Li, Z. Wei, X. Chen, Y. Jiao, A multi-forward-step lithium-ion battery SOH prediction method based on segmented gated recurrent unit and DualMamba, *J. Energy Storage*, **117** (2025), 116136. https://doi.org/10.1016/j.est.2025.116136

11. R. Mao, P. Han, W. Xu, N. Yuan, H. Zhang, Lithium-ion battery remaining useful life prediction using Bayesian filters with neural network pseudo observations, *J. Energy Storage,* 2026. https://doi.org/10.1016/j.est.2025.120179
12. K. Huang, K. Yao, Y. Guo, Z. Lv, State of health estimation of lithium-ion batteries based on fine-tuning or rebuilding transfer learning strategies combined with new features mining, *Energy*, **282** (2023), 128739. https://doi.org/10.1016/j.energy.2023.128739
13. X. Chen, Y. Qin, W. Zhao, Q. Yang, N. Cai, K. Wu, A self-attention knowledge domain adaptation network for commercial lithium-ion batteries state-of-health estimation under shallow cycles, *J. Energy Storage*, **86** (2024), 111197. https://doi.org/10.1016/j.est.2024.111197
14. L. Chen, X. Bao, A. M. Lopes, X. Li, H. Kong, Y. Chai, et al., Domain generalization-based state-of-health estimation of lithium-ion batteries, *J. Power Sources*, **610** (2024), 234696. https://doi.org/10.1016/j.jpowsour.2024.234696
15. B. Dou, S. Hou, H. Li, Y. Zhao, Y. Fan, L. Sun, et al., Cross-domain state of health estimation for lithium-ion battery based on latent space consistency using few-unlabeled data, *Energy*, **320** (2025), 135257. https://doi.org/10.1016/j.energy.2025.135257
16. D. Yu, J. Su, Y. Du, X. Qu, Cross-domain state-of-health estimation for lithium-ion batteries: A deep learning and similarity network fusion framework with transfer learning, *J. Energy Storage*, **143** (2026), 119706. https://doi.org/10.1016/j.est.2025.119706
17. N. Ahlawat, R. Pandit, A standardized comparative framework for machine learning techniques in lithium-ion battery state of health estimation, *Future Batter.*, **7** (2025), 100099. https://doi.org/10.1016/j.fub.2025.100099
18. E. S. A. Mousavi, F. Abdollahi, F. Barazandeh, F. Torabi, A. Hajizadeh, History-agnostic lithium-ion batteries state of health and remaining useful life estimation using fusion-based Bayesian neural network, *IET Renew. Power Gener.,* **20** (2026), e70297. https://doi.org/10.1049/rpg2.70297
19. S. Rizanov, A. Stoynova, G. Mihov, Toward trustworthy and transferable SOH/RUL estimation for lithium-ion batteries: A critical review and multi-fidelity validation framework from laboratory cells to real-world packs, *Batteries,* **12** (2026), 255. https://doi.org/10.3390/batteries12070255
20. Y. Ren, L. Zhang, P. Shi, Adaptive cooperative intra–inter domain adversarial network with stage division for RUL prediction of ship electric propulsion system, *IEEE Trans. Ind. Electron.,* 2026. https://doi.org/10.1109/TIE.2026.3672810
21. J. Hu, L. Wu, Early uncertainty quantification prediction of lithium-ion battery remaining useful life with transformer ensemble model, *IEEE Trans. Transp. Electrif.*, **10** (2024), 8618–8629. https://doi.org/10.1109/TTE.2024.3354302
22. Z. Song, H. Zhang, J. Jia, Data-Driven State of Health Interval Prediction for Lithium-Ion Batteries, *Electronics*, **13** (2024), 3991. https://doi.org/10.3390/electronics13203991
23. S. Wen, N. Lin, S. Huang, X. Li, Z. Wang, Z. Zhang, Lithium battery state of health estimation using real-world vehicle data and an interpretable hybrid framework, *J. Energy Storage,* **96** (2024), 112623. https://doi.org/10.1016/j.est.2024.112623
24. M. J. M. A. Rasul, A. Abbas, J. Baek, J. Kim, A hybrid ensemble learning framework with uncertainty quantification for state-of-health estimation in lithium-ion batteries, *Measurement,* 2026. https://doi.org/10.1016/j.measurement.2026.120528

25. Z. Zhao, M. Feng, Q. Yu, H. Yuan, F. Wang, Z. Zhai, Towards trustworthy lithium-ion battery remaining useful life prediction with uncertainty quantification, *J. Energy Storage,* **167** (2026), 122408. https://doi.org/10.1016/j.est.2026.122408
26. L. Basora, A. Viens, M. Arias Chao, X. Olive, A benchmark on uncertainty quantification for deep learning prognostics, *Reliab. Eng. Syst. Saf.,* **253** (2025), 110513. https://doi.org/10.1016/j.ress.2024.110513
27. F. Wang, Z. Zhao, Z. Zhai, Z. Shang, R. Yan, X. Chen, Explainability-driven model improvement for SOH estimation of lithium-ion battery, *Reliab. Eng. Syst. Saf.,* **232** (2023), 109046. https://doi.org/10.1016/j.ress.2022.109046
28. T. Tetik, Feature engineering and explainable artificial intelligence for state of health estimation of Lithium-ion batteries, *J. Energy Storage*, **144** (2026), 119873. https://doi.org/10.1016/j.est.2025.119873
29. M. Zhen, Z. Chen, B. Lu, Z. Chen, E. Pan, Net benefit-oriented condition-based maintenance for lithium-ion battery packs in SGLS systems: Combining degradation updating and decision-making, *Comput. Ind. Eng.*, **200** (2025), 110850. https://doi.org/10.1016/j.cie.2024.110850
30. Z. Wang, W. Shangguan, Z. Xu, C. Peng, E. Zio, B. Cai, A multi-objective optimization model for predictive opportunistic maintenance of lithium-ion batteries, *J. Energy Storage*, **13**2 (2025), 117902. https://doi.org/10.1016/j.est.2025.117902
31. M. Valizadeh, M. Servati, H. Ataei, F. Hamrang, From impedance and cycling physics to economic value: A novel multi-domain state-of-health–Markov framework for lithium-ion battery maintenance, *J. Energy Storage*, **169** (2026), 122642. https://doi.org/10.1016/j.est.2026.122642
32. G. Patrizi, L. Martiri, A. Pievatolo, A. Magrini, G. Meccariello, L. Cristaldi, et al., A review of degradation models and remaining useful life prediction for testing design and predictive maintenance of lithium-ion batteries, *Sensors,* **24** (2024), 3382. https://doi.org/10.3390/s24113382
33. Y. Che, Y. Zheng, F. E. Forest, X. Sui, X. Hu, R. Teodorescu, Predictive health assessment for lithium-ion batteries with probabilistic degradation prediction and accelerating aging detection, *Reliab. Eng. Syst. Saf.*, **241** (2024), 109603. https://doi.org/10.1016/j.ress.2023.109603
34. J. Chen, S. Jin, Y. Cui, C. Wei, K. Polat, F. Alenezi, Entropy-informed deep residual network for nonlinear EEG dynamics in fine-grained driver state recognition, *Appl. Soft Comput.,* **201** (2026), 115532. https://doi.org/10.1016/j.asoc.2026.115532
35. J. Chen, G. Yuan, C. Wei, Characterizing fatigue-related neural signatures via multi-band EEG features and a three-stage feature optimization strategy for driver fatigue detection, *Meas. Sci. Technol.,* **37** (2026), 306108. https://doi.org/10.1088/1361-6501/ae8d4b
36. J. Chen, Y. Zhao, C. Wei, Z. Qi, H. Wang, K. Polat, et al., Enhancing unseen driver state detection: An EEG-based framework with brain connectivity and depthwise separable convolutional neural networks, *Biomed. Signal Process. Control.,* **124** (2026), 110676. https://doi.org/10.1016/j.bspc.2026.110676
37. J. Chen, W. Zhao, Y. Cui, C. Wei, K. Polat, F. Alenezi, A review of EEG-based driver fatigue detection: Nonlinear dynamics, brain networks, and deep learning advances, *WIREs Data Min. Knowl. Discov.,* **16** (2026), e70099. https://doi.org/10.1002/widm.70099
38. D. Howey, C. Birkl, *Oxford battery degradation dataset 1*, University of Oxford, 2017. Available from: https://oxfordbatterydegdata.sweb.ox.ac.uk/.

39. B. Ali, I. Matloob, N. Alkhalifa, A. K. Alkhalifa, N. Asim, S. Kanwal, et al., Deep learning-based RUL and SOH prediction of lithium-ion batteries using LOCO, *Sci. Rep.*, **16** (2026), 27853. https://doi.org/10.1038/s41598-026-58379-0
40. K. Wu, R. Gan, J. Zhu, M. He, Uncertainty-aware transformer for lithium-ion battery state-of-health estimation, *Adv. Transdiscip. Eng.,* 2026. https://doi.org/10.3233/ATDE260884



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)