A Data-Centric Framework for OOD-Aware RUL Prediction of Lithium-Ion Batteries with Risk-Aware Evaluation

Yara Rodrigues Inácio[1], **Diego de Freitas Bezerra**[2]

[1]Technology Innovation – Accenture
Recife – PE – Brazil

[2]Recife Center for Advanced Studies and Systems – C.E.S.A.R
Recife – PE – Brazil

yara.inacio@accenture.com, dfb@cesar.org.br

Abstract. *Predicting the Remaining Useful Life (RUL) of lithium-ion batteries is a safety-critical machine learning task. A central obstacle is* domain shift*: models trained in one experimental context frequently fail to generalise to others, producing overoptimistic predictions that translate into direct operational risk, a failure mode that aggregate metrics such as mean absolute error (MAE) and root mean square error (RMSE) systematically conceal. We propose a data-centric framework that addresses this challenge through three contributions. First, we construct a* ***multi-domain OOD evaluation benchmark*** *integrating six heterogeneous public datasets spanning calendar ageing, industrial, and aeronautical degradation profiles under a unified schema and standardised evaluation protocols. Second, we introduce a* ***physics-informed inductive bias*** *via the* Shadow Impedance *proxy, an estimate of internal resistance derived directly from raw voltage and current signals, which enriches the feature space and acts as a routing signal in a* ***Regime-Aware Ensemble*** *of specialised XGBoost regressors following the Mixture-of-Experts principle. Third, we propose a* ***risk-aware evaluation framework*** *centred on the* Danger Bias *metric, which quantifies the fraction of operationally dangerous overestimates of remaining life. Evaluation under LOCO and LODO protocols shows strong intra-distribution performance (R^2=0.9662, MAE $< 5\%$), but reveals critical cross-domain degradation, with Danger Bias reaching 88.8% in EVERLASTING, a two-cell calendar-aging dataset, the most severe OOD scenario. These findings demonstrate that reliable AI for safety-critical physical domains requires not only expressive models, but also OOD-aware benchmarks and risk-sensitive evaluation metrics.*

1. Introduction

Lithium-ion batteries are the dominant energy-storage technology in electric vehicles and stationary systems [Zhai et al. 2024]. As cells age and their capacity falls below operational thresholds, a growing practice is to redirect them to less demanding *second-life* applications, such as grid storage and emergency power supply [Azad et al. 2026]. This reuse scenario imposes stringent requirements on predictive models: they must generalise reliably to cells with heterogeneous histories, chemistries, and usage profiles, conditions that systematically violate the i.i.d. assumptions underlying most learning algorithms.

In this context, *Remaining Useful Life* (RUL) prediction, the number of cycles remaining until a battery reaches its operational end-of-life, emerges as a central problem

for enabling safe and economically viable reuse strategies. Degradation is commonly quantified via the State of Health (SOH), defined as the ratio of current capacity to nominal capacity. RUL is then formalised as the cycles remaining until SOH reaches a predefined disposal threshold (typically 80% of nominal capacity).

Machine learning has become central to intelligent systems in physical-domain applications; however, its use in safety-critical contexts exposes a fundamental tension: models that perform well under controlled evaluation conditions frequently fail when confronted with data from distinct operational contexts. This phenomenon, known as *domain shift* or out-of-distribution (OOD) generalisation failure [Kouw and Loog 2021], is particularly critical in battery prognostics, since an overoptimistic prediction of remaining life can result in unexpected failures.

The battery prognostics literature has responded primarily by proposing increasingly complex model architectures [Lu et al. 2023, Zhang et al. 2025], while OOD generalisation has received comparatively little systematic attention. Two structural gaps persist: (i) evaluations are typically confined to a single dataset, and (ii) standard regression metrics (MAE, RMSE) aggregate errors symmetrically, masking the asymmetric risk structure of RUL prediction, where overestimating remaining life is operationally dangerous while underestimating it is merely conservative [Sambasivan et al. 2021].

This paper proposes a data-centric framework that addresses these gaps through three contributions:

- A **multi-domain OOD benchmark** unifying six heterogeneous public datasets under a canonical schema and rigorous evaluation protocols (LOCO and LODO);
- A **Regime-Aware Ensemble** that uses the *Shadow Impedance* as a physics-informed inductive bias and routing signal, following the Mixture-of-Experts principle [Jacobs et al. 1991];
- A **risk-aware evaluation framework** centred on the *Danger Bias* metric, which quantifies the fraction of operationally dangerous overestimates.

The body of this paper is organized as follows: Section 2 overviews the related work on machine learning approaches for battery State of Health (SOH) and Remaining Useful Life (RUL) estimation; Section 3 presents the data and methodology, detailing the multi-domain out-of-distribution (OOD) benchmark, physics-informed feature engineering via Shadow Impedance, the Regime-Aware Ensemble model, and the risk-aware evaluation framework; Section 4 describes the experimental setup, presents the results for both intra-domain and OOD cross-domain scenarios, and discusses the observed failure modes and theoretical implications; and finally Section 5 concludes our work and delineates some future research directions.

2. Related Work

Deep learning approaches have shown promising results for SOH and RUL estimation. Li et al. [Li and Chen 2025] propose hybrid models optimised via Particle Swarm Optimisation (PSO), demonstrating superior performance of LSTM architectures for modelling degradation time series. Lu et al. [Lu et al. 2023] employ deep learning combined with domain adaptation techniques to estimate SOH without new characterisation experiments. Zhang et al. [Zhang et al. 2025] propose BatLiNet, which exploits intra- and inter-cell relationships to improve generalisation under diverse ageing conditions. Mayemba et

al. [Mayemba et al. 2024] present a systematic comparison of machine learning approaches for capacity-fade modelling, finding that neural networks outperform empirical models in non-linear regimes at higher computational cost.

Severson et al. [Severson et al. 2019] demonstrated that battery lifetime can be estimated with high accuracy from statistics extracted from the first 100 cycles, including variations in discharge curves and incremental capacity. Celik et al. [Celik et al. 2022] extend this approach, combining machine learning with data management strategies to predict cycle life under varying charge conditions. Jiang and Song [Jiang and Song 2023] exploit discharge voltage curves as predictive feature sources, capturing the electrochemical signature of degradation directly from operational signals. Tan et al. [Tan et al. 2022] introduce sliding-window expansion to increase training set diversity from a reduced number of initial cycles.

Ni et al. [Ni et al. 2023] propose a hybrid model integrating features from charge curves with Support Vector Regression (SVR), obtaining good results even with limited training data. Takahashi et al. [Takahashi et al. 2023] apply Gaussian process regression to assess battery reuse feasibility, focusing on uncertainty quantification. Valizadeh and Amirhosseini [Valizadeh and Amirhosseini 2024] review ML applications in lithium-ion batteries, identifying the lack of standardised benchmarks and dataset heterogeneity as central obstacles.

The reviewed literature presents three recurrent limitations that motivate this work. First, most studies evaluate models within a single dataset, precluding systematic analysis of generalisation under domain shift. Second, widely used metrics (MAE, RMSE) do not explicitly distinguish errors with different operational implications. Third, the exploitation of electrochemical domain knowledge as inductive bias in scalable tabular models for RUL prediction remains limited.

3. Data and Method

The proposed framework comprises four complementary components: a Multi-domain OOD Benchmark, a physics-informed Feature Engineering strategy, a Regime-Aware Ensemble, and a Risk-aware Evaluation Framework.

3.1. Multi-Domain OOD Benchmark

Six datasets covering distinct degradation regimes were unified via schema alignment [Rahm and Bernstein 2001], mapping heterogeneous naming conventions onto five canonical physical variables: voltage (V), current (I), capacity (Q), temperature (T), and time (t). When capacity was not directly available, it was estimated via numerical integration of current:

$$Q_{\mathrm{Ah}}(t) \approx \frac{1}{3600} \sum_{k} I_k \, \Delta t_k \tag{1}$$

State of Health was then computed as:

$$\mathrm{SOH}(t) = \frac{Q_{\mathrm{Ah}}(t)}{Q_{\mathrm{Ah}}^{\mathrm{nom}}} \tag{2}$$

Data-leakage risk was mitigated structurally by adopting two complementary protocols. **LOCO** (Leave-One-Cell-Out) evaluates intra-domain generalisation by holding

out individual cells for testing. **LODO** (Leave-One-Dataset-Out) evaluates cross-domain generalisation by holding out entire datasets. In both protocols, train/test splits are strictly based on physical cell identity, ensuring that future cycles of a test cell are never visible during training. All normalisation statistics were computed exclusively on the training set.

Table 1 summarises the six datasets integrated in the benchmark, totalling approximately 366 unique cells spanning NMC, LFP, NCA/Si-Gr and $LiCoO_2$ chemistries.

Table 1. Multi-domain benchmark: characteristics of integrated datasets.

Dataset	Chemistry	Application	Degradation Profile	Cells
NASA [Fricke et al. 2023]	$LiCoO_2$	Generic	Constant CC/CV cycling	52
Oxford [Howey and Birkl 2017]	NMC	Laboratory	Constant current	8
Forklift [Vilsen and Stroe 2023]	LFP	Industrial	Dynamic forklift	3
EVERLASTING [Trad 2021]	NCA/Si-Gr	Laboratory	Calendar + mixed	2
eVTOL [Bills et al. 2021]	NMC	Aeronautical	Extreme C-rate	22
Multi-Stage [Stroebl et al. 2024]	NMC	Laboratory	Multi-stage (diverse)	279
Total				**≈366**

To ensure strict data hygiene and alignment across these heterogeneous sources, a standardized protocol for data inclusion and exclusion was enforced. The benchmark deliberately excludes structural metadata artifacts and secondary electrochemistry experiments, such as brief electrochemical impedance spectroscopy (EIS) measurements, which do not contribute directly to continuous cycling degradation trajectories. This data-centric curation process ensures that the joint feature space is constructed exclusively from unbroken, continuous time-series operational streams across all domains, protecting the downstream ensemble from synthetic optimization distortions.

Before applying the OOD evaluation protocols, the data were divided into a development set and an independent holdout test set. The holdout set was kept completely isolated from model development and was not used for hyperparameter selection, feature engineering decisions, or routing-threshold definition. LOCO and LODO evaluations were subsequently performed exclusively on the development set.

3.2. Feature Engineering and Shadow Impedance

The feature vector is grounded in electrochemical domain knowledge, comprising: normalised SOH, estimated internal resistance (Shadow Impedance), nominal capacity, maximum observed voltage, and SOH degradation rate (SOH_diff).

Among the indicators associated with battery ageing, the increase in internal resistance is one of the most consistently observed phenomena. Since direct impedance measurements are not consistently available across the analysed datasets, we define the *Shadow Impedance* proxy (Z_{shadow}) to capture this behaviour from available operational variables [Waag et al. 2014]:

$$Z_{\text{shadow}} = \frac{\Delta V}{\Delta I} \tag{3}$$

The variations ΔV and ΔI were calculated over a fixed temporal measurement window of 1 second during the pulse phases.

An Exponential Moving Average (EMA) filter is applied to voltage and current signals prior to extraction to mitigate sensor noise, ensuring the proxy predominantly captures ohmic resistance.

Z_{shadow} serves a dual role: it enriches the feature space with a physically meaningful degradation signal, and acts as the routing variable in the Regime-Aware Ensemble described next.

3.3. Regime-Aware Ensemble

The prediction target, *Fractional RUL* (RUL_{frac}), is defined as:

$$\text{RUL}_{\text{frac}} = \frac{\text{Cycle}_{\text{EOL}} - \text{Cycle}_{\text{current}}}{\text{Cycle}_{\text{EOL}}} \tag{4}$$

where $\text{Cycle}_{\text{EOL}}$ is the cycle at which SOH reaches the first-life disposal threshold (typically 80% of nominal capacity). Values are normalised to $[0, 1]$, enabling comparison across cells with different total lifetimes.

XGBoost [Chen and Guestrin 2016] was adopted as the base regressor due to its established performance on tabular regression problems. The routing function $r(Z_{\text{shadow}})$ distributes each sample to one of K=3 specialised regressors $\{E_1, E_2, E_3\}$:

$$r(Z_{\text{shadow}}) = \begin{cases} 1 & \text{if } Z_{\text{shadow}} < 0.06\,\Omega \quad \text{(high-performance)} \\ 2 & \text{if } 0.06 \leq Z_{\text{shadow}} < 0.12\,\Omega \quad \text{(industrial)} \\ 3 & \text{if } Z_{\text{shadow}} \geq 0.12\,\Omega \quad \text{(stationary)} \end{cases} \tag{5}$$

Routing thresholds were derived empirically via quantile analysis of the Z_{shadow} distribution on the training set. The final prediction is:

$$\hat{y}(\mathbf{x}) = E_{r(Z_{\text{shadow}})}(\mathbf{x}) \tag{6}$$

Hyperparameters (n_estimators=80, max_depth=4, learning_rate=0.05) were selected via grid search on NASA and Multi-Stage partitions and held fixed across all LODO and LOCO evaluations to prevent post-hoc fitting bias.

Although the routing thresholds in Equation 5 were empirically derived by quantitative analysis of the training set, this hard-threshold approach represents a deterministic baseline. In operational deployment, fixed boundaries can induce a variance collapse or regime saturation, as observed in the eVTOL and Oxford scenarios. To address this, transitioning from deterministic $if/else$ logic to a learned probabilistic gating function, where a soft routing signal weights the contribution of each expert, is a critical next step to enhance out-of-distribution robustness.

3.4. Risk-Aware Evaluation Framework

Standard regression metrics (MAE, RMSE, R^2) aggregate errors symmetrically, obscuring the asymmetric risk structure of RUL prediction. We formalise this asymmetry through

three complementary indicators defined over a set of predictions $\{(\hat{y}_i, y_i)\}_{i=1}^{N}$ with tolerance $\tau \geq 0$:

$$\text{SafeBias} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i < y_i - \tau] \tag{7}$$

$$\text{Neutral} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[|\hat{y}_i - y_i| \leq \tau] \tag{8}$$

$$\text{DangerBias} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i > y_i + \tau] \tag{9}$$

By construction, $\text{SafeBias} + \text{Neutral} + \text{DangerBias} = 1$. ***Danger Bias*** directly quantifies the fraction of predictions that would lead an operator to defer maintenance beyond the true end-of-life.

In this work, $\tau = 0$ is adopted, representing the most conservative theoretical limit where any overestimation of remaining life is penalised. To enable exact matching when evaluating the Neutral metric at $\tau = 0$, both the predicted and true Remaining Useful Life (RUL) values were discretized by rounding to the nearest integer cycle prior to calculation.

The choice of τ=0 reflects the operational asymmetry inherent to second-life battery reuse: in this context, any overestimate of remaining useful life may result in a cell being deployed beyond its safe operational boundary, whereas underestimation carries only an economic cost. Adopting τ=0 thus represents the risk posture appropriate for safety-critical deployment, and provides an upper bound on Danger Bias against which less conservative thresholds can be benchmarked.

Evaluation is conducted under three protocols: LODO for cross-domain OOD generalisation; LOCO for intra-domain cell-level variability; and Temporal Split for the end-of-life regime, defined by drawing a chronological boundary across the dataset, where the initial 50% of recorded cycles for each cell were assigned to the training set, and the remaining 50% were held out for testing to evaluate forecasting robustness over time.

4. Experiments and Results

4.1. Intra-Domain Performance and Ablation Study

On the holdout test set, the Regime-Aware Ensemble achieves R^2=0.9662, with predictions strongly aligned with the ideal diagonal (Figure 1). The error distribution is concentrated near zero, with mild asymmetric tails reflecting residual Danger Bias even within the training distribution (Figure 2).

Table 2 presents the ablation study, quantifying the marginal contribution of each feature (Figure 3). Removal of Shadow Impedance causes the largest performance drop (R^2: 0.9662 $\rightarrow$ 0.9124), confirming its vital dual role as both a predictive feature and a routing variable. Conversely, removing SOH yields a highly marginal drop (R^2: 0.9662 $\rightarrow$ 0.9632), while removing SOH_diff and max_voltage produces negligible effects ($\approx$0.965).

It is worth noting that the ablation study (Figure 3) evaluates the relative impact of the feature sets against the global baseline ($R^2 \approx 0.85$). The superior performance of

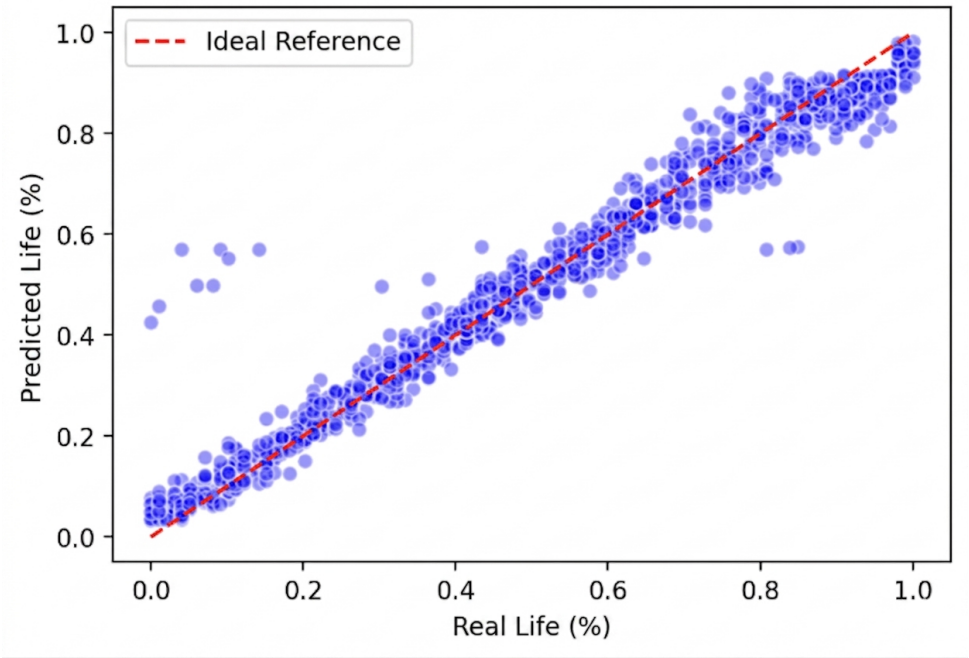


Figure 1. Predicted vs. actual RUL on the independent holdout test set. The dashed line represents the ideal prediction.

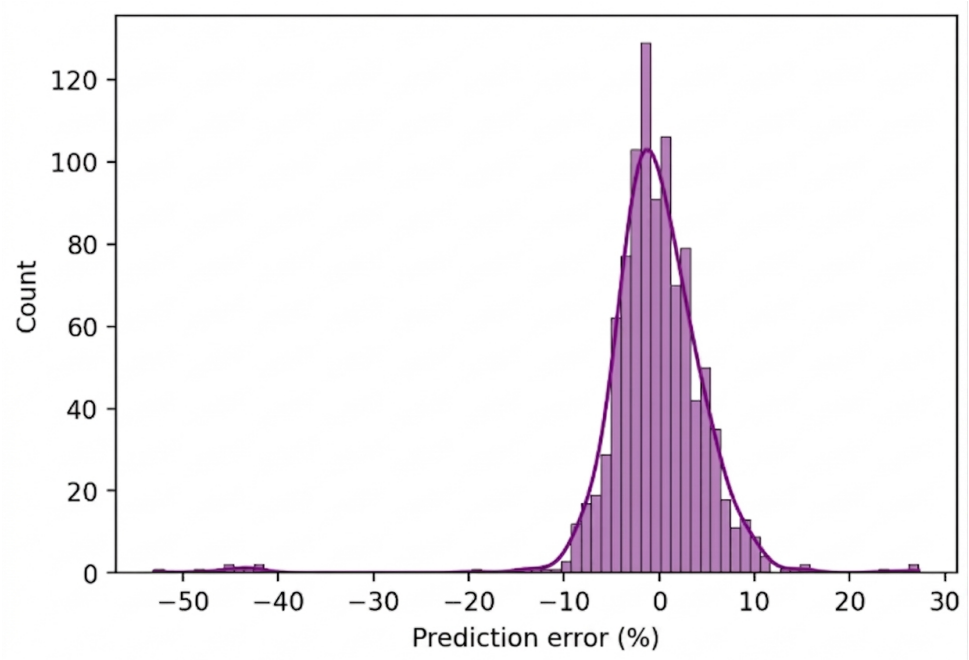


Figure 2. Prediction error distribution on the independent holdout test set. The left tail (negative errors) corresponds to safe underestimates; the right tail to dangerous overestimates captured by Danger Bias (τ=0).

$R^2 = 0.9662$ (Table 2) is achieved exclusively when the Regime-Aware Ensemble operates by combining domain experts across operational regimes.

This finding is highly consequential for real-world deployment: it demonstrates that the framework can be stripped of explicit SOH inputs without sacrificing predictive accuracy. By eliminating the operational dependency on real time SOH estimators, the model can successfully rely solely on non-invasive, raw voltage and current signals captured directly in the field.

Table 2. Ablation study: impact of individual feature removal on intra-domain R^2.

Feature Removed	R^2
None (full model)	0.9662
Shadow Impedance (internal resistance)	0.9124
SOH	0.9632
SOH_diff	0.965
max_voltage	0.965

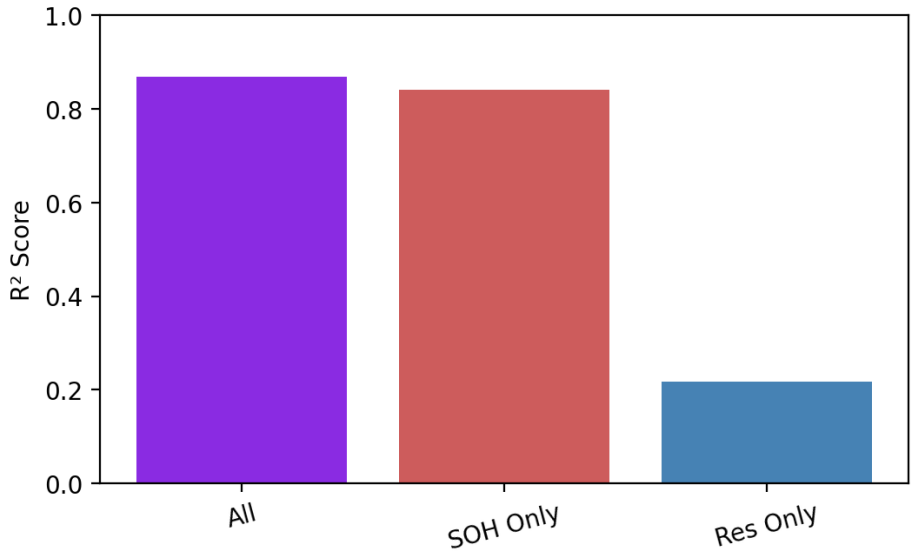


Figure 3. Feature-set comparison using SOH-only, Shadow Impedance-only, and combined feature configurations.

Under LOCO, detailed in Table 3, the model maintains MAE of 2.75% (NASA), 3.10% (Forklift), and 4.12% (Multi-Stage). The Temporal Split evaluation yields R^2=0.9381 and MAE of 7.14%, confirming the capture of non-linear degradation acceleration in the terminal regime.

Under the LOCO protocol, as detailed in Table 3, the model maintains a low error rate across cycling cohorts, achieving an MAE of 2.75% on NASA, 3.10% on Forklift, and 4.12% on Multi-Stage. Furthermore, the Temporal Split evaluation yields $R^2 = 0.9381$ and an MAE of 7.14%, confirming the model's capacity to capture non-linear degradation acceleration in the terminal regime.

Table 3. Intra-domain performance metrics under Leave-One-Cell-Out (LOCO) evaluation across all benchmark datasets.

Dataset	R^2	MAE (%)	RMSE (%)
NASA	0.9780	2.75	3.42
Forklift	0.9650	3.10	4.05
Multi-Stage	0.9410	4.12	5.18
eVTOL	0.9120	5.84	7.21
Oxford	0.8840	6.15	7.90
EVERLASTING	0.8210	8.92	11.30
Average Intra-Domain	**0.9168**	**5.15**	**6.51**

4.2. OOD Generalisation and Danger Bias Analysis

Table 4 presents the LODO results across all six datasets, directly measuring the OOD generalisation gap.

Table 4. OOD generalisation under LODO protocol. Danger Bias quantifies the fraction of operationally dangerous overestimates.

Dataset	R^2	MAE(%)	Safe(%)	Neutral(%)	Danger(%)
EVERLASTING	-2.60	47.24	7.0	4.2	**88.8**
eVTOL	-0.03	25.62	48.4	10.0	**41.6**
Forklift	0.97	3.76	14.2	71.0	14.7
Multi-Stage	0.93	5.77	36.0	56.8	7.2
NASA	0.98	3.63	6.2	72.7	21.1
Oxford	0.00	25.15	44.0	11.0	**45.0**

The results reveal three structurally distinct OOD failure modes:

Mechanism Shift (EVERLASTING, Danger Bias 88.8%): The EVERLASTING dataset is dominated by calendar ageing, passive capacity degradation driven by time and temperature, not cycling. Calendar ageing does not produce the ohmic resistance transients on which Shadow Impedance depends for routing, causing the router to misclassify these cells and the activated expert to produce overoptimistic life estimates.

Regime Saturation (eVTOL, Danger Bias 41.6%): eVTOL cells operate at extreme discharge rates with resistance values near $0.01\,\Omega$, below the lower routing threshold. This places all samples in the high-performance expert regardless of degradation level, collapsing predictive diversity.

Variance Collapse (Oxford, Danger Bias 45.0%): Oxford cells were cycled under rigidly controlled constant current, producing near-constant feature vectors with minimal variance. The model resorts to mid-range predictions that do not track actual degradation, yielding $R^2{=}0.00$.

Critically, even intra-distribution datasets exhibit residual Danger Bias: 21.1% for NASA. This demonstrates that Danger Bias exposes risks invisible to R^2 and MAE, even under favourable evaluation conditions.

It should be noted that the LODO (Leave-One-Dataset-Out) performance reported for the NASA and Multi-Stage datasets may be slightly optimistic, as these cohorts were previously utilized to establish the hyperparameter search bounds during the grid search phase.

4.3. Discussion

From a theoretical perspective, the observed generalisation failure is predominantly characterised as *Covariate Shift* [Kouw and Loog 2021]: the marginal distribution of input features changes between domains ($P_{\text{train}}(\mathbf{X}) \neq P_{\text{test}}(\mathbf{X})$) due to different operational charge and temperature profiles, while the conditional distribution, governed by the physics of electrochemical degradation ($P(Y|\mathbf{X})$), is assumed approximately stable. It is precisely this gap that the physics informed inductive bias via Z_{shadow} seeks to address: by injecting domain-invariant knowledge (ohmic resistance as a health proxy), model sensitivity to superficial covariate variations is reduced.

The experimental results underscore that high intra-domain accuracy provides no guarantees under domain shift, a phenomenon deeply rooted in the underlying electrochemistry. By inspecting the empirical dQ/dV curves, we can evaluate how distinct degradation signatures directly caused the observed out-of-distribution (OOD) failure modes. For instance, the severe variance collapse seen in the Oxford dataset ($R^2 = 0.00$) is visually explained by its highly uniform dQ/dV profiles, which offered insufficient electrochemical variance for the ensemble to track. Conversely, the massive Danger Bias in the EVERLASTING partition (88.8%) stems from the fact that passive calendar ageing shifts capacity without producing the pronounced dQ/dV peak transitions characteristic of cycling data, thereby blinding the resistance-driven routing function.

These diagnostic insights achieve a major objective of this work: they prove that through risk-aware metrics like Danger Bias and physically grounded visualizations, operators can look inside the model's blind spots and systematically trace algorithmic failures back to identifiable physical degradation mechanisms. Critically, these three OOD failure modes are not addressable by a single mitigation strategy. Calendar-ageing invisibility demands time-temperature features, regime saturation requires learned gating functions, whereas variance collapse may demand uncertainty quantification or confidence-based abstention.

Finally, we acknowledge that the EVERLASTING dataset contains only two physical cells, which statistically limits the robustness of the 88.8% Danger Bias reported under the LODO protocol. Rather than a definitive performance metric, conclusions drawn from this scenario should be interpreted as an empirical warning signal, a case study demonstrating how pure calendar ageing can render resistance, driven routing functions entirely blind. While the magnitude of this failure is consistent with the mechanistic explanation of calendar-ageing invisibility, these findings must be validated on larger calendar-ageing cohorts in future work.

5. Conclusion

This paper proposed and evaluated a data-centric framework for lithium-ion battery prognostics organised around three contributions: (i) a multi-domain benchmark spanning six heterogeneous datasets; (ii) a Regime-Aware Ensemble using Shadow Impedance as

a physics-informed inductive bias and routing signal; (iii) and a risk-aware evaluation framework centered on Danger Bias.

The results reveal a pronounced intra-/cross-domain performance gap. Although the Regime-Aware Ensemble achieves strong intra-distribution performance (R^2=0.9662, MAE $< 5\%$ under LOCO), the LODO evaluation exposes three structurally distinct OOD failure modes, with Danger Bias reaching 88.8% in the most critical scenario. These findings indicate that model-centric improvements alone are insufficient under strong domain shift: reliable battery prognostics requires an OOD-aware benchmark design, physically grounded inductive biases, and evaluation frameworks that quantify operational risk asymmetrically.

While this work focuses on establishing Danger Bias as a diagnostic evaluation tool, mitigating this operational risk ultimately requires shifting from symmetric optimization to risk-asymmetric training. Consequently, future work will focus on: (i) replacing the standard MSE loss in the base regressors with an asymmetric objective function (such as Linex or pinball loss) to explicitly penalize overestimations during training; (ii) developing learned gating functions to replace hard-threshold routing; (iii) applying domain adaptation to reduce Danger Bias under distributional shift; (iv) and incorporating uncertainty quantification methods to support confidence-based abstention.

References

Azad, A. et al. (2026). Second-life applications for retired electric vehicle batteries: challenges, opportunities, and future directions. *Journal of Energy Storage*, 110:114980.

Bills, A. et al. (2021). eVTOL battery dataset. *Kilthub – Carnegie Mellon University*. Available at: `https://kilthub.cmu.edu/articles/dataset/eVTOL_Battery_Dataset/14226830`.

Celik, B. et al. (2022). Prediction of battery cycle life using early-cycle data, machine learning and data management. *Batteries*, 8(12):266.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York. ACM.

Fricke, K. A. et al. (2023). Randomized and recommissioned battery dataset. Technical report, NASA Prognostics Center of Excellence. Available at: `https://data.nasa.gov/dataset/randomized-and-recommissioned-battery-dataset`.

Howey, D. A. and Birkl, C. (2017). Oxford battery degradation dataset 1. University of Oxford. `https://doi.org/10.5287/bodleian:KO2kdmYGg`.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Jiang, Y. and Song, W. (2023). Predicting the cycle life of lithium-ion batteries using data-driven machine learning based on discharge voltage curves. *Batteries*, 9(8):413.

Kouw, W. M. and Loog, M. (2021). A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):766–785.

Li, K. and Chen, X. (2025). Machine learning-based lithium battery state of health prediction research. *Applied Sciences*, 15(2):516.

Lu, J. et al. (2023). Deep learning to estimate lithium-ion battery state of health without additional degradation experiments. *Nature Communications*, 14(1):2760.

Mayemba, Q. et al. (2024). General machine learning approaches for lithium-ion battery capacity fade compared to empirical models. *Batteries*, 10(10):367.

Ni, Y. et al. (2023). Hybrid estimation of residual capacity for retired LFP batteries. In *Proceedings of ICEIV 2022*, volume 1018 of *Lecture Notes in Electrical Engineering*, pages 44–51. Springer, Heidelberg.

Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350.

Sambasivan, N. et al. (2021). Everyone wants to do the model work, not the data work: data cascades in high-stakes AI. In *Proceedings of the 2021 ACM CHI Conference on Human Factors in Computing Systems*, pages 1–15, New York. ACM.

Severson, K. A. et al. (2019). Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5):383–391.

Stroebl, F. et al. (2024). Multi-stage lithium-ion battery aging study. Figshare. `https://doi.org/10.6084/m9.figshare.25975315`.

Takahashi, A., Allam, A., and Onori, S. (2023). Evaluating the feasibility of batteries for second-life applications using machine learning. *iScience*, 26(4):106547.

Tan, Y. et al. (2022). Early prediction method for remaining useful life of retired batteries in second-life applications. In *Proceedings of 2022 IEEE International Conference on Energy Internet and Energy System Integration (EI2)*, pages 522–530.

Trad, K. (2021). Calendar ageing test results on commercial 18650 Li-ion cell @ 25°C and 45°C. 4TU.ResearchData. `https://doi.org/10.4121/13804304.v1`.

Valizadeh, A. and Amirhosseini, M. H. (2024). Machine learning in lithium-ion battery: applications, challenges, and future trends. *SN Computer Science*, 5(6):697.

Vilsen, S. B. and Stroe, D.-I. (2023). Lithium-ion battery degradation dataset based on a realistic forklift operation profile.

Waag, W., Fleischer, C., and Sauer, D. U. (2014). Critical review of the methods for monitoring of lithium-ion batteries in electric and hybrid vehicles. *Journal of Power Sources*, 258:321–339.

Zhai, Q. et al. (2024). Machine learning for full lifecycle management of lithium-ion batteries. *Renewable and Sustainable Energy Reviews*, 202:114647.

Zhang, H. et al. (2025). Battery lifetime prediction across diverse ageing conditions with inter-cell deep learning. *Nature Machine Intelligence*, 7(2):270–277.