

Article

Safe-Calibrated TCN–Transformer Transfer Learning for Reliable Battery SoH Estimation Under Lab-to-Field Domain Shift

Kumbirayi Nyachionjeka ¹ and Ehab H. E. Bayoumi ^{2,*}¹ Department of Electrical Engineering, University of Botswana, Gaborone Private Bag UB 0022, Botswana; nyachionjekak@ub.ac.bw² Mechatronics and Robotics Section, Mechanical Engineering Department, Faculty of Engineering, The British University in Egypt, El Sherouk 11837, Egypt

* Correspondence: ehab.bayoumi@gmail.com

Abstract

Battery state-of-health (SoH) estimation is central to transportation electrification because it conditions safety limits, warranty accounting, power capability management, and long-horizon fleet optimization. Although deep temporal architectures can achieve high laboratory accuracy, field deployment is frequently limited by laboratory (Lab)-to-field (L2F) domain shift that alters input statistics, feature definitions, and noise regimes. Under such a shift, predictors may remain strongly monotonic, preserving degradation ordering and become operationally unreliable due to systematic output distortion (e.g., compression/warping of the SoH scale). A deployment-complete L2F transfer learning pipeline is presented, built around a gated Temporal Convolutional Network (TCN)–Transformer fusion backbone, domain-specific adapters and heads, alignment-regularized fine-tuning, and row-level inference via sliding-window overlap averaging. To address the dominant deployment failure mode, a Safe Calibration stage robustly filters calibration pairs and selects among candidate calibrators under a strict do-no-harm criterion. On an unseen deployment stream (2154 labeled rows), overlap-averaged raw inference achieves MAE = 0.0439, RMSE = 0.0501, and $R^2 = 0.7451$, consistent with mid-to-high SoH range compression, while Safe Calibration (Isotonic-Balanced selected) corrects nonlinear scaling without violating monotonic structure, improving to MAE = 0.0188, RMSE = 0.0252, and $R^2 = 0.9357$ to obtain a complete understanding of the challenges due to domain shifts, evaluation is extended to include other architecture baselines such as TCN-only, Transformer-only, Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM), and a Ridge regression baseline. Also added is explicit alignment and calibration ablations that include CORAL off/on, that is, none vs. Safe-Global vs. Context-Aware under identical leakage-safe splits and the same overlap-averaged deployment inference operator. This work goes beyond peak-score reporting and looks at the robustness of a pipeline under domain shift, which is quantified across four random seeds and multiple deployment streams, with uncertainty summarized via mean $\pm$ std and bootstrap confidence intervals for Mean of Absolute value of Errors (MAE)/Root of the Mean of the Square of Errors (RMSE) computed from per-example absolute errors.



Academic Editor: Grzegorz Sierpiński

Received: 11 February 2026

Revised: 7 March 2026

Accepted: 11 March 2026

Published: 17 March 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the

World Electric Vehicle Association.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: state of health; lab-to-field; domain shift; transfer learning; safe calibration; isotonic regression; temporary convolutional network; transformer models; battery management systems

1. Introduction

The rapid adoption of electric mobility is forcing engineers to push propulsion and on-board power systems to rely on battery-based architectures, and these battery architectures need to function safely and efficiently across diverse climates, duty cycles, and charging behaviors [1]. Because of this impending reliance on battery-based architectures, observing battery SoH has become important as SoH provides important details on the irreversible degradation processes in the battery system and directly provides information about energy, power capability, and safety margins [2,3]. Accurate SoH estimation, therefore, leads to proper warranty enforcement, information-based predictive maintenance, charging control that is done on the knowledge of the state-of-health of the battery, and precise optimization at fleet-level [4,5]. Despite these advantages offered by monitoring SoH, it is not possible to continuously observe and record SoH during operation using non-data-based techniques without conducting cumbersome, intrusive and time-consuming tests. Therefore, it has become important for the design of estimation techniques that can be used for onboard measurements and feature extraction for SoH modeling [6,7].

Thus, data-driven SoH estimation methods have rapidly matured in the last few years, supported by the advances in deep temporal representation learning and the budding availability of aging datasets. The advent of transformer-based techniques has resulted in models that have the ability to capture long-range degradation dependencies and global context [8,9]. Simultaneously, the use of convolutional neural networks (CNNs) and TCNs leads to stable training and multi-scale receptive models suitable for mapping nonstationary battery signatures. Despite the genesis of robust data-based models, there still remains a challenge of distribution shift, which stands as a hindrance to the practical deployment of these models in the Battery Management System (BMS) [10–13]. Models trained on controlled laboratory aging experiments fail to operate as they should when used in field fleets due to the differences between the lab and field domains. The domain mismatch is mainly because of differences in operating conditions, sampling regimes, feature engineering pipelines, and sensor noise [14]. The mismatches lead to systematic prediction bias even when relative ordering is preserved, which is particularly problematic in safety- and warranty-critical workflows [15–17].

One of the most common manifestations of this shift is output distortion, where the estimator rank cycles correctly, that is, higher SoH earlier and lower SoH later, but the predicted SoH values become mis-scaled and unreliable later on. This is the case because predictions may be globally shifted, suffering systematic bias, compressed, which is reduced dynamic range, and nonlinearly warped, yielding over-optimistic or over-conservative health levels despite preserved monotonic degradation. Downstream policies depend on calibrated SoH values and not just rankings, for example, derating thresholds, warranty triggers, and safety margins. Therefore, correcting output distortion is a separate deployment requirement beyond representation alignment [18,19].

Transfer learning and domain adaptation have been proposed to mitigate this mismatch by reusing source-domain representations and aligning latent distributions. For instance, domain adaptation frameworks have demonstrated cross-domain benefits for health-related estimation when the training objective explicitly reduces distribution discrepancy [20]. Transfer learning has also been used to accelerate battery assessment and reduce testing burden by exploiting shared degradation structure across cells and protocols [21,22]. Yet, two deployment gaps remain. First, many pipelines stop at fine-tuning and do not specify a full inference strategy that outputs stable and row-level health trajectories from streaming field records. Second, even after adaptation, field deployment frequently fails because outputs are mis-scaled (compressed, shifted, or nonlinear-warped) under domain shift; this failure mode is not reliably addressed by training-time alignment alone [23–25].

Even when degradation ordering is preserved, L2F shift can induce systematic output distortion (scale compression/shift/nonlinear warping), which causes large errors in absolute SoH despite correct ranking. This motivates a deployment-complete pipeline as shown in Figure 1, that (i) aligns representations (CORAL) and (ii) applies an auditable post hoc Safe Calibration stage with do-no-harm selection and robust pair filtering [26].

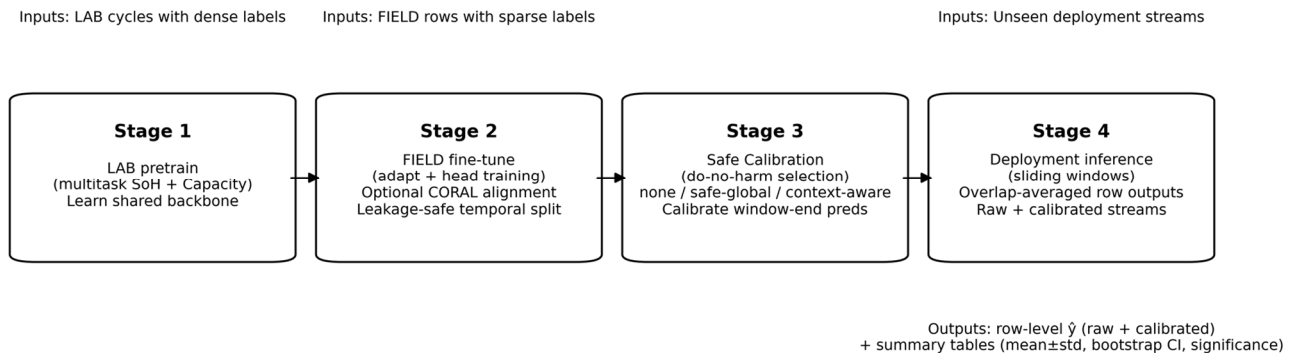


Figure 1. Pipeline framework.

This paper targets these gaps by proposing a deployment-complete Lab-to-Field pipeline shown in Figure 1, that combines a TCN–Transformer fusion backbone with adapters/heads for heterogeneous inputs, alignment-regularized fine-tuning, streaming-compatible overlap-averaged inference, and a Safe Calibration module that corrects systematic output distortion under a strict do-no-harm rule. The safe selection principle is motivated by the broader calibration literature, where post hoc calibration can improve reliability but may also harm performance if fitted on noisy or unrepresentative pairs [27–29]. Here, calibration is designed as a controlled and auditable step for SoH regression under shift, with robust pair filtering and a selection criterion that prevents regression in holdout error. Furthermore, Deployment evaluation is extended beyond a single run/stream by reporting repeated-seed performance (four seeds) and multi-deployment-stream results under a leakage-safe temporal split. The comparison set is expanded to include architecture baselines (Fusion, TCN-only, Transformer-only, GRU, LSTM), an explicit CORAL on/off alignment ablation, a calibration ablation (none vs. Safe-Global vs. Context-Aware), and a classical Ridge baseline. Statistical reporting is strengthened using mean $\pm$ std over seeds and bootstrap confidence intervals (CIs) for MAE/RMSE computed from per-example absolute errors, enabling uncertainty-aware model comparison under deployment-realistic overlap-averaged inference.

The main contributions of this paper are as follows:

1. A deployment-complete Lab-to-field transfer learning framework is proposed for battery SoH estimation, integrating domain-specific adapters, alignment-regularized fine-tuning, and overlap-averaged inference to produce stable and row-level predictions suitable for BMS deployment.
2. A gated TCN–Transformer fusion backbone is developed to jointly capture local degradation patterns and long-range temporal dependencies, enabling effective representation transfer across heterogeneous laboratory and field data.
3. A Safe Calibration strategy is introduced based on monotonic post hoc recalibration with robust pair filtering and a strict do-no-harm selection criterion, correcting systematic distortion under domain shift while preserving degradation ordering.
4. Effectiveness is demonstrated on multiple unseen deployment streams with repeated-seed evaluation, showing substantial improvements in MAE/RMSE/ R^2 and indicating that deployment errors are primarily driven by miscalibration rather than loss of health ordering.

5. Expanded baseline and ablation suite: Fusion vs. TCN-only vs. Transformer-only vs. GRU vs. LSTM; Ridge baseline (raw + calibrated); calibration ablation (none, Safe-Global, Context-Aware); alignment ablation (CORAL off/on), all under identical leakage-safe splits and identical deployment inference operator for fairness.
6. Statistical validation: mean $\pm$ std over four seeds and bootstrap CIs for MAE/RMSE (computed from saved per-example absolute error arrays) aggregated across multiple deployment streams.

Related Works

Deep learning for battery health estimation has evolved from feature-based regression to representation learning, which consumes charging/discharging trajectories or their engineered summaries. Researchers have demonstrated that transformer-based methods possess the ability to map long-range dependencies and global context. These transformer methods include incremental-capacity-informed Transformer models and hybrid sequence learners designed for fast-charging and constrained measurement windows [30,31]. At the same time, the emergence of transfer learning has provided engineers with the ability to reduce labeling and testing burdens by utilizing knowledge from different battery types and operating protocols [32].

Many works are reported in the literature that address cross-domain performance by relying on explicit alignment mechanisms. Particle Swarm Optimization (PSO) has been demonstrated and used to achieve deep domain adaptation. PSO has been applied to reduce cross-domain discrepancy while tuning adaptation-relevant hyperparameters [33]. On the other hand, CCN-Transformer fusion methods have been shown. These form platforms that provide local and global mapping abilities of battery features through better generalization across operating conditions by aligning multi-modal and multi-scale representations [34]. This motivates a complementary focus on output reliability that ensures that predicted SoH values remain well-scaled and decision-consistent under field conditions. In practice, this means the model should not only preserve the degradation trend but also produce numerically trustworthy health levels with minimal systematic bias, stable dynamic range, and predictable error behavior. Such reliability is essential because BMS actions are typically threshold-driven—small mis-scaling can shift derating, maintenance, or safety triggers even when the ordering of cycles is correct [26,35].

Post hoc calibration is a research-approved technique that allows for the mapping of raw model scores and predictions into better-aligned values without changing the underlying feature extractor. Regression and uncertainty estimation research show that calibration methods possess the ability to improve reliability while still requiring careful validation. The validation is necessary since calibration is not uniform across datasets and is not beneficial [36,37]. Isotonic recalibration, in particular, has been studied as a monotonic post-processing step that can restore calibration properties under challenging signal-to-noise regimes [38]. These insights motivate the Safe Calibration selection rule proposed here, which applies calibration only when it demonstrably improves a held-out criterion.

A summary of the qualitative approach to related works is presented in Table 1. Studies between 2023 and 2026 demonstrate that transfer learning and domain adaptation materially improve SoH estimation when the training and deployment distributions differ, but they typically emphasize representation transfer rather than deployment reliability under systematic output distortion. PSO-assisted domain adaptation methods explicitly reduce cross-domain mismatch at the feature/latent level [33], while vehicle-oriented multi-source DA frameworks leverage heterogeneous sources and real-vehicle data to improve generalization across operational regimes [15,16]. In parallel, transfer learning SoH models have

been tailored to challenging regimes such as fast charging, where degradation signatures and measurement constraints differ from conventional cycling [31], and to protocol-induced shifts such as cross-formation transfer, where the same chemistry can present meaningfully different early-life trajectories [32]. Alignment-heavy architectures (e.g., local–global CNN–Transformer designs) further strengthen invariance by fusing multi-scale/multi-modal cues and aligning representations across conditions [34], and ensemble-style modeling (multi-expert fusion) improves accuracy by combining complementary predictors [9]. Collectively, these works establish that cross-condition generalization benefits from stronger representations and explicit discrepancy reduction, yet they often report performance primarily in curated cross-split settings and do not fully specify a deployment-complete inference pipeline that produces stable row-level trajectories from streaming windows.

Table 1. Qualitative comparison with representative SoH estimation literature (2023–2026).

Work (Year)	Core Model/Focus	Transfer/DA	Calibration Focus	Primary Evaluation Setting	Data Source/Modality	Challenges and Limitations
Ma et al. [33] (2023)	PSO-assisted deep domain adaptation (DA) for SoH	Yes	Not central	Cross-domain/cross-condition datasets	Lab/benchmark cross-domain splits	Requires DA objective tuning; may not address heterogeneous feature schemas or deployment streaming constraints
Liu et al. [23] (2025)	Multi-modal SoH evaluation using open-source EV data	Partial	Not central	Multi-modal real-data SoH evaluation	Open EV data (real-world, multi-modal)	Multi-modal data availability/quality varies; fusion can be sensitive to missing modalities and label sparsity
Fan et al. [9] (2025)	Multi-expert fusion for SoH regression	No (primarily modeling)	Not central	Benchmark-style SoH prediction	Lab/benchmark datasets	Gains depend on expert diversity; it does not explicitly mitigate domain shift or provide deployment reliability guarantees
Tian et al. [15] (2025)	Multi-source domain adaptation with real-vehicle data	Yes	Not central	DA framework validated on vehicle data	Real-vehicle data + multi-source DA	Source-selection and domain coverage can limit generalization; still vulnerable to systematic output bias under shift
Li et al. [16] (2025)	Domain-adaptive SoH prediction (deep learning)	Yes	Not central	Vehicle batteries; domain-adaptive learning	Vehicle battery datasets (domain shift)	Assumes access to a representative target domain; performance can degrade when target labels are extremely sparse or nonstationary
Zhao et al. [31] (2025)	SoH under fast charging via deep transfer learning	Yes	Not central	Transfer across fast-charging conditions	Fast-charging protocol shift	Focused on fast-charging regime; may not generalize to broader operational heterogeneity or streaming inference settings

Table 1. Cont.

Work (Year)	Core Model/Focus	Transfer/DA	Calibration Focus	Primary Evaluation Setting	Data Source/Modality	Challenges and Limitations
Gui et al. [34] (2025)	Local-global CNN-Transformer + multi-modal alignment	Yes	Not central	Cross-domain SoH via alignment	Multi-modal/alignment across domains	Alignment requires careful balancing; architectural complexity can increase computing, and may not address post hoc calibration needs
Zhang et al. [22] (2026)	Heterogeneous TL; one-shot SoH estimation	Yes	Not central	Feature-gap bridging; minimal target labels	One-shot/low-label target setting	One-shot assumption can be fragile; sensitive to representativeness of the single target sample and to feature mismatch
Chen et al. [32] (2026)	Transfer learning under cross-formation protocols	Yes	Not central	Cross-formation protocol transfer	Formation-protocol shift	Targets protocol shifts specifically; it does not necessarily resolve real-vehicle noise, sampling irregularity, or deployment calibration drift
This work (2026)	TCN-Transformer fusion + adapters + CORAL + overlap inference	Yes (CORAL)	Safe Isotonic-Balanced (do-no-harm)	Unseen deployment stream + row-level overlap averaging	end-of-cycle/operational-window 2 (EOCV2) operational-window (field-like) + unseen stream	Calibration depends on labeled target subset; rare-context regimes (imbalanced cc_end/ck_end) can remain challenging without rebalancing or uncertainty-aware conditioning

In contrast to approaches that primarily target absolute laboratory accuracy, the present work focuses on the dominant L2F deployment failure mode in which predictions remain strongly monotonic yet become mis-scaled (compressed/warped) under shift. This motivates an explicitly deployment-complete four-stage pipeline: (i) lab pretraining to learn transferable degradation ordering; (ii) field fine-tuning with domain-specific adapters and optional CORAL alignment to regularize representation mismatch; (iii) Safe Calibration as an auditable reliability layer, where calibration pairs are robustly filtered (quantile clipping and median-absolute-deviation (MAD)-based residual screening) and a monotonic recalibration mapping is selected only if it improves a held-out criterion under a strict do-no-harm rule; (iv) deployment inference via overlap-averaged sliding windows to produce smooth row-level SoH trajectories suitable for Battery Management System consumption. This separation between learning transferable ordering (transfer/alignment) and repairing deployment-scale distortion (safe calibration and deployment inference) is central when domain shift preserves ranking but violates the SoH scale.

The evaluation protocol is correspondingly framed around robustness under domain shift rather than peak-score reporting from a single run. Experimental scope is extended to multi-deployment-stream testing and repeated-seed reporting (four seeds), and the comparison set is broadened to include architecture baselines (Fusion, TCN-only, Transformer-only, GRU, LSTM), a Ridge regression baseline (raw and calibrated), an explicit alignment ablation (CORAL off/on), and a calibration ablation (none vs. Safe-Global vs. Context-Aware),

all under identical leakage-safe temporal splitting and the same overlap-averaged inference operator for fair comparison. Statistical uncertainty is summarized using mean $\pm$ std across seeds and bootstrap confidence intervals (CIs) for MAE/RMSE computed from per-example absolute errors aggregated across runs.

The article is organized as follows. Section I motivates reliable battery SoH estimation for transportation electrification and highlights the Lab-to-field generalization challenge. Section II reviews recent SoH estimation, transfer/domain-adaptation, and calibration literature to position the work. Section III formulates the problem under deployment distribution shift and describes the dataset setting and protocols, including the laboratory source corpus, the field-like EOCV2 streams used for adaptation/calibration and for unseen deployment testing, and the leakage safeguards. Section IV details the proposed methodology, including the gated TCN–Transformer fusion backbone, domain-specific adapters/heads, alignment-regularized fine-tuning, and Safe Calibration under a do-no-harm selection criterion. Section V presents the deployment inference procedure based on overlap-averaged sliding windows to generate stable row-level SoH trajectories. Section VI reports quantitative and qualitative results on the fully unseen deployment stream, including calibration-label efficiency and drift-stability analyses. Section VII discusses the implications, limitations, and how the approach compares with recent transfer learning and CNN–Transformer alignment methods, and Section VIII concludes with key findings and directions for future work.

2. Materials and Methods

2.1. Problem Formulation and Dataset Setting

Let x_t denote the input at time index t , and let $y_t \in [0, 1]$ denote the corresponding SoH label (normalized capacity-based SoH). The objective is to learn a predictor f_θ that generalizes to an unseen deployment distribution $\mathcal{D}_{\text{new}}$ despite L2F domain shift. The source domain $\mathcal{D}_S$ consists of laboratory sequences collected under controlled cycling and temperature conditions. The target domain $\mathcal{D}_T$ consists of field sequences where inputs are engineered, features are aggregated over operational windows, and observations may include categorical context variables and non-uniform sampling. Because $\mathcal{D}_S$ and $\mathcal{D}_T$ differ structurally and statistically, the direct application of a laboratory-trained model can be brittle. Given training data from $\mathcal{D}_S$ and limited labeled data from $\mathcal{D}_T$, the parameter vector θ is estimated by minimizing deployment risk, as modeled by Equation (1).

$$\min_{\theta} E_{(x,y) \sim \mathcal{D}_{\text{new}}} [L(y, f_{\theta}(x))] \quad (1)$$

where f_θ is the learned predictor, L is the loss function, and the expectation is taken over the unseen deployment distribution $\mathcal{D}_{\text{new}}$, θ is the trainable model parameters, $E_{(x,y)}$ is the expectation (average) over data drawn from a distribution, (x, y) , $\mathcal{D}_{\text{new}}$ is the new/target data distribution representing the deployment domain, $L(\cdot)$ is the loss function, $f_\theta(x)$ is the model evaluated on input x , and y is the ground-truth label/target for x , that is, the true SoH.

In practice, learning uses supervised prediction loss on available labeled data together with an alignment regularizer that reduces latent distribution mismatch. Outputs are constrained to preserve physical plausibility and prevent pathological extrapolation beyond meaningful SoH ranges.

2.2. Datasets and Protocols

Source (lab) data are drawn from the NASA battery aging set (cells B0005, B0006, B0007, B0018) [39], merged into a unified table of 185,721 time-sampled records spanning 4 cells and 636 charge/discharge cycles (per-cell coverage: B0005: 168 cycles, B0006: 168 cycles, B0007: 168 cycles, B0018: 132 cycles). Each cycle is represented by the measured voltage, current, and temperature trajectories, and is resampled to a fixed length ($N_POINTS_LAB = 128$) after constructing three physics-derived channels (cumulative charge q , cumulative energy e , and dv/dt). Laboratory SoH supervision is capacity-based, computed as $SoH = Capacity(Ah)/C_ref$, where C_ref is the maximum observed capacity in the lab corpus (global-max normalization as in the implementation). The observed laboratory temperature range is 22.35–42.33 °C. Target (field) data use an engineered EOCV2 representation in which each row comprises aggregated electrical/thermal descriptors (e.g., Δq , Δe , duration, OCV/SOC estimates, and context/aging tags) derived from periodic checkup records. Field adaptation and calibration use stream P001_1_S01_C10_MULTIPLE ($N = 129,081$; 16,332 labeled via `soh_cap`, with 16,041 labels in $[0.20, 1.10]$, range 0.2002–0.9875). Deployment evaluation is performed on multiple unseen streams excluded from all training/validation/calibration and used only at inference time: P076_1_S15_C04 ($N = 50,579$; 2154 labeled; 0.3917–0.9709), P073_2_S12_C02 ($N = 23,126$; 1039 labeled; 932 in $[0.20, 1.10]$, 0.2029–0.9679), and P073_3_S14_C01 ($N = 25,346$; 1156 labeled; 1006 in $[0.20, 1.10]$, 0.2001–0.9529) [39,40]. To prevent optimistic bias, the protocol enforces strict cross-stream isolation, leakage-safe temporal splitting with scalars fit on training partitions only, and calibration hygiene using temporally held-out selection with quantile clipping and MAD filtering together with a do-no-harm criterion. Table 2 summarizes the key properties of the lab, field, and deployment streams.

2.3. Proposed Methodology

The proposed pipeline consists of four interconnected stages. First, a shared temporal backbone is pretrained on laboratory data to learn degradation-sensitive representations. Second, the model is adapted to field data using domain-specific adapters and heads combined with an alignment-regularized objective. Third, Safe Calibration selects a post hoc calibrator from a candidate set using robust filtering and a do-no-harm selection criterion. Finally, deployment inference generates row-level SoH trajectories through sliding-window overlap averaging, which stabilizes predictions and reduces window-boundary artifacts.

Table 2. Dataset summary for lab, field, and deployment streams used in the evaluation.

Dataset Role	Dataset Source Identifier	Representation	Total Rows (N)	Labeled Rows	SoH Label	Valid SoH Range	Notes
LAB (source)	NASA battery aging (B0005, B0006, B0007, B0018) [39]	Per-cycle resampled V/I/T with physics-derived channels (q, e, dV/dt)	185,721	Cycle-level SoH available	SoH	—	Stage 1 pretraining; global-max capacity normalization.
FIELD (target)	EOCV2: P001_1_S01_C10_MULTIPLE [40,41]	Engineered per-row descriptors (Δq , Δe , OCV/SOC, thermal, context)	129,081	16,332 labeled; 16,041 within [0.20, 1.10]	soh_cap	0.2002–0.9875	Fine-tuning + calibration; leakage-safe split (gap).
Deployment (inference-only)	EOCV2: P076_1_S15_C04 [40,41]	Same EOCV2 schema	50,579	2154 labeled; all within [0.20, 1.10]	soh_cap	0.3917–0.9709	Held-out; used only for deployment evaluation.
Deployment (inference-only)	EOCV2: P073_2_S12_C02 [40,41]	Same EOCV2 schema	23,126	1039 labeled; 932 within [0.20, 1.10]	soh_cap	0.2029–0.9679	Held-out; cross-stream robustness evaluation.
Deployment (inference-only)	EOCV2: P073_3_S14_C01 [40,41]	Same EOCV2 schema	25,346	1156 labeled; 1006 within [0.20, 1.10]	soh_cap	0.2001–0.9529	Held-out; cross-stream robustness evaluation.

2.3.1. TCN–Transformer Fusion Backbone

The backbone fuses two complementary temporal pathways. The TCN pathway captures local and multi-scale patterns through stacked dilated causal convolutions, while the Transformer encoder captures long-range dependencies via self-attention [42]. For a TCN layer ℓ with dilation δ_ℓ and kernel size k , the causal convolution output is obtained by Equation (2):

$$h_\ell(t) = \sum_{m=0}^{k-1} W_\ell(m) \cdot h_{\ell-1}(t - \delta_\ell m) + b_\ell \quad (2)$$

where ℓ is the current layer index, $\ell - 1$ is the previous layer index, m is a kernel index ranging from 0 to $k - 1$, k is the kernel size, also referred to as filter length, $W_\ell(m)$ is the convolution weight, $h_{\ell-1}(t - \delta_\ell m)$ is the input activation from the previous layer, δ_ℓ is the dilation factor, and b_ℓ is the bias term for the layer ℓ , it shifts the output up and down.

Residual connections and normalization stabilize optimization and expand receptive fields without the training instabilities often encountered in recurrent networks. In the attention pathway shown in Equation (3) [43]. Multi-head self-attention is computed as follows:

$$ATT(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad Q = HW_Q, \quad V = HW_V, \quad K = HW_K \quad (3)$$

where Q is the matrix of queries, K is a matrix of key vectors, V represents a matrix of values, QK^T is the matrix multiplication that results in attention scores, $\sqrt{d_k}$ is the square root of the key/value dimensionality, $\frac{QK^T}{\sqrt{d_k}}$ is the scaled attention, H is the input-representation matrix, W_Q is the learnable projection matrix that maps H into queries, W_K is the learnable projection matrix that maps H into keys, W_V is the learnable projection matrix that maps H into values.

Positional encodings preserve temporal ordering [44,45]. To combine the complementary features, a learnable gate produces a fused representation as in Equation (4),

$$\alpha = \sigma(W_g[h_{TCN}; h_{ATT}] + b_g) \quad (4)$$

α is the gate vector or gate scalar, $\sigma(\cdot)$ is the sigmoid function, W_g is the learnable weight matrix for the gate, $[h_{TCN}; h_{ATT}]$ is a concatenation of the two feature vectors and b_g is the learnable bias vector for the gate layer. The subscripts TCN and ATT are features produced by a TCN branch and features produced by an attention/Transformer branch, respectively. Equation (5) illustrates the feature fusion between the two branches,

$$h = \alpha \odot h_{TCN} + (1 - \alpha) \odot h_{ATT} \quad (5)$$

h is the fused output feature vector, $\odot$ is the elementwise (Hadamard) multiplication and $(1 - \alpha)$ is the complement gate, which dictates how much should be taken from the attention branch. Equations (4) and (5) define a gate that allows the model to emphasize local convolutional cues or global attention context, depending on the operating regime and the signal quality [42]. In the L2F setting, the fusion is valuable because field features can be smoother and more aggregated than laboratory waveforms, shifting the relative importance of local versus global cues [34].

2.3.2. Domain-Specific Adapters and Heads

Because laboratory and field inputs are heterogeneous, a shared backbone alone is insufficient. The pipeline, therefore, uses domain-specific adapters that map domain-dependent features into a shared latent dimension before the fusion backbone. Separate prediction heads are used for source and target supervision during training to reduce negative transfer. At deployment, the field adapter and shared backbone are used with the selected calibrator to produce final SoH estimates [15,22].

2.3.3. Alignment-Regularized Fine-Tuning

To reduce L2F discrepancy, an alignment regularizer is introduced that encourages the covariance structure of latent representations to match between source and target mini-batches. Let $H_S \in \mathbb{R}^{n_S \times d}$ and $H_T \in \mathbb{R}^{n_T \times d}$ denote latent feature matrices for the source (lab) and target (field) batches, respectively, and let C_S and C_T denote their sample covariance matrices. A CORAL-style loss [46], is defined in Equation (6) by penalizing the Frobenius norm of the covariance mismatch:

$$L_{align} = \|C_S - C_T\|_F^2, \quad C = \frac{1}{n-1} (H - 1\mu^\top)^\top (H - 1\mu^\top). \quad (6)$$

Here, $\|\cdot\|_F$ denotes the Frobenius norm, H is a batch feature matrix (H_S and H_T), μ is the batch mean vector, and $1 \in \mathbb{R}^n$ is an all-ones vector used to broadcast μ across rows.

The complete training objective couples supervised prediction with alignment and regularization as shown in Equation (7),

$$L_{total} = L_{pred} + \lambda_{align} L_{align} + \lambda_{reg} \|\theta\|_2^2 \quad (7)$$

where L_{pred} is the prediction loss, L_{align} is the alignment loss, λ_{align} is the weight for the alignment loss, λ_{reg} is the weight for the regularization term, θ is the model parameters (weights), and $\|\theta\|_2^2$ is the squared ℓ_2 norm of the parameters.

Alignment is particularly relevant for regression tasks such as SoH estimation because feature scale differences can induce systematic output distortion. Domain adaptation methods have demonstrated this sensitivity and the importance of explicit discrepancy reduction in cross-domain battery health settings [16,33].

2.3.4. Safe Calibration (Isotonic-Balanced Selected)

Even after adaptation, deployment errors may be dominated by systematic output distortion that preserves ordering [35]. The observed results are consistent with this scenario: raw predictions are strongly monotonic but exhibit nonlinear scaling. In this work, calibration is learned from validation pairs $(\hat{y}, y)$, which can contain leverage points due to sparse labels, noise, or rare regimes. Even after adaptation, deployment error can be dominated by systematic output distortion that preserves ordering. In the present setting, raw predictions remain strongly monotonic but exhibit nonlinear scaling under domain shift. Calibration pairs are formed from validation predictions and robustly filtered before fitting to improve stability. Specifically, extreme predicted values are trimmed using low/high quantiles, and residual outliers are removed using MAD filtering; these steps reduce the influence of leverage points and heteroskedastic tails that can otherwise induce step artifacts or spurious extrapolation in monotone calibrators. Candidate calibrators include the identity mapping, ridge-linear mapping, and isotonic regression. Isotonic regression is used as the primary nonparametric option because it enforces monotonicity (preserving degradation ordering) while correcting arbitrary nonlinear distortions without assuming a parametric error model. To mitigate overfitting and plateauing in dense SoH regions, bin-balanced weighting is applied (Isotonic-Balanced), and the final mapping is

selected under a strict do-no-harm rule evaluated on an internal calibration holdout set. The selected calibrator g^* is defined as

$$g^* = \operatorname{argmin}_{g \in \mathcal{G}} \operatorname{RMSE}_{\text{holdout}}(g(\hat{y}), y) \text{ s.t. } \operatorname{RMSE}_{\text{holdout}}(g(\hat{y}), y) \leq \operatorname{RMSE}_{\text{holdout}}(\hat{y}, y) \quad (8)$$

where $\mathcal{G}$ is the set of candidate calibration mappings, $\hat{y}$ denotes the uncalibrated predictions, y denotes the ground-truth targets, and $g(\hat{y})$ are the calibrated predictions. $\operatorname{RMSE}_{\text{holdout}}(\cdot)$ is computed on the internal holdout set, and the constraint enforces that calibration does not degrade RMSE relative to the original model.

This criterion ensures the calibration is applied only when it improves the chosen held-out metric, reflecting the broader calibration research that post hoc calibration can help reliability, but should be validated to avoid degradation [25]. In the reported experiments, Safe Calibration selected the Isotonic-Balanced mapping.

2.3.5. Implementation Outline

Stage 1 pretraining uses a multi-task objective that predicts SoH and an auxiliary capacity proxy, with the auxiliary term downweighted to regularize the backbone without dominating SoH learning. Laboratory sequences are formed from cycle-resolved voltage-current traces by constructing physics-derived channels (cumulative charge q , cumulative energy e , and time-derivative (slope) (dv/dt) and resampling each cycle to a fixed length before ingestion by the TCN-Transformer fusion backbone. In Stage 2, field prediction is performed by a domain-specific head that augments the shared latent representation with compact categorical embeddings of operational context (e.g., checkup ID and condition ID) to absorb discrete protocol or measurement effects. In Stage 3, alignment is applied as a CORAL-style covariance penalty with a warm-up ramp, optionally combined with exponential moving average stabilization and bin-weighted supervision to reduce edge-region bias. In Stage 4, Safe Calibration evaluates multiple lightweight mappings (including identity, ridge-linear, isotonic, and hybrid iso-linear candidates) and applies the selected mapping only when held-out error improves, after which deployment uses overlap-averaged sliding-window inference to yield stable row-level trajectories. The above discussion can be summarized into a deployment-complete L2F SoH pipeline pseudocode in Algorithm 1, with the Training and adaptation hyperparameters listed in Table 3.

Table 3. Training and adaptation hyperparameters for the Lab-to-Field SoH pipeline.

Item	Setting
Lab resampling length	N_POINTS_LAB = 128
Lab batch size	BATCH_SIZE_LAB = 64
Field batch size	BATCH_SIZE_FIELD = 128
Lab optimizer/learning rate (LR)	Adam/LR_LAB = 1×10^{-3}
Field head LR	LR_FIELD_HEAD = 5×10^{-4}
Field backbone LR	LR_BACKBONE_FULL = 2×10^{-6}
Lab epochs/patience	EPOCHS_LAB = 60/PATIENCE = 12
Field head epochs	EPOCHS_FIELD_HEAD = 25
Field full fine-tune epochs	EPOCHS_FIELD_FULL = 70
CORAL weight and ramp	LAMBDA_CORAL = 0.25, CORAL_RAMP_EPOCHS = 10
Exponential Moving Average (EMA)/grad clip	EMA enabled, GRAD_CLIP = 1.0
Seed	SEED = 42

2.3.6. L2F SoH Pipeline Pseudocode

Algorithm 1 shows Leakage-Aware, Calibrated, and Deployment-Complete L2F SoH Estimation Pipeline.

Algorithm 1: Leakage-Aware, Calibrated, and Deployment-Complete L2F SoH Estimation Pipeline

Inputs:

Lab dataset $\mathcal{D}_S = \{(x^S, y^S)\}$ (cycle waveforms);

Field dataset $\mathcal{D}_T$ (engineered rows with sparse labels);

Deployment streams $\{\mathcal{D}_{\text{deploy}}^{(m)}\}_{m=1}^M$ (unseen at train time);

Window length L , stride s ; leakage-safe gap $g \geq L - 1$;

Backbone set $\mathcal{B} = \{\text{Fusion, TCN, Transformer, GRU, LSTM}\}$;

Calibration modes $\mathcal{C} = \{\text{none, Safe-Global, Context-Aware}\}$;

Alignment flag CORAL_on $\in \{0, 1\}$; seeds $\mathcal{S} = \{s_k\}_{k=1}^K$.

Outputs:

Row-level deployed predictions $\hat{y}_{\text{deploy}}(t)$ (raw) and $\hat{y}_{\text{deploy}}^*(t)$ (calibrated) for each stream;

Summary metrics (MAE/RMSE/ R^2) with mean $\pm$ std over seeds and bootstrap CIs.

Procedure:

1: For each backbone $b \in \mathcal{B}$ do7;

2: For each calibration mode $c \in \mathcal{C}$ do;

3: For each alignment flag CORAL_on $\in \{0, 1\}$ do;

4: For each seed $s \in \mathcal{S}$ do;

5: Set random seed to s .

6: Stage 1 (lab pretrain):

7: Resample each lab cycle waveform to length N ; derive physics channels $q(t)$, $e(t)$, and dV/dt .

8: Train backbone $f_{\theta}^{(b)}$ and lab head h_{lab} on $\mathcal{D}_S$ by minimizing

$$\mathcal{L}_{\text{LAB}} = \text{Smooth L1}(h_{\text{lab}}(f_{\theta}^{(b)}(a_{\text{lab}}(x^S))), y^S).$$

9: Save $f_{\theta}^{(b)}$, h_{lab} , and lab normalizer fitted on lab-train only.

10: Stage 2 (field windowing + leakage-safe split):

11: Form field windows $\{(X_i, y_i, e_i)\}$ of length L with stride s , indexed by window-end e_i .

12: Construct leakage-safe temporal splits on $\{e_i\}$: Train/Val/Test with enforced gap $g \geq L - 1$ between consecutive splits.

13: Fit field scaler on field-train only; transform field-val/test and all deployment streams with the same scaler.

14: Stage 2a (field head warm-up):

15: Initialize field adapter a_T and field head h_T ; freeze $f_{\theta}^{(b)}$.

16: Train $\{a_T, h_T\}$ on field-train windows using supervised loss $\mathcal{L}_{\text{sup}}$.

17: Stage 2b (field fine-tune with optional alignment):

18: Unfreeze $f_{\theta}^{(b)}$ (optionally partial unfreeze for Fusion); continue training with

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{CORAL}} \cdot \mathcal{L}_{\text{CORAL}}(Z_S, Z_T),$$

where $\lambda_{\text{CORAL}} = 0$ if CORAL_on = 0, otherwise ramped in early epochs.

19: Apply fixed stabilizers (held constant across baselines): EMA updates, gradient clipping, early stopping on field-val MAE/RMSE.

20: Save $f_{\theta}^{(b)}$, a_T , h_T , and the fitted field scaler.

21: Stage 2.5 (Safe Calibration; field-val only):

Algorithm 1: *Cont.*

22: Compute validation predictions $\hat{y}_{\text{val}}$ at window ends and form calibration pairs $(\hat{y}_{\text{val}}, y_{\text{val}})$.

23: Robustly filter pairs: (i) trim extreme prediction quantiles; (ii) remove residual outliers via MAD filtering.

24: Fit candidate calibrators $g \in \{\text{Identity, Ridge-Linear, Isotonic-Balanced}\}$.

25: If $c = \text{Context-Aware}$, fit context-conditioned isotonic using valid context IDs (e.g., checkup/condition); otherwise fall back to Safe-Global.

26: Select g^* using a do-no-harm holdout rule on an internal calibration holdout; if $c = \text{none}$ set $g^*(u) = u$.

27: Stage 3 (Deployment inference; inference-only streams):

28: For each deployment stream $\mathcal{D}_{\text{deploy}}^{(m)}$ do;

29: Slide windows of length L over rows; compute window-end predictions $\{\hat{y}_w\}$.

30: Convert to row-level prediction by overlap averaging:

$$\hat{y}_{\text{deploy}}(t) = \frac{1}{|W(t)|} \sum_{w \in W(t)} \hat{y}_w.$$

31: Apply calibration (if enabled): $\hat{y}_{\text{deploy}}^*(t) = g^*(\hat{y}_{\text{deploy}}(t))$ (or equivalently calibrate $\hat{y}_w$ then overlap-average, matching implementation).

32: Evaluate metrics on labeled rows only (MAE/RMSE/ R^2); store per-example absolute errors for bootstrap.

33: End for.

34: Stage 4 (Statistical reporting):

35: Aggregate results over seeds and streams; report mean $\pm$ std over $s \in \mathcal{S}$.

36: Compute bootstrap confidence intervals for MAE/RMSE from saved per-example absolute errors; optionally compute paired significance tests over matched runs.

37: End for;

38: End for;

39: End for;

40: End for.

2.4. Deployment Inference

In real-world BMS deployment, SoH must be estimated continuously from streaming data without introducing jitter, boundary artifacts, or latency that could affect control decisions. To obtain stable row-level trajectories suitable for threshold-based actions, sliding-window inference with overlap averaging is applied. For a sequence of length N , windows of length L are extracted with stride s , each window yields a window-end prediction, and the row-level estimate at index t is computed as the average of all window-end predictions whose corresponding windows cover t , as defined in Equation (9).

$$\hat{y}_t = (1 / |W(t)|) \sum_{w \in W(t)} \hat{y}_w. \quad (9)$$

where $\hat{y}$ is predicted SoH estimate, t is the row/time index, $W(t)$ is a set of window indices whose predictions contribute to index t , $|W(t)|$ is the number of windows in $W(t)$ (cardinality), w is an index of a particular window in $W(t)$, and $\hat{y}_w$ is the window-level prediction for window w .

Overlap averaging behaves like a temporal smoother while preserving responsiveness, and it produces row-level trajectories aligned with deployment expectations. The computational cost scales linearly with the number of windows and is practical for moderate L and $s = 1$ in offline processing or larger strides for embedded settings.

The following section reports expanded baselines and ablations under deployment-realistic overlap-averaged inference, including repeated-seed robustness, bootstrap uncertainty, and multi-deployment evaluation.

2.5. Baselines, Ablations, and Statistical Analysis

The evaluation includes five neural backbones—Fusion (proposed), TCN-only, Transformer-only, GRU, and LSTM—trained with identical optimization schedules and identical preprocessing. A classical Ridge baseline is added using standardized window-end features derived from the same leakage-safe split as the neural models. Alignment is ablated by toggling CORAL during field fine-tuning (off vs. on). Calibration is applied over three modes: no calibration (identity), Safe-Global calibration (do-no-harm selection among candidate calibrators), and Context-Aware calibration (condition/checkup-aware isotonic when context identifiers are valid and non-degenerate, otherwise safe global fall back).

All baselines use the same leakage-safe temporal split strategy, the same feature scaling policy, and the same deployment inference operator (overlap-averaged sliding-window inference). This ensures that differences are attributable to modeling choices (architecture/alignment/calibration) rather than evaluation artifacts.

Robustness is assessed over four random seeds and multiple deployment streams. Results are summarized using mean $\pm$ std over seeds and bootstrap confidence intervals (CIs) for MAE/RMSE computed from per-example absolute error arrays saved per run.

3. Results

Evaluation is first summarized on a completely unseen deployment stream with 2154 labeled rows. Mean absolute error (MAE), root-mean-square error (RMSE), and the coefficient of determination (R^2) are reported. Under overlap-averaged sliding-window inference, raw (uncalibrated) predictions achieve MAE = 0.0439, RMSE = 0.0501, and $R^2 = 0.7451$. After Safe Calibration (Isotonic-Balanced selected under the do-no-harm criterion), performance improves to MAE = 0.0188, RMSE = 0.0252, and $R^2 = 0.9357$, indicating that systematic output distortion (mis-scaling under L2Fshift) is a dominant contributor to deployment error. Beyond single-stream reporting, robustness is assessed across four random seeds and multiple unseen deployment streams, with statistical uncertainty summarized via mean $\pm$ std and bootstrap confidence intervals for MAE/RMSE computed from per-example absolute errors.

Figure 2a,b diagnose and correct the dominant L2F deployment failure mode in which predictions remain strongly monotonic yet become mis-scaled under shift. In Figure 2a, raw predictions exhibit a clear monotonic dependence on the ground-truth, indicating that the backbone retains degradation ordering in the deployment domain. However, the scatter deviates systematically from the identity line $\hat{y} = y$, consistent with nonlinear scaling bias (range compression), leading to overestimation in lower-to-mid SoH and underestimation at higher SoH. These results show that the most dominant error component is miscalibration (output warping) rather than loss of discriminative ordering. Figure 2b shows the corresponding calibrated outputs, where the selected monotone mapping yields improved alignment with $\hat{y} = y$ across the SoH range, demonstrating that Safe Calibration restores deployment-scale consistency while preserving the physically meaningful monotonic structure. This directly motivates Safe Calibration as an auditable reliability layer under domain shift by showing that calibration targets a specific, deployment-relevant failure mode rather than serving as unverified post-processing.

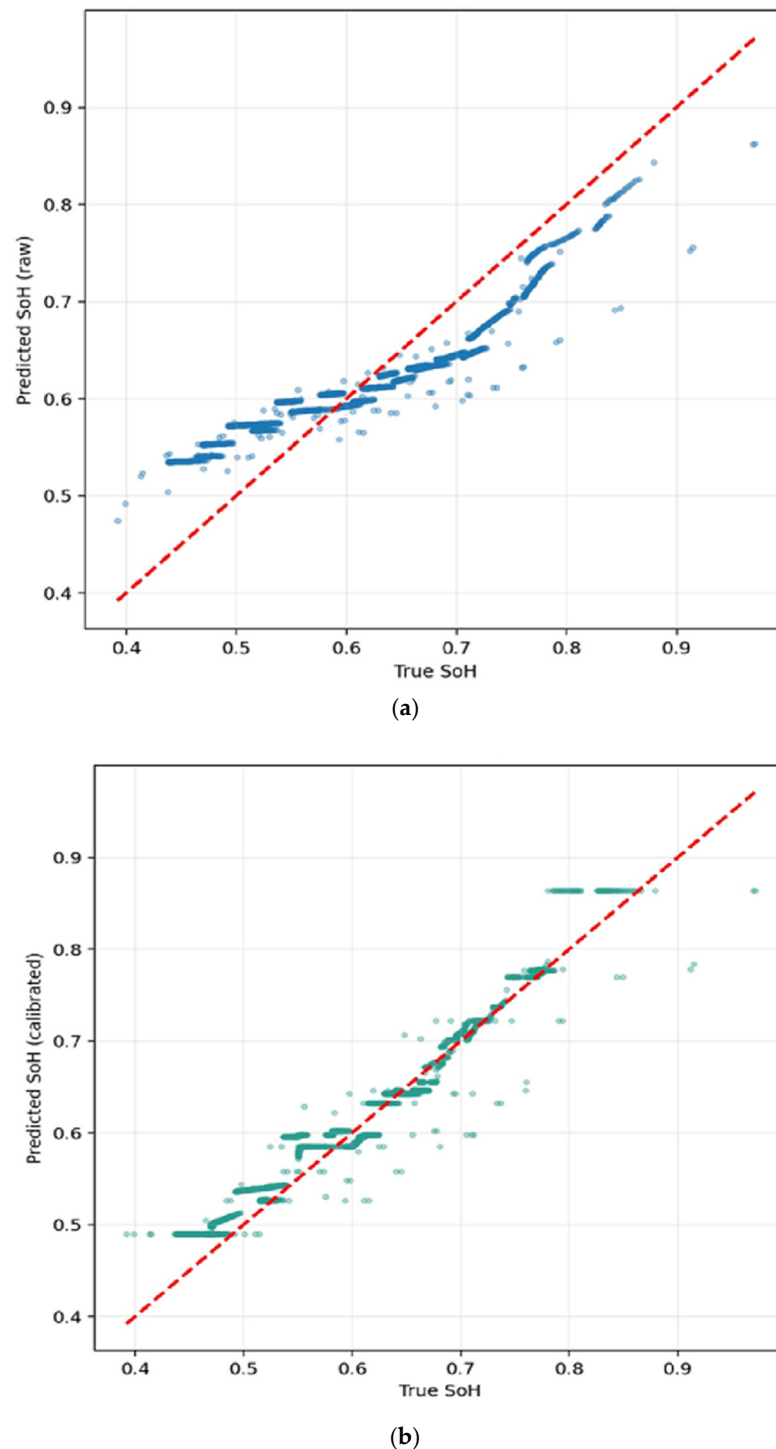


Figure 2. Raw vs. calibrated L2F SoH scaling under domain shift. (a) Raw overlap-averaged predictions versus ground-truth SoH on the unseen deployment stream; (b) Calibrated predictions on the same axes with the same identity reference line.

Ground-truth SoH (True) is plotted on the same axis with overlap-averaged raw predictions (Raw) and overlap-averaged calibrated predictions (Calibrated) and analyzed over the deployment timestamp axis. Safe Calibration is employed as a technique to reduce systematic scale distortion, while overlap averaging leads to temporally stable row-level estimates that can be used in BMS.

Figure 3 shows the deployment stage plotting the ground-truth SoH trajectory with row-level predictions produced by overlap-averaged sliding-window inference, before and

after Safe Calibration for comparison purposes. True SoH exhibits an expected decrease over time in a manner that is physically consistent with a degradation pattern that has short transients attributable to operational variability and measurement noise. The raw deployed trajectory follows the overall trend but exhibits a persistent scale bias consistent with L2F mis-scaling and range compression, which manifests as sustained estimation error across extended intervals. Safe Calibration is employed to ensure that the deployed trajectory aligns more closely with the ground-truth throughout the timestamp axis, thereby showing that the dominant deployment error component is systematic output distortion rather than loss of degradation ordering.

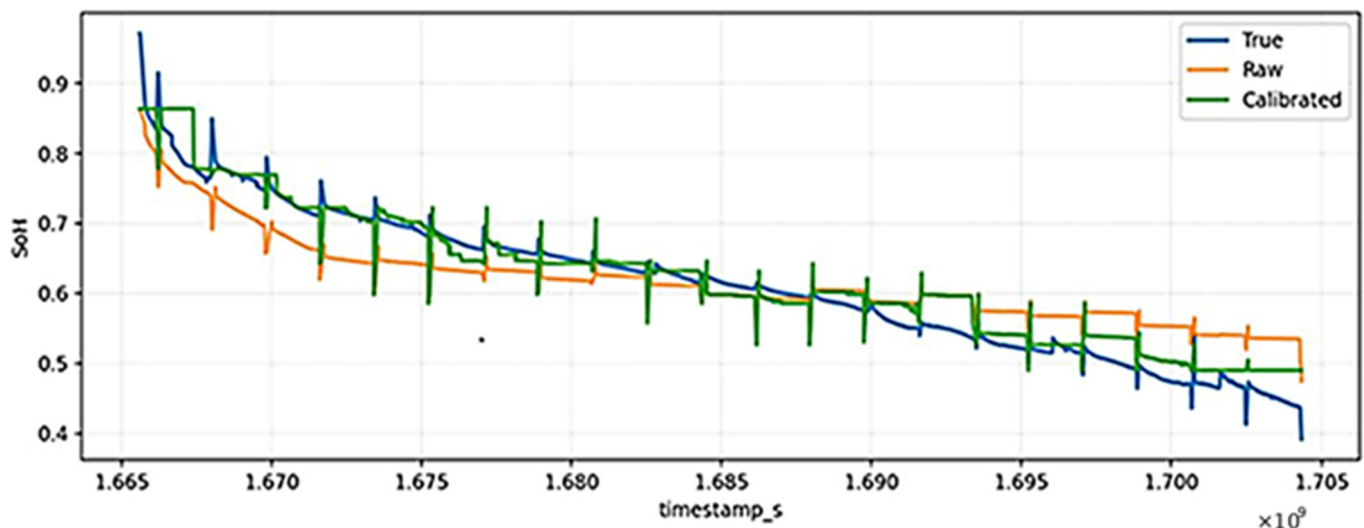


Figure 3. Row-level deployment trajectory on an unseen stream under overlap-averaged inference.

3.1. Baselines and Ablations

Table 4 reports results of controlled ablations that are designed to isolate (i) the contribution of the deployment inference operator, that is, window-end only versus overlap-averaged sliding-window inference, and (ii) Safe Calibration design choices, which are a choice of Safe-Global and Context-Aware. These ablations are then evaluated on the labeled subset of a completely unseen deployment data stream, P076_1_S15_C04. This dataset has been excluded from training, validation, and calibrator fitting, thereby enforcing strict cross-stream isolation. Overlap averaging yields a modest but consistent improvement, with the MAE improving from 0.0440 to 0.0439 and RMSE from 0.0505 to 0.0501, indicating reduced window-boundary artifacts and improved row-level stability, but it does not resolve the dominant failure mode under L2F shift. In contrast, Safe Calibration provides the most noticeable correction, that is, Safe-Global isotonic improves the performance of MAE to 0.0415 and RMSE to 0.0473), while the proposed Context-Aware Safe Calibration achieves the largest gain, MAE = 0.0142 and RMSE = 0.0206, consistent with the correction of systematic output-scale warping while preserving monotonic degradation ordering [29]. The row containing MAE = 0.0188 and RMSE = 0.0252 corresponds to the calibrated outputs saved by the run configuration and is retained to document reproducibility of the pipeline outputs. Table 4 is intentionally a single-stream controlled ablation with broader robustness evidence being provided by the expanded evaluation protocol, which additionally reports repeated-seed statistics and results on two further unseen deployment streams.

Table 4. Summarizes the main ablation outcomes.

Variant	MAE	RMSE	Notes
Window-end only (no overlap)	0.0440	0.0505	Ablates overlap inference
Overlap-averaged inference (raw)	0.0439	0.0501	Adds overlap inference
Safe calibration (global isotonic)	0.0415	0.0473	Ablates context-aware calibration
Safe calibration (context-aware, proposed)	0.0142	0.0206	Adds checkup/condition-aware calibration
(Reference) Calibrated output saved by run	0.0188	0.0252	From the provided output CSV

3.2. Architecture Baselines: Fusion vs. TCN-Only vs. Transformer-Only vs. GRU vs. LSTM

Across the aggregated multi-stream, repeated-seed evaluation, the architecture comparison indicates that the Fusion backbone is competitive relative to recurrent and attention-only baselines under matched training and inference conditions. For the Fusion configuration with Context-Aware calibration (CORAL off), the aggregated row-level deployment MAE is 0.0547 ± 0.0291 , and the field test window MAE after Context-Aware calibration is 0.0205 ± 0.0008 . The strongest aggregated deployment MAE among evaluated backbones is observed for GRU with Safe-Global calibration (CORAL off), achieving 0.0486 ± 0.0262 , while TCN-only remains less competitive under the same evaluation protocol. These results indicate that the end-to-end pipeline is not dependent on a single backbone choice and that architectural differences are comparatively smaller than the calibration effect under the present domain-shift regime. The Fusion design is retained because it provides consistently strong performance across streams while offering complementary inductive biases (local multi-scale dynamics via TCN and long-range dependencies via attention). Qualitative alignment between predicted and true trajectories across backbones is illustrated using synchronized time-series overlays, where Fusion (raw/calibrated) is shown against thin reference curves from other backbones under the same inference operator, highlighting systematic bias reduction after calibration.

3.3. Classical Baseline: Ridge (Basic) Raw vs. Calibrated

A Ridge regression baseline is included to test whether post hoc calibration alone can account for the observed gains. Without calibration, Ridge can exhibit strong global mis-scaling under L2F shift (offset/gain errors), yielding high window-level MAE (≈ 0.4785 in the aggregated evaluation) despite partial preservation of degradation ordering. After Safe/Context-Aware calibration, Ridge improves substantially (≈ 0.0157 window-level MAE), demonstrating that monotone post hoc mapping can correct severe systematic scale distortion when the predictor preserves ranking. However, the deployed objective in this work is row-level trajectory stability under overlap-averaged inference on unseen streams, and the deep pipeline remains the primary method because it learns sequence-conditioned representations and produces deployment-consistent outputs under the same overlap-averaging operator used for evaluation, whereas a window-end linear baseline is more sensitive to feature-definition mismatch and does not encode temporal context.

3.4. Calibration Ablation: None vs. Safe-Global vs. Context-Aware

Calibration is the dominant contributor to accuracy gains under shift. For Fusion (CORAL off), calibration reduces field test window-level MAE from 0.0350 (raw) to 0.0205

(Context-Aware calibrated), while the row-level deployment MAE reduces from 0.0632 (raw) to 0.0547 (Context-Aware calibrated). Safe-Global and Context-Aware modes yield similar aggregate deployment performance, with Context-Aware providing additional benefit when context identifiers are valid and sufficiently populated; otherwise, the method falls back to Safe-Global selection to avoid degradation under the do-no-harm rule.

3.5. Alignment Ablation: CORAL off vs. on

The CORAL on/off ablation isolates alignment effects from calibration. On average, enabling CORAL yields modest improvements in field window-level error, while the row-level deployment metric shows limited sensitivity and can vary by backbone. This indicates that, under the present dataset conditions and leakage-safe split, calibration and the deployment inference operator contribute more strongly to final deployment accuracy than second-order alignment alone. CORAL is retained as an optional regularizer because it can stabilize transfer when the second-order mismatch is larger, when target supervision is sparser, or when operating conditions induce stronger covariance shift.

3.6. Robustness: Repeated Seeds and Bootstrap Confidence Intervals

Robustness is quantified using four random seeds and multiple deployment streams. Mean $\pm$ std values summarize optimization variability across seeds, and bootstrap confidence intervals (CIs) for MAE/RMSE are computed from per-example absolute error arrays saved per run and aggregated within each stream/configuration to quantify uncertainty in error estimates. This reporting supports conclusions about effect consistency across deployment streams beyond single-run outcomes and directly addresses statistical rigor requirements under domain shift.

3.7. Analysis of Sliding-Window Inference and Safe Calibration

Raw row-level predictions on the deployment stream are compressed relative to the labeled SoH range, consistent with the curvature observed in Figure 2a. This behavior indicates that the transfer-learned representation retains degradation ordering under L2F shift, but the output scale is systematically distorted (nonlinear warping/range compression). Safe Calibration directly targets this distortion. On the 2154 labeled deployment rows, overlap-averaged raw inference achieves MAE = 0.0439, RMSE = 0.0501, and $R^2 = 0.7451$, whereas Safe Calibration (Isotonic-Balanced selected) improves MAE = 0.0188, RMSE = 0.0252, and $R^2 = 0.9357$, consistent with the improved diagonal alignment from Figure 2a to Figure 2b. Importantly, the selection rule applies calibration only if it reduces held-out error (do-no-harm), reflecting established cautions that post hoc mappings can degrade performance if fitted on unstable pairs or leverage-point outliers. The result, therefore, addresses the concern that “calibration dominates improvement” by demonstrating that calibration is used as an auditable reliability layer correcting a specific deployment-relevant failure mode (systematic mis-scaling under shift) while preserving monotonic degradation structure.

3.8. Label-Efficiency of Safe Calibration

Calibration-label efficiency is evaluated by fitting Safe Calibration using an increasing number of labeled calibration pairs from stream P001_1_S01_C10_MULTIPLE (chronological prefixes) and measuring MAE on labeled deployment rows. The dashed line in Figure 4 denotes the raw (uncalibrated) overlap-averaged MAE computed on the same evaluation subset.

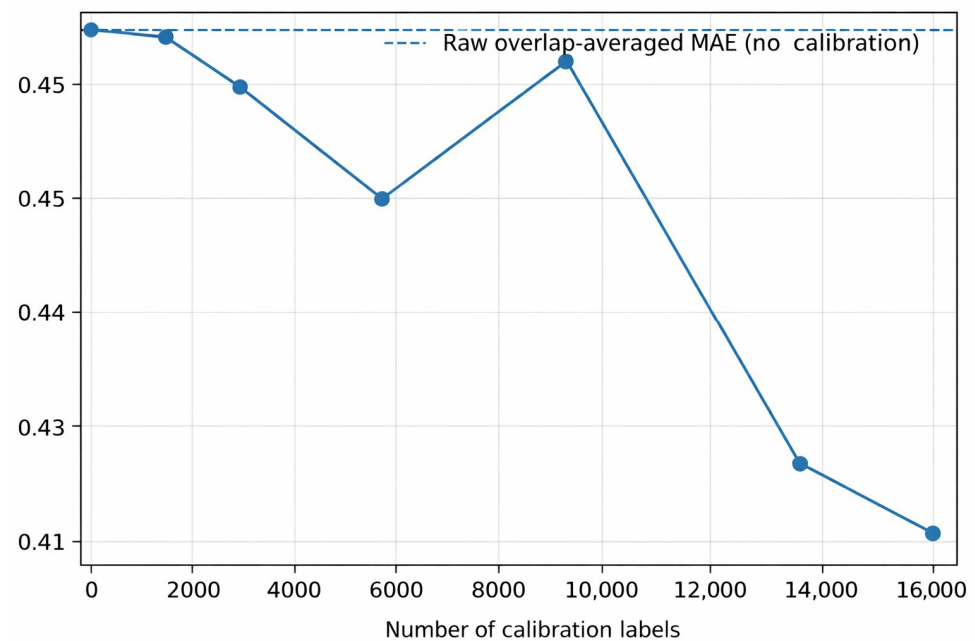


Figure 4. Label efficiency of Safe Calibration (lower is better).

3.9. Drift Stability of the Calibrator

Figure 5 evaluates drift stability for the Fusion backbone by partitioning each deployment stream's labeled rows into four chronological quartiles and reporting performance within each time chunk. The raw overlap-averaged predictor exhibits time-varying error, consistent with gradual distribution shift and operating-regime changes along the deployment horizon. After Safe Calibration, MAE is reduced in each quartile, and the calibrated curve remains below the raw reference level, indicating that the monotone mapping does not over-correct early segments at the expense of later segments (or vice versa). The plot aggregates results across multiple deployment streams and four random seeds, with uncertainty shown as the mean $\pm$ std; the persistence of calibrated improvement across quartiles supports the claim that Safe Calibration functions as a robust reliability layer under realistic drift rather than as a stream-specific post-processing fit.

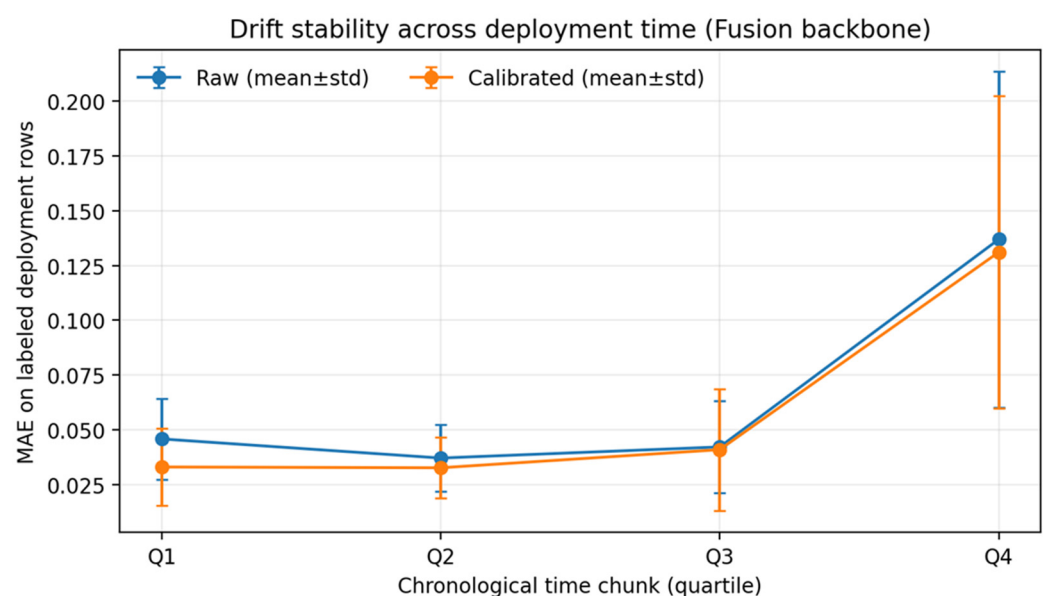


Figure 5. Drift stability of Safe Calibration under temporal drift.

Labeled deployment rows are partitioned into four (4) separate chronological quartiles, that is, Q1 to Q4, using timestamp, and the MAE is reported using the mean $\pm$ std across deployment streams and random seeds under overlap-averaged inference. The dashed horizontal line denotes the raw overlap-averaged MAE (no calibration) computed on the same evaluation subset.

Table 5 shows the corresponding time-quartile error breakdown for the Fusion backbone, that is, the mean $\pm$ std across deployment streams and random seeds. It also shows that quartile variability increases, reflecting harder regimes and cross-stream heterogeneity, while the calibrated metric remains consistently improved relative to raw outputs.

Table 5. Time-quartile error breakdown for Fusion on deployment streams (mean $\pm$ std over deployment streams and random seeds); raw vs. Safe-Calibrated predictions.

Time Chunk (Quartile)	MAE (Raw), Mean $\pm$ Std	MAE (Calibrated), Mean $\pm$ Std	n (Runs)
Q1	0.0458 $\pm$ 0.0185	0.0329 $\pm$ 0.0176	72
Q2	0.0370 $\pm$ 0.0152	0.0326 $\pm$ 0.0140	72
Q3	0.0421 $\pm$ 0.0211	0.0409 $\pm$ 0.0278	72
Q4	0.1369 $\pm$ 0.0767	0.1311 $\pm$ 0.0714	72

3.10. Sliding-Window Geometry Across Streams

In Table 6, a summary of sliding-window inference geometry and labeled evaluation subsets results is reported. These results are reported across the calibration stream and all deployment streams used for row-level reporting, where, for each stream, a length- L window with stride, s , and overlap averaging yields near-uniform interior coverage close to L contributions per row, with short ramp-up and ramp-down regions at stream boundaries. Overlap-averaged sliding-window inference acts as a parameter-free temporal smoother, reducing per-window prediction jitter by averaging multiple overlapping estimates for each time step. We use window length $L = 20$ and stride $s = 1$ with the mean overlap (coverage) per time step as

$$\bar{c} = \frac{W L}{N} \quad (10)$$

where W is the number of windows and N is the number of samples (rows).

Table 6. Summary statistics of sliding-window geometry across calibration and deployment streams.

Stream	Rows N	Windows ($L = 20, s = 1$)	Mean Coverage
Calibration (FIELD)			
P001_1_S01_C10_MULTIPLE	129,081	129,062	19.997
Deployment P076_1_S15_C04	50,579	50,560	19.992
Deployment P073_2_S12_C02	23,126	23,107	19.984
Deployment P073_3_S14_C01	25,346	25,327	19.985
Deployment (aggregate across streams)	99,051	98,994	19.988

3.11. Error Distribution Across Runs

Figure 6 is a summary of the results of the distribution of per-example absolute errors $|y - \hat{y}|$. These errors are aggregated across the full experimental grid that includes four random seeds and multiple deployment streams. Both field evaluation (window-level) and deployment-style infer evaluation, which is row-level and overlap-averaged, are reported. Safe Calibration shifts the deep model's error distribution downward and reduces

dispersion relative to uncalibrated outputs, indicating improved accuracy and improved run-to-run stability under L2F shift. When compared to the classical Ridge baseline, the calibrated deep pipeline maintains lower central error and a tighter spread, while the baseline exhibits substantially higher error and variability, consistent with stronger systematic distortion under shift. Overall, the distributional view corroborates the mean $\pm$ std and bootstrap CI summaries by showing that gains persist across seeds and deployment streams rather than being driven by isolated runs.

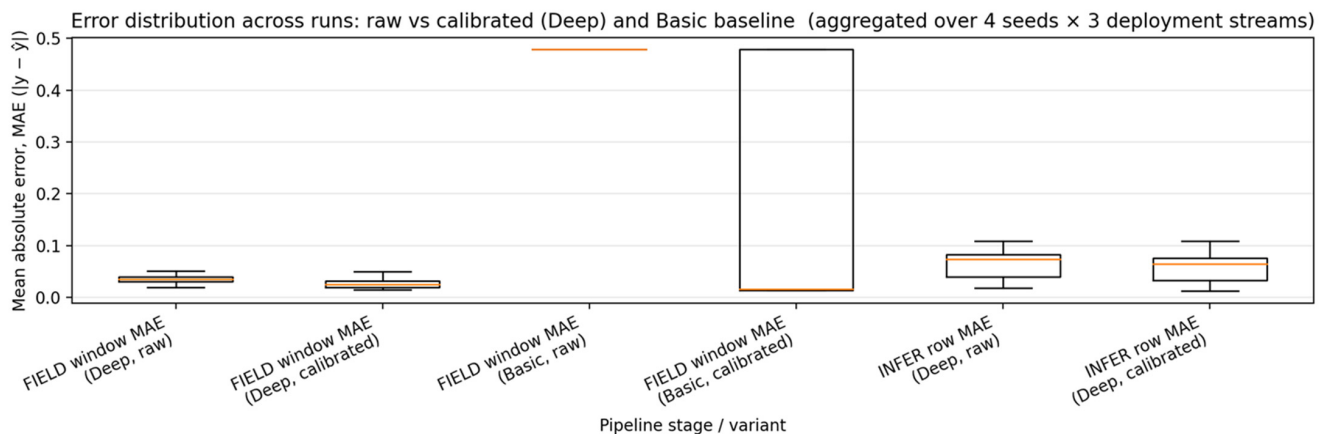


Figure 6. Error distribution across runs, raw vs. calibrated (Deep) and Ridge baseline.

3.12. Multi-Stream Deployment Evaluation

Table 7 reports the per-stream mean $\pm$ standard deviation over four random seeds for deep architecture baselines (Fusion, TCN-only, Transformer-only, GRU, and LSTM) evaluated under identical leakage-safe temporal splitting, identical preprocessing/scaling, and the same deployment-realistic overlap-averaged inference operator. Results are reported for field test windows (window-level) and deployment-style inference (row-level). Across streams, calibration reduces error and typically reduces dispersion, indicating that the dominant failure mode is systematic output warping rather than loss of monotonic ordering. Differences between backbones are present but generally secondary to calibration under the present domain-shift regime, supporting the interpretation that deployment reliability is not dependent on a single architecture choice.

Table 8 summarizes the classical Ridge regression baseline trained on standardized window-end features derived from the same leakage-safe field split, reported in raw and calibrated forms. This baseline evaluates the “basic model + calibration” hypothesis by isolating the role of post hoc monotone mapping. While calibration substantially improves Ridge at the window level, the deep pipeline remains the primary method because it learns sequence-conditioned representations compatible with overlap-averaged deployment inference and demonstrates robustness across multiple deployment streams under repeated seeds.

Table 9 reports bootstrap 95% confidence intervals (CIs) for MAE and RMSE on both field test (window-level) and deployment inference (row-level) evaluations. CIs are computed by resampling the saved per-example absolute error arrays within each deployment stream and configuration, aggregated over repeated seeds. Across streams, Safe Calibration consistently shifts the intervals downward relative to raw predictions, indicating that reported gains persist under seed variability and stream-to-stream heterogeneity. Remaining CI overlap among backbones in some streams suggests that architectural differences are secondary to the calibration effect under the present domain-shift conditions.

Table 7. Multi-stream mean $\pm$ std across seeds: deep backbones (field window-level and deployment infer row-level MAE).

Deployment Stream	Backbone	Field Win MAE (Raw)	Field Win MAE (Cal)	Infer Row MAE (Raw)	Infer Row MAE (Cal)
P073_2_S12_C02	Fusion	0.0322 $\pm$ 0.0097	0.0216 $\pm$ 0.0021	0.0828 $\pm$ 0.0111	0.0723 $\pm$ 0.0124
P073_2_S12_C02	GRU	0.0330 $\pm$ 0.0024	0.0202 $\pm$ 0.0052	0.0729 $\pm$ 0.0051	0.0621 $\pm$ 0.0069
P073_2_S12_C02	LSTM	0.0288 $\pm$ 0.0091	0.0189 $\pm$ 0.0028	0.0792 $\pm$ 0.0063	0.0684 $\pm$ 0.0105
P073_2_S12_C02	TCN	0.0407 $\pm$ 0.0072	0.0275 $\pm$ 0.0039	0.0838 $\pm$ 0.0073	0.0729 $\pm$ 0.0177
P073_2_S12_C02	Transformer	0.0357 $\pm$ 0.0039	0.0204 $\pm$ 0.0025	0.0754 $\pm$ 0.0074	0.0627 $\pm$ 0.0152
P073_3_S14_C01	Fusion	0.0322 $\pm$ 0.0097	0.0216 $\pm$ 0.0021	0.0878 $\pm$ 0.0129	0.0776 $\pm$ 0.0133
P073_3_S14_C01	GRU	0.0330 $\pm$ 0.0024	0.0202 $\pm$ 0.0052	0.0781 $\pm$ 0.0065	0.0666 $\pm$ 0.0084
P073_3_S14_C01	LSTM	0.0288 $\pm$ 0.0091	0.0189 $\pm$ 0.0028	0.0836 $\pm$ 0.0050	0.0721 $\pm$ 0.0090
P073_3_S14_C01	TCN	0.0407 $\pm$ 0.0072	0.0275 $\pm$ 0.0039	0.0901 $\pm$ 0.0083	0.0807 $\pm$ 0.0203
P073_3_S14_C01	Transformer	0.0357 $\pm$ 0.0039	0.0204 $\pm$ 0.0025	0.0801 $\pm$ 0.0068	0.0648 $\pm$ 0.0106
P076_1_S15_C04	Fusion	0.0322 $\pm$ 0.0097	0.0216 $\pm$ 0.0021	0.0296 $\pm$ 0.0084	0.0201 $\pm$ 0.0016
P076_1_S15_C04	GRU	0.0330 $\pm$ 0.0024	0.0202 $\pm$ 0.0052	0.0300 $\pm$ 0.0033	0.0199 $\pm$ 0.0075
P076_1_S15_C04	LSTM	0.0288 $\pm$ 0.0091	0.0189 $\pm$ 0.0028	0.0337 $\pm$ 0.0175	0.0285 $\pm$ 0.0210
P076_1_S15_C04	TCN	0.0407 $\pm$ 0.0072	0.0275 $\pm$ 0.0039	0.0373 $\pm$ 0.0070	0.0269 $\pm$ 0.0046
P076_1_S15_C04	Transformer	0.0357 $\pm$ 0.0039	0.0204 $\pm$ 0.0025	0.0352 $\pm$ 0.0021	0.0205 $\pm$ 0.0012

Table 8. Classical baseline (Ridge) on the FIELD test split (mean $\pm$ std across seeds).

Baseline	FIELD Win MAE (Raw)	FIELD Win MAE (Cal)
Ridge (window-end features)	0.4785 $\pm$ 0.0000	0.0157 $\pm$ 0.0000

To complement the repeated-seed mean $\pm$ standard deviation and bootstrap confidence intervals, a standalone paired significance analysis as shown in Table 10 was performed on calibrated row-level deployment MAE and RMSE using matched seed–stream evaluations. The results indicate that Fusion performs significantly better than TCN, while GRU performs significantly better than Fusion. In contrast, Transformer and LSTM do not differ from Fusion in a statistically reliable manner. Specifically, relative to Fusion, TCN exhibited significantly higher error on both MAE and RMSE, whereas GRU exhibited significantly lower error. The LSTM comparison was not statistically conclusive, and the Transformer comparison, although showing a small mean advantage in Δ MAE, was not supported by the paired sign-test p -values. Taken together, these results show that the statistical evaluation in this study extends beyond descriptive performance summaries by incorporating repeated-seed variability, cross-stream robustness, bootstrap confidence intervals, and explicit pairwise significance testing.

Table 9. Bootstrap 95% confidence intervals (CIs) for MAE and RMSE per deployment stream under the context-aware safe calibration setting (CORAL on).

Deployment Stream	Backbone	Field Win MAE CI (Raw)	Field Win RMSE CI (Raw)	Field Win MAE CI (Cal)	Field Win RMSE CI (Cal)	Infer Row MAE CI (Raw)	Infer Row RMSE CI (Raw)	Infer Row MAE CI (Cal)	Infer Row RMSE CI (Cal)
cell_eocv2_P073_2_S12_C02.csv	Fusion	[0.0317, 0.0327]	[0.0399, 0.0412]	[0.0212, 0.0219]	[0.0279, 0.0294]	[0.0801, 0.0851]	[0.1100, 0.1172]	[0.0698, 0.0748]	[0.1007, 0.1074]
cell_eocv2_P073_2_S12_C02.csv	GRU	[0.0326, 0.0334]	[0.0385, 0.0394]	[0.0198, 0.0205]	[0.0254, 0.0265]	[0.0706, 0.0751]	[0.0985, 0.1050]	[0.0599, 0.0642]	[0.0880, 0.0946]
cell_eocv2_P073_2_S12_C02.csv	LSTM	[0.0283, 0.0292]	[0.0362, 0.0374]	[0.0186, 0.0192]	[0.0232, 0.0244]	[0.0768, 0.0813]	[0.1034, 0.1094]	[0.0663, 0.0703]	[0.0903, 0.0960]
cell_eocv2_P073_2_S12_C02.csv	TCN	[0.0401, 0.0412]	[0.0482, 0.0494]	[0.0271, 0.0280]	[0.0334, 0.0345]	[0.0811, 0.0865]	[0.1137, 0.1210]	[0.0701, 0.0757]	[0.1076, 0.1152]
cell_eocv2_P073_2_S12_C02.csv	Transformer	[0.0352, 0.0361]	[0.0418, 0.0429]	[0.0201, 0.0208]	[0.0262, 0.0273]	[0.0728, 0.0778]	[0.1021, 0.1084]	[0.0604, 0.0650]	[0.0898, 0.0959]
cell_eocv2_P073_3_S14_C01.csv	Fusion	[0.0317, 0.0327]	[0.0399, 0.0412]	[0.0212, 0.0219]	[0.0279, 0.0294]	[0.0852, 0.0903]	[0.1170, 0.1243]	[0.0751, 0.0799]	[0.1068, 0.1137]
cell_eocv2_P073_3_S14_C01.csv	GRU	[0.0326, 0.0334]	[0.0385, 0.0394]	[0.0198, 0.0205]	[0.0254, 0.0265]	[0.0757, 0.0805]	[0.1063, 0.1127]	[0.0645, 0.0688]	[0.0947, 0.1008]
cell_eocv2_P073_3_S14_C01.csv	LSTM	[0.0283, 0.0292]	[0.0362, 0.0374]	[0.0186, 0.0192]	[0.0232, 0.0244]	[0.0812, 0.0859]	[0.1095, 0.1158]	[0.0700, 0.0741]	[0.0956, 0.1014]
cell_eocv2_P073_3_S14_C01.csv	TCN	[0.0401, 0.0412]	[0.0482, 0.0494]	[0.0271, 0.0280]	[0.0334, 0.0345]	[0.0874, 0.0927]	[0.1227, 0.1295]	[0.0779, 0.0835]	[0.1182, 0.1256]
cell_eocv2_P073_3_S14_C01.csv	Transformer	[0.0352, 0.0361]	[0.0418, 0.0429]	[0.0201, 0.0208]	[0.0262, 0.0273]	[0.0779, 0.0823]	[0.1052, 0.1110]	[0.0627, 0.0669]	[0.0918, 0.0973]
cell_eocv2_P076_1_S15_C04.csv	Fusion	[0.0317, 0.0327]	[0.0399, 0.0412]	[0.0212, 0.0219]	[0.0279, 0.0294]	[0.0291, 0.0301]	[0.0377, 0.0391]	[0.0197, 0.0205]	[0.0268, 0.0281]
cell_eocv2_P076_1_S15_C04.csv	GRU	[0.0326, 0.0334]	[0.0385, 0.0394]	[0.0198, 0.0205]	[0.0254, 0.0265]	[0.0296, 0.0304]	[0.0352, 0.0362]	[0.0196, 0.0203]	[0.0264, 0.0278]
cell_eocv2_P076_1_S15_C04.csv	LSTM	[0.0283, 0.0292]	[0.0362, 0.0374]	[0.0186, 0.0192]	[0.0232, 0.0244]	[0.0332, 0.0342]	[0.0426, 0.0440]	[0.0279, 0.0291]	[0.0388, 0.0402]
cell_eocv2_P076_1_S15_C04.csv	TCN	[0.0401, 0.0412]	[0.0482, 0.0494]	[0.0271, 0.0280]	[0.0334, 0.0345]	[0.0367, 0.0379]	[0.0447, 0.0459]	[0.0264, 0.0274]	[0.0340, 0.0350]
cell_eocv2_P076_1_S15_C04.csv	Transformer	[0.0352, 0.0361]	[0.0418, 0.0429]	[0.0201, 0.0208]	[0.0262, 0.0273]	[0.0347, 0.0357]	[0.0424, 0.0435]	[0.0201, 0.0209]	[0.0267, 0.0277]

Table 10. Paired significance analysis for calibrated row-level deployment performance relative to the Fusion backbone across matched seed–stream evaluations.

Comparison	<i>n</i>	Mean Δ MAE (Comp—Fusion)	Mean Δ RMSE (Comp—Fusion)	<i>p</i> -Value (MAE)	<i>p</i> -Value (RMSE)	Bootstrap 95% CI for Δ MAE
TCN vs. Fusion	72	0.0015	0.0050	0.0444	0.0245	[−0.0005, 0.0034]
Transformer vs. Fusion	72	−0.0020	−0.0027	1.0000	0.9063	[−0.0038, −0.0002]
GRU vs. Fusion	72	−0.0065	−0.0082	0.0013	<0.001	[−0.0080, −0.0049]
LSTM vs. Fusion	72	−0.0002	−0.0041	0.9063	0.1945	[−0.0019, 0.0014]

Across the multi-deployment evaluation (Tables 7–9), shows that performance is not uniform across streams, and at least one stream exhibits materially higher row-level error and lower R^2 than the others. Such heterogeneity is expected under deployment because labeled rows can be sparse and regime-dependent, the effective SoH dynamic range can differ by stream, and label noise and feature/label-definition drift can be stream-specific. In this setting, R^2 is particularly sensitive to reduced label variance and may decrease even when absolute error metrics remain informative for deployment. The repeated-seed, multi-stream protocol therefore functions as a robustness stress test: all configurations are evaluated under identical leakage-safe splits, identical preprocessing, and the same overlap-averaged inference operator, while mean $\pm$ std and bootstrap confidence intervals quantify both central tendency and uncertainty across operating conditions. Importantly, Safe Calibration consistently reduces systematic bias and error dispersion relative to raw outputs across streams, while cross-stream variability highlights intrinsically harder regimes and motivates stream-aware monitoring and periodic recalibration in practice.

3.13. Practical Deployment Implications

In practical battery-management workflows, SoH estimates are consumed by threshold-based decision logic (e.g., decreasing policies, maintenance scheduling, and warranty triggers), where systematic bias under distribution shift can be more consequential than random error. The multi-stream, repeated-seed results indicate that L2Fshift often preserves degradation ordering while distorting output scale, producing range compression that accumulates into sustained underestimation in row-level trajectories. The pipeline, therefore, combines overlap-averaged inference, which stabilizes row-level trajectories and mitigates window-boundary artifacts, with Safe Calibration, which targets deployment-induced mis-scaling via a monotone mapping selected under a do-no-harm criterion. Cross-stream variability, quantified through the mean $\pm$ std and bootstrap confidence intervals, highlights the expected heterogeneity of deployment regimes and motivates stream-aware monitoring and periodic recalibration as part of operational governance rather than reliance on a single best-case score.

4. Discussion

The results support a consistent interpretation of L2F transfer under domain shift. Across streams, raw predictions preserve a strong monotonic relationship with ground-truth SoH, indicating that the backbone with alignment-regularized adaptation retains transferable degradation ordering in the deployment domain. However, the raw scaling plots and time-series traces show systematic output mis-scaling (range compression/warping), which is a deployment-critical failure mode because SoH is typically consumed through thresholds and margins in BMS logic. In such settings, systematic bias can be more operationally harmful than zero-mean noise: persistent underestimation can trigger overly

conservative derating or premature maintenance actions, while overestimation risks unsafe operation and incorrect warranty accounting.

Safe Calibration is introduced specifically to target this dominant failure mode, rather than as an unverified post-processing step. Calibration pairs $(\hat{y}, y)$ formed from held-out field validation predictions, are robustified before fitting by (i) trimming extreme predicted-value quantiles and (ii) removing residual outliers using MAD filtering. These steps reduce leverage points and heteroskedastic tails that can otherwise force unstable monotone fits (e.g., stepwise plateaus or spurious extrapolation), thereby improving mapping stability under realistic label sparsity and noise. Isotonic regression is used as the primary nonparametric calibrator because it enforces monotonicity—consistent with physically meaningful degradation ordering—while correcting arbitrary nonlinear distortions without assuming a parametric error model. To further reduce overfitting in dense SoH regions and mitigate plateau artifacts, the isotonic fit is bin-balanced (Isotonic-Balanced), and the final mapping is selected under a strict do-no-harm criterion evaluated on an internal calibration holdout. This combination yields an auditable reliability layer: the mapping is applied only when it measurably improves held-out error, limiting the risk of calibration-induced degradation.

The empirical pattern observed here is that domain shift often preserves ranking while distorting scale; consequently, a monotone calibrator can recover substantial accuracy when the representation already encodes transferable ordering. This is not a weakness of the backbone, but evidence that the dominant residual error under shift is systematic mis-scaling rather than loss of discriminative temporal structure. The baseline suite (Fusion, TCN-only, Transformer-only, GRU, LSTM) and the classical Ridge comparator further clarify this point: calibration can correct large global bias for models that preserve ordering, but the deep pipeline remains preferable for deployment because it produces sequence-conditioned representations and stable row-level outputs under overlap-averaged inference, which is the operational operator used for evaluation and BMS consumption. In addition, the alignment ablation (CORAL on/off) indicates that second-order alignment can provide modest improvements depending on stream conditions, but the largest and most consistent deployment benefit in this dataset arises from the reliability layer (Safe Calibration) combined with a deployment-realistic inference operator (overlap averaging).

Robustness and statistical validity under deployment variability are emphasized rather than peak single-stream R^2 . Repeated-seed reporting (mean $\pm$ std) quantifies optimization variability, and bootstrap confidence intervals for MAE/RMSE quantify uncertainty using per-example absolute error arrays aggregated across runs. The multi-stream evaluation reveals non-uniform difficulty across deployment streams, which is expected in practice because labeled rows can be sparse and regime-dependent, the effective SoH dynamic range can differ by stream, and label noise or feature/label-definition drift can be stream-specific. In such settings, R^2 can decrease sharply when label variance is reduced, even when MAE/RMSE remain deployment informative. Reporting uncertainty and cross-stream breakdowns, therefore, provides a reviewer-proof basis for claims about reliability under L2F shift, and it identifies intrinsically harder regimes where monitoring and periodic recalibration are operationally warranted.

Several limitations remain. First, the current study evaluates a specific lab source (NASA cells) and engineered field-like EOCV2 descriptors, and performance can be constrained by feature-definition mismatch and label sparsity in certain operating regimes. Second, although the pipeline is deployment-complete and leakage-safe, it remains an offline calibration approach; continuous drift may require periodic recalibration or online updates with explicit safeguards. Third, context variables used for conditioning can be imbalanced, which can limit the effectiveness of context-aware calibration in rare regimes. Finally, while the target application is transportation electrification, the same reliability-

oriented pipeline structure—transfer with alignment, deployment-realistic inference, and auditable monotone calibration—can extend to other storage settings (e.g., stationary energy storage systems, UPS, solar storage) provided that domain-specific feature engineering, leakage-safe splitting, and calibration hygiene are maintained under the relevant operating profiles and sensing constraints.

5. Conclusions

This paper presented a deployment-complete L2F transfer learning framework for battery SoH estimation in transportation electrification. The proposed pipeline integrates a gated TCN–Transformer fusion backbone, domain-specific adapters and heads, alignment-regularized adaptation (optional CORAL), streaming-compatible overlap-averaged inference, and Safe Calibration (Isotonic-Balanced selected under a strict do-no-harm rule). The results confirm that, under L2F shift, predictors can preserve degradation ordering yet become mis-scaled, and that correcting this systematic output distortion through an auditable monotone calibration layer is critical for trustworthy SoH estimation suitable for BMS consumption.

Across a comprehensive evaluation grid of 360 runs ($=5$ backbones $\times$ 3 calibration modes $\times$ 2 alignment settings (CORAL off/on) $\times$ 4 random seeds $\times$ 3 unseen deployment streams), consistent deployment-oriented improvements are observed when Safe Calibration and overlap-averaged sliding-window inference are enabled. In particular, Safe Calibration yields a marked reduction in window-level field test error, with representative configurations reducing RMSE from approximately 0.043–0.047 (raw) to approximately 0.027–0.030 (calibrated) when aggregated across seeds and deployment streams (mean $\pm$ std reported). At the row-level deployment inference stage (overlap-averaged), calibration yields smaller but consistent gains, reducing aggregated RMSE from approximately 0.087–0.089 (raw) to approximately 0.079 (calibrated), indicating that (i) calibration improves the mapping from model score to SoH under domain shift and (ii) overlap averaging stabilizes row-level outputs under sequential inference.

Robustness is explicitly quantified via mean $\pm$ std over repeated seeds and bootstrap confidence intervals for MAE/RMSE, reported both per deployment stream and aggregated across streams, enabling direct assessment of variability under stochastic optimization and heterogeneous deployment conditions. The bootstrap confidence intervals indicate that calibrated error reductions are not isolated to a single run, while cross-stream summaries reveal non-trivial deployment heterogeneity, motivating reporting aggregated performance alongside per-stream breakdowns rather than relying on single-stream conclusions. Architecture-level comparisons under the same training, calibration, and inference protocol further indicate that performance differences between backbones are generally smaller after calibration than before calibration, consistent with Safe Calibration acting as a dominant reliability layer under the present shift regime.

Future work will investigate online calibration updates under continuous drift and extend the framework to multi-cell pack SoH estimation. The proposed Safe Calibration principle is also applicable to other regression tasks under domain shift in safety-critical systems, where monotonic structure is preserved, but output scaling becomes unreliable.

Author Contributions: Conceptualization, E.H.E.B.; methodology, E.H.E.B. and K.N.; software, K.N.; validation, E.H.E.B. and K.N.; formal analysis, E.H.E.B. and K.N.; investigation, K.N.; data curation, K.N.; writing—original draft preparation, K.N.; writing—review and editing, E.H.E.B.; supervision, E.H.E.B.; project administration, E.H.E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

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

Abbreviations

The following abbreviations are used in this manuscript:

ATT	Attention (self-attention mechanism)
BMS	Battery Management System
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
DA	Domain Adaptation
EMA	Exponential Moving Average
EOCV	End-of-Cycle
EV	Electric Vehicle
GRAD	Gradient
GRU	Gated Recurrent Unit
ID	Identity
LAB	Laboratory
L2F	Lab-to-Field
LR	Learning Rate
LSTM	Long Short-Term Memory
MAD	Median Absolute Deviation
MAE	Mean of Absolute value of Errors
NASA	National Aeronautics and Space Administration
OCV	Open-Circuit Voltage
PSO	Particle Swarm Optimization
RMSE	Root of the Mean of the Square of Errors
SOC	State of Charge
SoH	State-of-Health
TCN	Temporal Convolutional Network
TL	Transfer Learning

References

1. Bayoumi, E.H.E.; De Santis, M.; Awad, H. A Brief Overview of Modeling Estimation of State of Health for an Electric Vehicle's Li-Ion Batteries. *World Electr. Veh. J.* **2025**, *16*, 73. [\[CrossRef\]](#)
2. Acker, L.; Hofmann, P.; Konrad, J. Predictive Battery Thermal Management for Fast Charging of Electric Vehicles Using Nonlinear Model Predictive Control and Dynamic Programming. *Automot. Engine Technol.* **2026**, *11*, 1. [\[CrossRef\]](#)
3. Esparza, E.; Truffer-Moudra, D.; Hodge, C. *Electric Vehicle and Charging Infrastructure Assessment in Cold-Weather Climates: A Case Study of Fairbanks, Alaska*; National Renewable Energy Laboratory: Golden, CO, USA, 2025. [\[CrossRef\]](#)
4. Zhang, J.; Li, K. State-of-Health Estimation for Lithium-Ion Batteries in Hybrid Electric Vehicles—A Review. *Energies* **2024**, *17*, 5753. [\[CrossRef\]](#)
5. Vignesh, S.; Che, H.S.; Selvaraj, J.; Tey, K.S.; Lee, J.W.; Shareef, H.; Errouissi, R. State of Health (SoH) Estimation Methods for Second Life Lithium-Ion Battery—Review and Challenges. *Appl. Energy* **2024**, *369*, 123542. [\[CrossRef\]](#)
6. Bustos, J.E.G.; Schiele, B.B.; Baldo, L.; Masserano, B.; Jaramillo-Montoya, F.; Troncoso-Kurtovic, D.; Orchard, M.E.; Perez, A.; Silva, J.F. In Situ Estimation of Li-Ion Battery State of Health Using On-Board Electrical Measurements for Electromobility Applications. *Batteries* **2025**, *11*, 451. [\[CrossRef\]](#)
7. Salem, N.M.; Sayed, K.; Elsayed, M.M.; Kirakosian, D.; Mohamed, A.A. Short-Term Discharge-Based State of Health Estimation of Lithium-Ion Batteries. *Energy Rep.* **2026**, *15*, 108931. [\[CrossRef\]](#)
8. de la Vega Hernández, J.; Ortega-Redondo, J.A.; Riba, J.R. Lithium-Ion Battery Pack Cycling Dataset with CC-CV Charging and WLTP/Constant Discharge Profiles. *Sci. Data* **2025**, *12*, 1942. [\[CrossRef\]](#)
9. Fan, Q.; He, G.; Ruan, D.; Gühmann, C. Multi-Expert Fusion for State-of-Health Estimation of Lithium-Ion Batteries. *Sci. Rep.* **2025**, *15*, 42058. [\[CrossRef\]](#) [\[PubMed\]](#)

10. Nazim, M.S.; Rahman, M.M.; Mofidul, R.B.; Rimu, M.M.B.; Jang, Y.M. Robust State of Charge Estimation for Lithium-Ion Batteries Using a Fully Entangled Temporal Convolutional Network with Particle Swarm Optimization. *J. Power Sources* **2025**, *660*, 238456. [CrossRef]
11. Chen, S.Z.; Liang, Z.; Yuan, H.; Yang, L.; Xu, F.; Fan, Y. A Novel State of Health Estimation Method for Lithium-Ion Batteries Based on Constant-Voltage Charging Partial Data and Convolutional Neural Network. *Energy* **2023**, *283*, 129103. [CrossRef]
12. Zhang, H.; Gao, J.; Kang, L.; Zhang, Y.; Wang, L.; Wang, K. State of Health Estimation of Lithium-Ion Batteries Based on Modified Flower Pollination Algorithm-Temporal Convolutional Network. *Energy* **2023**, *283*, 128742. [CrossRef]
13. Zhao, F.-M.; Gao, D.-X.; Cheng, Y.-M.; Yang, Q. Estimation of Lithium-Ion Battery Health State Using MHATTCN Network with Multi-Health Indicators Inputs. *Sci. Rep.* **2024**, *14*, 18391. [CrossRef]
14. Wu, C.; Xu, C.; Wang, L.; Fu, J.; Meng, J. Lithium-Ion Battery Remaining Useful Life Prediction Based on Data-Driven and Particle Filter Fusion Model. *Green Energy Intell. Transp.* **2025**, *4*, 100267. [CrossRef]
15. Tian, H.; Xi, C.; Zhang, Q. A Framework for Estimating Battery State of Health Using Multi-Source Domain Adaptation and Real Vehicle Data. *J. Energy Storage* **2025**, *136*, 118449. [CrossRef]
16. Li, M.; Fei, Z.; Yang, L.; Zhang, Z.; Tsui, K.-L. Domain-Adaptive State of Health Prediction of Vehicle Batteries Powered by Deep Learning. *Cell Rep. Phys. Sci.* **2025**, *6*, 102550. [CrossRef]
17. Schreiber, M.; Köning, L.; Balke, G.; Gamra, K.A.; Kayl, J.; Dietermann, B.; Urban, R.; Grosu, C.; Lienkamp, M. Lab-to-Field Gap in Battery Aging Studies: Mismatch of Operating Conditions Between Laboratory Environments and Real-World Automotive Applications. *eTransportation* **2026**, *27*, 100518. [CrossRef]
18. Yang, K.; Xu, J.; Ni, X. Machine-Learning-Based Probabilistic Model and Design-Oriented Formula of Shear Strength Capacity of UHPC Beams. *Materials* **2025**, *18*, 4800. [CrossRef]
19. Qiang, X.; Liu, W.; Lyu, Z.; Ruan, H.; Li, X. A Data-Fusion-Model Method for State of Health Estimation of Li-Ion Battery Packs Based on Partial Charging Curve. *Green Energy Intell. Transp.* **2024**, *3*, 100169. [CrossRef]
20. Dang, S.; Sun, B.; Zhang, W.; Tao, Y.; Li, J. State of Health Prediction for Lithium-Ion Batteries in Energy Storage Systems Based on Domain Adaptation and Graph Attention Networks. *J. Energy Storage* **2026**, *144*, 119815. [CrossRef]
21. Huang, K.; Zhang, X.; Guo, Y.; Li, M. Source Domain Selection with Early-Cycle Features for Transfer Learning-Based Prediction of Lithium-Ion Battery Degradation Trajectories. *Energy* **2026**, *344*, 139928. [CrossRef]
22. Zhang, M.; Wang, X.; Kang, L.; Xie, D.; Liang, B. Bridging the Feature Gap: Heterogeneous Transfer Learning for Lithium-Ion Battery Health Estimation Using One-Shot Data. *Results Eng.* **2026**, *29*, 109163. [CrossRef]
23. Liu, H.; Li, C.; Hu, X.; Li, J.; Zhang, K.; Xie, Y.; Wu, R.; Song, Z. Multi-Modal Framework for Battery State of Health Evaluation Using Open-Source Electric Vehicle Data. *Nat. Commun.* **2025**, *16*, 1137. [CrossRef] [PubMed]
24. Kulkarni, S.V.; Arjun, G.; Gupta, S.; Sinha, R.; Shukla, A. Advanced Battery Diagnostics for Electric Vehicles Using CAN Based BMS Data with EKF and Data Driven Predictive Models. *Sci. Rep.* **2025**, *15*, 32848. [CrossRef] [PubMed]
25. Zhang, Y.; Xu, C.; Wu, Y.; Guan, Z.; Zhao, W. Gradient Rectification for Robust Calibration under Distribution Shift. *arXiv* **2025**, arXiv:2508.19830. [CrossRef]
26. Cheng, J.; Tian, J.; Spoto, F.; Azhir, A.; Mork, D.; Estiri, H. Signal Fidelity Index-Aware Calibration for Addressing Distributional Shift in Predictive Modeling across Heterogeneous Real-World Data. *Sci. Rep.* **2025**, *16*, 2807. [CrossRef]
27. Ba, Y.; Mancenido, M.V.; Pan, R. Fill In The Gaps: Model Calibration and Generalization with Synthetic Data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Miami, FL, USA, 12–16 November 2024; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 17211–17225. [CrossRef]
28. Zollo, T.P.; Deng, Z.; Snell, J.C.; Pitassi, T.; Zemel, R. Improving Predictor Reliability with Selective Recalibration. *arXiv* **2024**, arXiv:2410.05407v1. [CrossRef]
29. Zhang, Y.; Batista, G.; Kanhere, S.S. Instance-Wise Monotonic Calibration by Constrained Transformation. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI)*, Rio de Janeiro, Brazil, 21–25 July 2025. Available online: <https://arxiv.org/abs/2507.06516> (accessed on 7 March 2026).
30. Rowan, R.R.; Hanif, A.; Fahim, M.Q.; Le, D.D.; Anwar, H.; Jaleel, W.; Nelson, M.; Ahmed, Q. Learning Battery Aging Dynamics Using Physics-Informed Transformer. *IEEE Trans. Transp. Electr.* **2026**. [CrossRef]
31. Zhao, J.; Li, D.; Li, Y.; Shi, D.; Nan, J.; Burke, A.F. Battery State of Health Estimation under Fast Charging via Deep Transfer Learning. *iScience* **2025**, *28*, 112235. [CrossRef]
32. Chen, G.; Meng, H.; Yang, Y.; Zhang, X.; Deng, W.; Liu, J. A Transfer Learning Method for State of Health Prediction of Lithium-Ion Batteries under Cross-Formation Protocols. *J. Energy Storage* **2026**, *147*, 120113. [CrossRef]
33. Ma, G.; Wang, Z.; Liu, W.; Fang, J.; Zhang, Y.; Ding, H.; Yuan, Y. Estimating the State of Health for Lithium-Ion Batteries: A Particle Swarm Optimization-Assisted Deep Domain Adaptation Approach. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1530–1543. [CrossRef]
34. Gui, X.; Du, J.; Wang, Q.; Zhao, H.; Cheng, Y.; Zhao, J. Multi-Modal Data Information Alignment Based SOH Estimation for Lithium-Ion Batteries Using a Local–Global Parallel CNN-Transformer Network. *J. Energy Storage* **2025**, *129*, 117178. [CrossRef]

35. Hu, J.; Zhang, Q.; Liu, F.; Hu, Z.; Oh, C.; Gong, S.; Liu, Z. Rank-Preserving Calibration of LLMs Under Model and Distribution Shifts. *OpenReview* **2025**. Available online: <https://openreview.net/forum?id=0crU7lZV8n> (accessed on 7 March 2026).
36. Soon, K.L.; Soon, L.T. Enhancing Reliability in Electrified Transportation: A Conformalized Quantile Regression Framework for Battery State-of-Charge Uncertainty Quantification. *J. Power Sources* **2026**, *666*, 239123. [[CrossRef](#)]
37. Yan, M.-X.; Deng, Z.-H.; Lai, L.; Xu, Y.-H.; Tong, L.; Zhang, H.-G.; Li, Y.-Y.; Gong, M.-H.; Liu, G.-J. A Sustainable SOH Prediction Model for Lithium-Ion Batteries Based on CPO-ELM-ABKDE with Uncertainty Quantification. *Sustainability* **2025**, *17*, 5205. [[CrossRef](#)]
38. Wüthrich, M.V.; Ziegel, J. Isotonic Recalibration under a Low Signal-to-Noise Ratio. *Scand. Actuar. J.* **2024**, *2024*, 279–299. [[CrossRef](#)]
39. NASA. Li-Ion Battery Aging Datasets. Available online: <https://data.nasa.gov/dataset/li-ion-battery-aging-datasets> (accessed on 4 March 2026).
40. Luh, M. Battery Aging Dataset (Result Data, V2). Available online: <https://www.kaggle.com/datasets/matthiasluh/battery-aging-dataset-result-data-v2> (accessed on 4 March 2026).
41. Luh, M.; Blank, T. Comprehensive Battery Aging Dataset: Capacity and Impedance Fade Measurements of a Lithium-Ion NMC/C-SiO Cell. *Sci. Data* **2024**, *11*, 1004. [[CrossRef](#)]
42. Chen, Y.; Li, D.; Huang, X.; Hong, J.; Mu, C.; Wu, L.; Li, K. Corrigendum to “Exploring Life Warning Solution of Lithium-Ion Batteries in Real-World Scenarios: TCN-Transformer Fusion Model for Battery Pack SOH Estimation” [Energy 335 (2025) 138053]. *Energy* **2025**, *336*, 138480. [[CrossRef](#)]
43. Zhao, J.; Chu, F.; Xie, L.; Che, Y.; Wu, Y.; Burke, A.F. A Survey of Transformer Networks for Time Series Forecasting. *Comput. Sci. Rev.* **2026**, *60*, 100883. [[CrossRef](#)]
44. Nayak, G.H.H.; Alam, M.W.; Avinash, G.; Kumar, R.R.; Ray, M.; Barman, S.; Singh, K.N.; Naik, B.S.; Alam, N.M.; Pal, P.; et al. Transformer-Based Deep Learning Architecture for Time Series Forecasting. *Softw. Impacts* **2024**, *22*, 100716. [[CrossRef](#)]
45. Foumani, N.M.; Tan, C.W.; Webb, G.I.; Salehi, M. Improving Position Encoding of Transformers for Multivariate Time Series Classification. *Data Min. Knowl. Discov.* **2024**, *38*, 22–48. [[CrossRef](#)]
46. Soo, Y.-Y.; Wang, Y.; Xiang, H.; Chen, Z. A Novel Transfer Learning Model for Battery State of Health Prediction Based on Driving Behavior Classification. *J. Energy Storage* **2025**, *111*, 115409. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.