ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.2467

Liquid Neural Networks for Battery Prognostics: Multi-Metric Evaluation and Interpretability Analysis

Anthony José Arguedas-Rodríguez[1]; Juan José Montero-Jiménez[2]
[1,2]Tecnológico de Costa Rica, Costa Rica, antarguedas@estudiantec.cr, juan.montero@itcr.ac.cr

***Abstract*—Data-driven battery prognostics models are increasingly expected to be accurate, efficient, robust, and interpretable, yet existing benchmarks often report accuracy alone using random train/test splits that permit temporal leakage. This study presents a comprehensive multi-metric evaluation of liquid neural networks (LNNs) and neural circuit policies (NCPs) for capacity-based state-of-health prediction on the four-unit NASA Prognostics Center of Excellence (PCoE) lithium-ion battery dataset under a chronological splitting protocol designed to prevent future-cycle information leakage. On the shared raw-capacity benchmark, compact continuous-time architectures (under 10K parameters) obtain accuracy comparable to selected larger baselines with approximately 5–550× more parameters in the evaluated same-protocol and contextual comparisons; the main LNN-TS configuration trains 4–20× faster than conventional raw time-series baselines. An accuracy-robustness trade-off emerges across model families, with the most accurate models exhibiting 4.6–5.5% noise degradation and NCP sparsity providing a controllable trade-off between prediction quality and noise insensitivity. However, intrinsic uncertainty estimation through tau variability and hidden-state stability yields poorly calibrated confidence signals, identifying reliable self-assessment of prediction quality as an open challenge. Complementary interpretability experiments (confidence estimation, hidden-state trajectory analysis, phase-wise position importance, and counterfactual perturbation) suggest that LNNs learn representations correlated with physically plausible degradation behavior, while fixed-sparsity NCP experiments characterize accuracy–interpretability trade-offs without claiming learned causal pathways. Together, these results support compact continuous-time architectures as efficient and interpretable candidates for battery prognostics benchmarks that jointly evaluate accuracy, robustness, and model transparency.**

***Keywords*—prognostics, liquid neural networks, neural circuit policies, interpretability, multi-metric model evaluation**

I. Introduction

Predictive maintenance requires data-driven prognostics models to be as reliable and interpretable as the physical systems and components they oversee. These high expectations for prognostics models are justified by the fact that the performance of a system cannot be guaranteed without first determining the state-of-health (SOH) of individual components [1]. However, this is a complex task that traditional interpretable models perform poorly on, influenced by the continuous, multivariate, non-linear and noisy nature of the data streams commonly used as inputs.

As a result, the field of predictive maintenance has produced an extensive literature on machine learning models for battery prognostics. Physics-informed neural networks (PINNs) incorporate domain knowledge through partial differential equation loss terms, constraining model predictions to respect monotonic capacity degradation [2]. This architecture offers interpretability through constrained physical behavior but potentially limits flexibility.

Transformer-based models have emerged as powerful alternatives for time series prognostics. The Convolutional Transformer-based Multiview Information Perception framework (MVIP-Trans) processes battery data through a transformer encoder with masked attention to handle variable-length sequences, employing learnable positional embeddings and masked mean pooling [3].

Hybrid architectures attempt to combine the strengths of multiple approaches. Closed-form Continuous-depth Hybrid Difference Features Re-representation Networks (CfC-F2RN) use a long short-term memory (LSTM) encoder with hard attention to select important hidden states, followed by closed-form continuous-time decoder layers [4].

Liquid neural networks (LNNs) model continuous-time neural dynamics with adaptive time constants using a closed-form approximation of ordinary differential equations [5]. LNNs offer five key advantages for prognostics: physical interpretability through differential equation-based modeling, strong noise resistance and robustness to out-of-distribution data, computational efficiency through reduced parameter counts, dynamic adaptability suitable for time-varying environments, and balanced handling of dependencies with prioritization of short-term causal information [6]. This compact architecture achieves complex temporal modeling with minimal hidden units, naturally supporting interpretability due to small parameter counts and explicit time-scale parameters.

Neural Circuit Policies (NCPs) employ brain-inspired sparse connectivity patterns from the *C. elegans* nervous system alongside the same closed-form continuous-time dynamics as LNNs. Their structured sparsity makes them relevant for examining accuracy–interpretability trade-offs in compact continuous-time prognostics models.

Across data-driven prognostics models, a common limitation of existing benchmarks is the emphasis on accuracy alone without systematic evaluation of robustness, efficiency, or explainability [7]. Moreover, many battery prognostics models are evaluated on random train/test splits, which can allow models to see future degradation patterns during training and inflate reported performance. This temporal leakage obscures the true generalization capability required for real-world deployment where only historical data is available.

Recent work has applied LNNs to battery health prediction

for edge deployment [8] and as components within hybrid architectures [4], but these approaches either obscure interpretability through distillation pipelines or trade off inherent interpretability for increased model complexity.

Thus, systematic interpretability analysis for LNNs in battery prognostics applications remains unexplored. Interpretability has become a point of concern as machine learning methods gain academic attention and industrial relevance, both for regulatory compliance and for scientific understanding of model behavior [6].

For other data-driven models, interpretability approaches for battery prognostics have primarily focused on attention mechanisms and perturbation-based analysis, with attention-based approaches identifying critical phases in discharge cycles [9], Partial Dependence Plots for convolutional neural network–recurrent neural network (CNN-RNN) hybrid models [10], and saliency analysis methods for chemistry-agnostic lifetime prediction [11].

In this work, we evaluate LNNs alongside Neural Circuit Policies (NCPs) in the task of battery prognostics using a multi-metric framework and conduct a systematic interpretability analysis. The contributions of this work are:

1) A chronological evaluation protocol for the NASA PCoE battery prognostics benchmark, designed to prevent future-cycle temporal leakage and to test whether reference implementations using random splits could report inflated performance.
2) A multi-metric benchmark of compact continuous-time architectures (LNNs, NCPs) against raw time-series PINN, transformer-based, and hybrid models on a common capacity target, showing that, within this controlled benchmark and contextual reference comparisons, models with under 10K parameters can reach accuracy comparable to selected larger architectures with approximately 5–550$\times$ more parameters.
3) Quantification of training efficiency advantages, with the main LNN-TS compact continuous-time model training 4–20$\times$ faster than conventional raw time-series baselines in the evaluated comparisons.
4) Identification of an accuracy-robustness trade-off across model families and NCP sparsity levels, where the most accurate models exhibit the highest noise sensitivity.
5) A negative result establishing that intrinsic uncertainty estimation through tau variability and hidden-state stability does not yield reliable confidence signals for LNN prognostics.
6) Convergent interpretability evidence that LNN representations are correlated with capacity/SOH progression and physically plausible degradation indicators, obtained through hidden-state trajectory analysis, phase-wise position importance, and counterfactual perturbation experiments. NCP sparsity is analyzed as a fixed-topology accuracy–interpretability trade-off.

Because the NASA PCoE benchmark contains only four battery units, we position the study as a controlled benchmark under a leakage-aware chronological protocol rather than evidence of broad generalization across chemistries, operating conditions, or deployment settings.

The remainder of the paper is organized as follows: Section II describes the experimental setup, including the dataset and model architectures. Section III presents the experimental results, including the multi-metric benchmarking and interpretability analysis. Section IV discusses the implications of the results, and Section V concludes the paper.

II. Methodology

Following [7], this study uses the NASA Prognostics Center of Excellence (PCoE) lithium-ion battery dataset, comprising discharge cycles from four battery units (B0005, B0006, B0007, B0018) [12]. Each cycle contains variable-length time series of voltage, current, temperature, and relative time. The primary supervised target for the raw time-series models is measured discharge capacity in ampere-hours (Ah), used as a continuous proxy for state-of-health (SOH). The feature-based LNN-PINN variant follows the PINN4SOH preprocessing and predicts normalized capacity (capacity divided by nominal capacity). Consequently, direct accuracy comparisons focus on models trained on the shared raw-capacity target, while normalized-SOH feature variants are reported on their native scale for context. Unless explicitly stated otherwise, train/validation splits within training units are chronologically ordered, so models are trained on earlier cycles and validated on later cycles while the test unit remains held out.

A. Model Architectures

1) Liquid Neural Networks: The raw time-series LNN implementation processes each discharge cycle timestep by timestep using a closed-form continuous-time recurrent update with $\Delta t = 1$. At timestep t, the four measured inputs are projected to u_t, optionally augmented with a learned positional embedding p_t, and combined with the previous hidden state h_{t-1}:

$$\begin{aligned} u_t &= W_{\text{in}} x_t + b_{\text{in}} + p_t, \quad z_t = [h_{t-1}; u_t], \\ \tau_t &= \tau_{\min} + (\tau_{\max} - \tau_{\min})\sigma(W_\tau z_t + b_\tau), \\ g_t &= \tanh(W_g z_t + b_g), \quad \alpha_t = \exp(-1/\tau_t), \\ h_t &= \alpha_t \odot h_{t-1} + (1 - \alpha_t) \odot g_t. \end{aligned} \tag{1}$$

Here σ denotes the logistic sigmoid and $\odot$ denotes elementwise multiplication. Padding masks leave $h_t = h_{t-1}$ after a sequence's true length, and the final hidden state is passed to an MLP readout.

We evaluate LNNs in three data-processing variants. LNN-TS processes raw time series with positional embeddings up to `max_seq_len`=500 to handle variable-length sequences, followed by an LNN layer and a single-layer readout network. LNN-PINN uses the statistical features from the PINN reference implementation [2], including cycle-level statistics computed from discharge cycles, and follows that pipeline's normalized-SOH target scale. LNN-MVIP uses the fixed-length battery data used for training MVIP-Trans models [3].

2) Neural Circuit Policies: Neural Circuit Policy (NCP) models incorporate brain-inspired sparse connectivity patterns inspired by the *C. elegans* nervous system [13], [14]. This topology provides hierarchical temporal dynamics through sensory projections, interneurons, command neurons, and a readout. We implemented NCP-TS using the same closed-form continuous-time dynamics as LNNs but with hierarchical sparse linear projections: four input features are projected to 16 interneurons, then to 8 command neurons, followed by an MLP readout for capacity prediction. Connection sparsity is controlled by binary masks sampled at initialization and stored as non-trainable buffers. Training updates the surviving weights whereas masks remain constant. We evaluate NCP-TS at three sparsity levels: dense (0% target sparsity), moderate (50% target sparsity), and high (80% target sparsity) to assess the trade-off between connectivity and performance. Because sparsity is implemented through multiplicative binary masks applied to fixed-size weight matrices, the total allocated parameter count (9,977) remains constant across all sparsity configurations, with only the number of effective active connections changing.

3) CfC-F2RN: Closed-form Continuous-depth Hybrid Difference Features Re-representation Networks (CfC-F2RN) combine an LSTM encoder with hard attention and closed-form continuous-time decoder layers [4]. We evaluate CfC-F2RN with 8 and 16 hidden units, 1–3 CfC layers, and varying learning rates and batch sizes.

4) Time Series-based PINN and MVIP-Trans: The PINN-Raw variant extends the reference PINN approach [2] by adding an LSTM encoder to process raw time series data. We evaluate two PINN-Raw configurations with LSTM hidden sizes of 16 and 32, yielding 13K and 24K parameters respectively. Similarly, the MVIP-Raw variant derived from [3] processes raw time series directly but requires substantial modifications including linear patch embedding and careful sequence length handling. These variants are evaluated to provide a baseline for comparison with LNNs and NCPs.

B. Evaluation Framework

We employ the multi-metric evaluation framework introduced by [7] for model benchmarking. Leave-one-unit-out cross-validation (LOOCV) is applied across the four battery units. For each fold, three units are used for training with a chronological train/validation split (80%/20%) performed by cycle index, and the fourth unit serves as test data. This procedure keeps validation cycles later than training cycles within the training units, and the held-out test unit is never used for fitting.

The four metrics used for model evaluation are: accuracy via root-mean-square error (RMSE) and mean absolute error (MAE) on test predictions in each model's native target scale, generalization gap (RMSE_{test} - RMSE_{train}) measuring performance degradation from train to test, robustness via average ΔRMSE under four noise types following [15], and efficiency via training time in seconds. Robustness is computed as the relative change $\Delta\text{RMSE} = (\text{RMSE}_{\text{noisy}} - \text{RMSE}_{\text{clean}})/\text{RMSE}_{\text{clean}}$, where ΔRMSE is averaged across four noise profiles and four LOOCV folds. Early stopping with a patience of 50 epochs is enabled for all training runs (unless another patience value is stated explicitly), up to a maximum of 500 training epochs.

For context, we refer to MVIP-Trans and PINN reference implementations reported in previous literature. These reference implementations use the same NASA PCoE dataset but employ random train/test splits without chronological ordering, allowing models to see future cycles during training. As an additional protocol-sensitivity check, we evaluate an LNN-TS random-split ablation in which training and validation cycles are shuffled within the same LOOCV setup while the fourth battery unit remains fully held out for testing. This ablation uses the same LNN-TS architecture and changes only the train/validation ordering. It tests split-ordering effects within our pipeline, but it is not treated as a direct reproduction of single-battery random train/test protocols used in some prior work.

Accordingly, only models retrained in this work under the shared chronological LOOCV protocol are treated as direct baselines. Literature-reported random-split MVIP-Trans and PINN results are included as reference points to illustrate the effect of evaluation protocol rather than fully matched comparisons.

C. Hyperparameter Sweep

Hyperparameters were explored through structured sequential search while maintaining a consistent training protocol across all model families (Section II-B). Rather than exhaustive grid search across all hyperparameter combinations, which would be computationally prohibitive given the high-dimensional search space, we employed systematic one-dimensional sweeps. This approach allows meaningful exploration of each parameter's effect while maintaining computational feasibility [16]. The search space intervals listed in Table I define the candidate value sets tested across multiple configurations following domain-specific tuning strategies [17]. Models not subjected to a hyperparameter sweep were trained using fixed configurations as specified in their respective reference implementations to ensure faithful reproduction of baseline settings. All experiments were implemented in Python using PyTorch.

D. Interpretability Analysis

Beyond multi-metric trade-off benchmarks, we investigate the interpretability properties of LNNs and the structural trade-offs of NCPs through complementary diagnostic experiments. Recent prognostics studies [11], [18] likewise motivate combining multiple interpretability views rather than relying on a single post-hoc explanation.

1) Tau-Based Confidence Estimation: The adaptive time constants τ in LNNs capture the temporal integration patterns of each hidden unit throughout the discharge cycle. We explore whether tau variability can serve as an internal confidence signal by hypothesizing that predictions with stable, consistent tau dynamics reflect well-modeled patterns, while high

TABLE I
Sequential hyperparameter search ranges and number of evaluated configurations per model family. Candidate value sets were explored one dimension at a time, while fixed values indicate inherited or held-constant settings.

Model family	**# variants**	**Hyperparameters explored (sequential) / fixed**
LNN-TS	27	Hidden units $\in \{8, 16, 24\}$; learning rate $\in \{0.001, 0.005, 0.01, 0.02\}$; dropout $\in \{0.0, 0.1\}$; monotonicity loss weight $\in \{0.0, 0.1, 0.5, 1.0\}$; batch size $\in \{4, 8, 16\}$; $\tau_{\min} \in \{0.0001, 0.001\}$; $\tau_{\max} \in \{0.1, 1.0, 10.0, 20.0\}$; readout hidden size $\in \{16, 32, 64\}$; positional embedding ablation (enabled/disabled).
LNN-PINN	3	Hidden units $\in \{8, 16, 24\}$; fixed learning rate $= 0.005$; fixed dropout $= 0.1$; fixed monotonicity loss weight $= 0.1$.
NCP-TS	14	Hidden units $= 16$; sparsity $\in \{0\%, 50\%, 80\%\}$; learning rate $\in \{0.001, 0.005\}$; dropout $\in \{0.0, 0.1\}$; monotonicity loss weight $\in \{0.0, 0.1\}$; batch size $\in \{8, 16\}$.
CfC-F2RN	9	Hidden units $\in \{8, 16\}$; CfC layers $\in \{1, 2, 3\}$; learning rate $\in \{0.0005, 0.001, 0.005\}$; batch size $\in \{4, 8, 16\}$; monotonicity loss weight $\in \{0.0, 0.1\}$.
MVIP-Raw	2	Model dimension $\in \{32, 64\}$; fixed depth $= 4$; fixed learning rate $= 0.001$; fixed batch size $= 8$.
PINN-Raw	2	LSTM hidden size $\in \{16, 32\}$; fixed encoder layers $= 3$; fixed F network layers $= 3$; fixed $\alpha = 0.7$; fixed $\beta = 0.2$; fixed learning rate $= 0.001$.

tau variance indicates uncertainty or regime shifts. For each prediction we compute confidence as $c = 1/(1+\sigma_\tau^2)$, where σ_τ^2 is the variance of tau values across timesteps and hidden units for that sample. We bin predictions by confidence and compute expected calibration error (ECE) and maximum calibration error (MCE) to quantify average and worst-case miscalibration.

2) Hidden State Stability as Confidence Signal: As a complementary approach to tau-based confidence estimation, we investigate hidden state stability as an alternative intrinsic uncertainty signal. The hypothesis is that lower variance in hidden state activations across timesteps indicates more stable internal representations and thus higher prediction confidence. We test this approach across three network sizes (8, 16, and 24 hidden units) to assess whether the signal is consistent across model capacities. For each configuration, we compute the variance of hidden state activations across timesteps for each sample, derive confidence scores using the same inverse-variance formulation as tau-based estimation ($c = 1/(1+\sigma_h^2)$, where σ_h^2 is the variance of hidden state activations), and evaluate calibration via ECE.

3) Hidden State Trajectory Analysis across SOH Progression: To understand how LNNs represent battery degradation, we extract hidden state trajectories across the full battery lifespan for each unit. We project high-dimensional hidden states into two dimensions using principal component analysis (PCA), analyze the first two principal components (PC1 and PC2), and label trajectories by the measured capacity/SOH proxy. We compute the linear correlation between position along PC1 and the true target value, quantifying how well the learned latent space captures degradation progression. Because PCA component signs are arbitrary, aggregate comparisons use absolute PC1-target correlations.

To establish statistical significance, we employ permutation testing with 10,000 label randomizations, generating a null distribution of PC1-target correlations under the null hypothesis of no relationship. We report both standard parametric p-values and permutation p-values, the latter being non-parametric and assumption-free. Bootstrap 95% confidence intervals (CIs; 10,000 resamples) quantify the precision of correlation estimates and account for finite-sample uncertainty. As a control, we perform identical PCA on LSTM encoder hidden states from PINN-Raw models.

4) Phase-wise Position Importance: Discharge cycles exhibit distinct phases with varying diagnostic value for SOH estimation. We divide each cycle into early (first 20%), mid (20–80%), and late (last 20%) phases and aggregate position importance scores within each phase. Position importance is computed as the magnitude of the gradient of the loss with respect to the input features at each timestep (Euclidean (L_2) norm across input features). Note that gradients scale with loss magnitude, so this metric reflects gradient sensitivity rather than signal quality.

5) Fixed NCP Sparsity Analysis: NCPs are designed for interpretability through structured, biologically inspired connectivity [13], [14]. In the implementation evaluated here, the sparse masks are fixed after initialization rather than learned. We therefore treat NCP connectivity as an architectural sparsity control rather than as a learned causal pathway explanation. For each NCP variant (dense, moderate, high sparsity), we report the resulting active connection counts, effective sparsity, prediction performance, and noise sensitivity under the same LOOCV protocol used for the other models. This analysis quantifies the trade-off between structural simplicity and predictive accuracy without assigning causal meaning to individual masked pathways.

6) Counterfactual Perturbation Analysis: To identify which battery measurements are most critical for accurate capacity/SOH prediction, we perform counterfactual perturbation analysis by systematically introducing controlled perturbations to each input feature. For each of the four input features (Voltage, Current, Temperature, Relative Time), we generate perturbed samples at six magnitude levels: $\pm 10\%$, $\pm 20\%$, and $\pm 30\%$ in physical units. For each perturbed sample we compute tau variance sensitivity ($\Delta\sigma_\tau^2$) and prediction impact ($\Delta\hat{y}$), where $\hat{y}$ is the model output on its target scale. Features inducing larger absolute changes in both metrics are identified as more critical for the model's decision process. We aggregate results across all LOOCV folds using z-score normalization

and compute correlations between $\Delta\sigma_\tau^2$ and $|\Delta\hat{y}|$ to test whether changes in internal temporal dynamics correlate with prediction sensitivity.

E. Interpretability Experimental Design

For the LNN-focused confidence, trajectory, phase-importance, and counterfactual experiments, we use LNN-TS-16-1L-lr001-p50 (16 hidden units, learning rate 0.001, patience 50 epochs) trained on each LOOCV fold, yielding four independent model instances. The accuracy-oriented LNN-TS-16-1L-lr001-p30 configuration is retained as the main LNN accuracy baseline, whereas p50 is used for interpretability diagnostics that require non-degenerate internal dynamics over the discharge trajectory. The quantitative RMSE and tau-dynamics comparison between p30 and p50 is reported in Section III. For the NCP sparsity analysis, we analyze three NCP-TS sparsity variants (dense, moderate, high) trained on the same LOOCV splits and aggregate active connection counts, prediction errors, and robustness metrics across folds.

III. Results

Table II summarizes selected representative configurations, including the best-performing variants and sparsity/size alternatives used for trade-off analysis, across the multi-metric evaluation framework.

A. Hyperparameter Ablation Insights

Table III summarizes the targeted LNN-TS ablation highlights. The results indicate that weak monotonicity regularization was preferable to either removing the term or enforcing stronger constraints, positional embeddings improved accuracy despite their parameter overhead, lower learning rates were more stable, and the 8-unit model reduced parameters substantially at measurable accuracy cost.

B. Chronological versus Random Split Ablation

The random train/validation ablation under the same held-out-unit LOOCV test protocol achieved RMSE 0.0865 ± 0.0561 and robustness degradation 0.0617 ± 0.0711. This did not improve over the chronological LNN-TS-16-1L-lr001-p30 result in Table II (RMSE 0.0795 ± 0.0410) and showed higher noise sensitivity. Therefore, random-split results are used only to contextualize protocol sensitivity rather than as direct baselines.

C. Effect of Learned Tau Dynamics

Comparing the compact-tau LNN-TS-16-1L-lr001-p30 run (mean $\tau = 0.051$, variance $= 0.00017$, maximum observed $\tau = 0.089$) with the interpretability-oriented LNN-TS-16-1L-lr001-p50 run (mean $\tau = 10.36$, variance $= 15.35$, maximum observed $\tau = 19.53$) shows substantially larger learned tau dynamics in the latter, with modest RMSE degradation (from 0.0795 to 0.0839, +6%). Phase-wise importance analysis revealed that early discharge phase importance increased from approximately 10^{-25} to 1.4×10^{-7}, mid-phase importance from 2×10^{-12} to 1.3×10^{-5}, while late-phase importance remained unchanged at 0.0014.

D. Tau-Based Confidence Estimation Results

The tau-based confidence estimation reveals that tau variability alone does not provide a reliable signal for estimating prediction accuracy. The expected calibration error (ECE) of 0.78 exceeds the typical threshold of 0.05 for well-calibrated models, indicating substantial miscalibration. The maximum calibration error (MCE) was 0.89, representing the worst deviation between predicted confidence and actual accuracy across confidence bins.

The correlations were not informative: Pearson r = -0.205 ($p = 1.9 \times 10^{-7}$) and Spearman $\rho = 0.0003$ ($p = 0.99$). The confidence distribution was highly concentrated with mean approximately 0.001 and minimal variance, indicating that the model assigns uniformly low confidence across all samples regardless of prediction accuracy. This uniform distribution prevents effective discrimination between reliable and unreliable predictions.

E. Hidden State Stability as Confidence Signal

Testing across three network sizes (8, 16, and 24 hidden units) revealed performance comparable to the tau-based approach, with ECE values of 0.673, 0.733, and 0.656 respectively. The 16-unit configuration showed moderate positive Pearson correlation ($r = +0.399$, $p = 1.2 \times 10^{-25}$) and positive Spearman correlation ($\rho = +0.358$, $p = 1.1 \times 10^{-20}$). The 8-unit model produced a positive correlation ($r = 0.360$, $\rho = 0.369$), while the 24-unit model showed no meaningful correlation ($r = 0.056$, $\rho = 0.028$).

Confidence distributions were narrow and concentrated across models, preventing discrimination between reliable and unreliable predictions. The high miscalibration across all network sizes indicates that hidden state variance does not provide a reliable uncertainty signal for LNN prognostics.

F. Hidden State Trajectory Analysis Results

The PCA-based trajectory analysis showed strong associations between the first hidden-state principal component and the capacity/SOH target. Table IV reports the per-battery trajectory statistics. Across all four battery units, the first principal component (PC1) captured the majority of hidden-state variance, and all observed PC1-target correlations were significantly different from random under permutation testing. The sign changes across batteries reflect PCA sign indeterminacy rather than opposite degradation semantics. The mean absolute PC1-target correlation was $|r| = 0.926$ across units, while the mean signed correlation was near zero ($r = -0.063$). B0018 remained the weakest trajectory but still showed a substantial PC1-target relationship, consistent with its shorter cycle record (132 versus 168 cycles for the other units). Figure 1 shows the B0005 trajectory dashboard.

Comparison to LSTM encoder hidden states from PINN-Raw models showed different representation geometry, but the current results do not support a statistically significant superiority claim for LNNs. Absolute PC1-target correlations were higher on average for LNNs than for LSTMs (mean $|r| = 0.926\pm0.126$ vs $|r| = 0.747\pm0.143$), but with only four

TABLE II

Representative configurations from each model family. All metrics are reported as mean ± standard deviation across four LOOCV folds with chronological train/validation/test splits. RMSE and MAE are in Ah for raw-capacity models, while the LNN-PINN feature baseline is reported on normalized SOH scale. Robustness denotes mean relative ΔRMSE under four input-noise types.

Model	**Params**	**RMSE**$_{test}$	**MAE**$_{test}$	**Gen. Gap**	**Robust.**	**Train Time (s)**
NCP-TS-16-Dense	9,977	0.0754 ± 0.0441	0.0616 ± 0.0361	**0.0303** ± 0.0266	0.0555 ± 0.0567	278.28 ± 116.48
LNN-TS-16-1L-lr001-p30‡	9,425	0.0795 ± 0.0410	0.0699 ± 0.0364	0.0422 ± 0.0321	0.0554 ± 0.0518	101.37 ± 41.99
NCP-TS-16-1L	9,977	0.0978 ± 0.0554	0.0814 ± 0.0525	0.0562 ± 0.0413	0.0165 ± 0.0158	262.22 ± 103.32
NCP-TS-16-Sparse	9,977	0.1807 ± 0.0626	0.1572 ± 0.0517	0.1036 ± 0.0569	0.0013 ± 0.0017	217.51 ± 5.33
MVIP-Raw-32	51,041	0.0758 ± 0.0377	0.0547 ± 0.0266	0.0393 ± 0.0165	0.0458 ± 0.0485	445.51 ± 139.57
CfC-F2RN-16-bs4	5,138	0.0883 ± 0.0372	0.0793 ± 0.0387	0.0439 ± 0.0373	0.0018 ± 0.0053	171.92 ± 57.97
LNN-PINN-16-1L†	865	0.0975 ± 0.0121	0.0831 ± 0.0093	0.0259 ± 0.0169	0.0013 ± 0.0023	**2.53** ± 0.36
LNN-MVIP-16-1L	961	0.1978 ± 0.0377	0.1734 ± 0.0350	0.0527 ± 0.0256	–	180.98 ± 16.06
PINN-Raw-16	13,062	0.1246 ± 0.0560	0.1101 ± 0.0556	0.0564 ± 0.0486	-0.0008 ± 0.0017	2016.57 ± 815.78
MVIP-Raw-64	169,665	0.1157 ± 0.0567	0.0928 ± 0.0561	0.0191 ± 0.0118	–	1137.89 ± 503.08
PINN-Raw-32	24,262	0.1982 ± 0.0850	0.1664 ± 0.0810	0.0630 ± 0.0386	–	1760.08 ± 501.67

†LNN-PINN follows the PINN4SOH normalized-SOH target scale [2]; direct raw-capacity accuracy comparisons should use the other rows.
‡LNN-TS-16-1L-lr001-p30 τ variance: 0.000165. LNN-TS-16-1L-lr001-p50 was selected for interpretability experiments (confidence estimation, trajectory analysis, phase-wise importance, counterfactual evaluation) due to its higher τ variance (15.35), which enables meaningful confidence analysis.
§NCP sparsity is implemented via multiplicative binary masks, so total allocated parameters remain constant across sparsity levels while effective active connections vary: Dense (∼192), 1L (∼83), Sparse (∼33).

TABLE III

Targeted LNN-TS hyperparameter ablation highlights. Rows are local one-dimensional sweeps and should be compared within each ablation group rather than across groups. RMSE values are mean ± standard deviation across four LOOCV folds when fold-level deviations were reported. ΔRMSE is relative to the best setting within each local comparison.

Ablation	**Setting**	**Params**	**RMSE**$_{test}$	**ΔRMSE**	**Interpretation**
Monotonicity weight	0.0	9,425	0.1880 ± 0.0237	+0.0373	Removing weak physics regularization hurt accuracy.
	0.1	9,425	**0.1507 ± 0.0468**	0.0000	Best setting in targeted sweep.
	0.5	9,425	0.1816 ± 0.0154	+0.0309	Stronger constraint over-regularized.
	1.0	9,425	0.1625 ± 0.0445	+0.0118	Still worse than weak constraint.
Positional embedding	Enabled	9,425	**0.1507 ± 0.0468**	0.0000	Timestep information improved accuracy.
	Disabled	1,425	0.1868 ± 0.0193	+0.0361	Fewer parameters, lower accuracy.
Learning rate	0.001	–	**0.0915**	0.0000	More stable optimization.
	0.02	–	0.2021	+0.1106	High learning rate degraded accuracy.
Hidden size	16 units	9,425	**0.0795 ± 0.0410**	0.0000	Main accuracy baseline.
	8 units	4,473	0.1016 ± 0.0502	+0.0221	52.5% fewer parameters with accuracy loss.

TABLE IV

Per-battery hidden-state PCA trajectory statistics for LNN-TS-16-1L-lr001-p50. PCA component signs are arbitrary; aggregate comparisons use absolute PC1-target correlations.

Battery	**PC1 Var.**	r**(PC1,target)**	**95% CI**	**Perm.** p
B0005	87.1%	−0.983	[−0.987, −0.979]	< 0.001
B0006	88.9%	−0.995	[−0.996, −0.994]	< 0.001
B0007	85.3%	+0.990	[+0.987, +0.992]	< 0.001
B0018	78.3%	+0.738	[+0.659, +0.799]	< 0.001

batteries the paired test was not significant ($t(3) = 1.36$, two-sided $p = 0.268$). LSTM hidden states had PC1 variance shares of 100.0%, 100.0%, 99.9%, and 81.5% across batteries B0005, B0006, B0007, and B0018 respectively, whereas LNN hidden states had PC1 variance shares of 87.1%, 88.9%, 85.3%, and 78.3%.

The second principal component showed weak or no correlation with the target across units (mean $r = -0.079$, range -0.167 to -0.022). Trajectory smoothness analysis yielded moderate monotonicity (mean 0.638) with frequent direction changes (mean 56 changes across 132–168 cycles).

G. Phase-wise Position Importance Results

The phase-wise importance analysis reveals that the late phase of discharge cycles contributes most to capacity/SOH predictions. Across all four battery units, the late phase (80–100% of discharge) showed consistently higher position importance scores compared to early (0–20%) and mid (20–80%) phases. The aggregated mean importance was 0.0014 for the late phase, while early and mid phases showed minimal but non-zero importance (early = 1.4×10^{-7}, mid = 1.3×10^{-5}). The standard deviation across batteries was 0.0008, indicating some variation but consistent dominance of the late phase.

The early and mid phases show non-negligible values when comparing the compact-tau and expanded-tau learned dynamics. The late-phase score is approximately 100–10,000 times larger than the early- and mid-phase scores. Statistical tests showed that late-phase importance exceeded early phase (t = -21.11, p ¡ 1e-30) and mid phase (t = -15.75, p ¡ 1e-30) for the highest-weighted battery B0006, with similar patterns

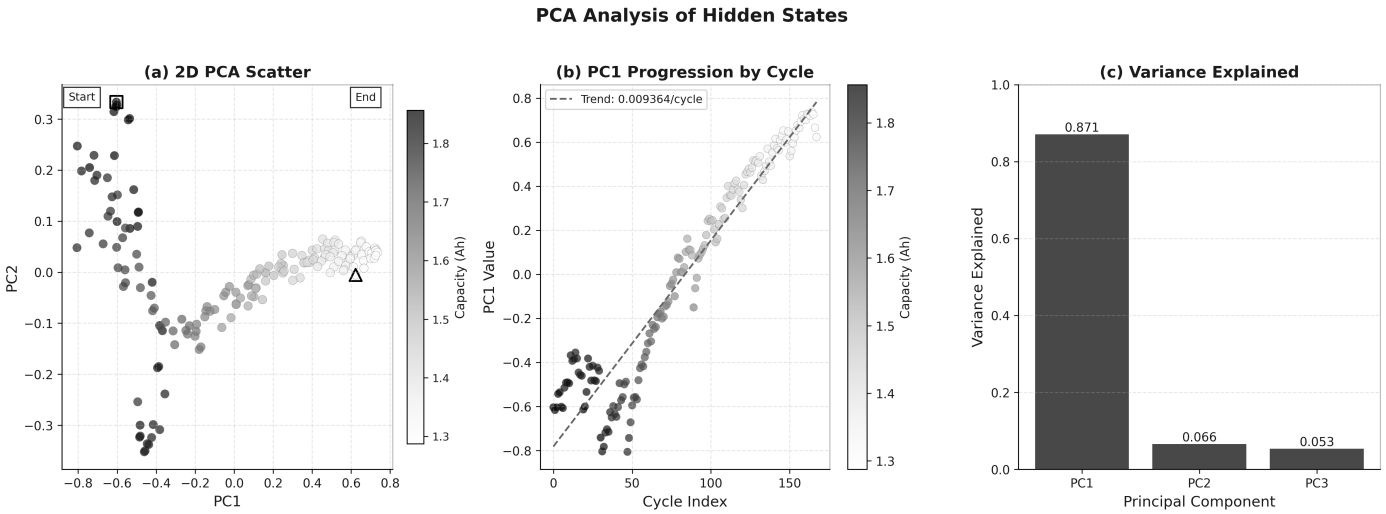

Fig. 1. PCA trajectory dashboard for LNN-TS-16-1L-lr001-p50 on battery B0005. (a) Hidden-state trajectory in the PC1–PC2 plane colored by measured capacity, (b) PC1 progression by cycle index, and (c) variance explained by the first three principal components of the hidden states.

observed across other units. Bootstrap confidence intervals (10,000 resamples) quantified uncertainty as late [0.0013, 0.0015], early [0, 0], mid [0, 0].

Under noise perturbation, phase-importance patterns remained highly stable. For noise levels $\sigma \in [0.05, 0.30]$, Pearson correlations between clean and noisy phase importance consistently exceeded $r > 0.99$ (mean $r = 1.000 \pm 0.000$). Per-battery analysis showed B0018 had the highest late-phase importance (0.0026), followed by B0005 (0.0014), B0006 (0.0011), and B0007 (0.0005). Temporal importance curves show a sharp increase in importance after approximately 80% of discharge completion. Phase importance versus SOH degradation monotonicity revealed that late-phase importance increases consistently from high-SOH through low-SOH stages.

H. Fixed NCP Sparsity Trade-off Results

The NCP sparsity sweep quantifies how fixed sparse topology affects accuracy, robustness, and structural simplicity under the same held-out-unit LOOCV protocol. Dense NCP-TS-16-Dense achieved the best test performance (RMSE 0.0754 $\pm$ 0.0441, MAE 0.0616 $\pm$ 0.0361) with 192 active connections and 0% effective sparsity. Moderate-sparsity NCP-TS-16-1L used 83 active connections (56.8% effective sparsity) and maintained usable but lower accuracy (RMSE 0.0978 $\pm$ 0.0554, MAE 0.0814 $\pm$ 0.0525). High-sparsity NCP-TS-16-Sparse used 33 active connections (82.8% effective sparsity) and showed substantially degraded accuracy (RMSE 0.1807 $\pm$ 0.0626, MAE 0.1572 $\pm$ 0.0517).

As sparsity increased, the number of effective connections and robustness degradation decreased, while prediction error increased. Robustness degradation decreased from 0.0555 for dense NCPs to 0.0013 for high-sparsity NCPs. Figure 2 visualizes this Pareto trade-off between accuracy, noise sensitivity, and structural sparsity.

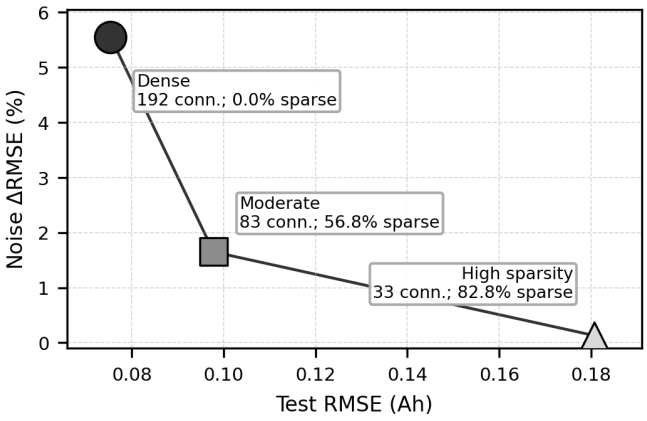

Fig. 2. Pareto view of NCP sparsity trade-offs. Lower-left is preferable for RMSE and noise degradation, while labels report the effective active connection count and sparsity. Increasing fixed sparsity improves noise insensitivity but moves away from the accuracy optimum.

I. Counterfactual Perturbation Analysis Results

The counterfactual analysis reveals distinct sensitivity profiles across input features. Temperature emerged as the most critical feature for capacity/SOH prediction (z-score = +1.13). Relative Time showed the second-highest sensitivity (z-score = +0.37), while Voltage (z-score = -0.73) and Current (z-score = -0.77) demonstrated minimal impact on model predictions under perturbation.

The correlation between tau variance changes ($\Delta\sigma_\tau^2$) and prediction change ($|\Delta\hat{y}|$) was weak (Pearson $r = 0.08 \pm 0.26$, Spearman $\rho = 0.08 \pm 0.13$). While permutation testing ($n = 10,000$) indicates statistical significance ($p < 0.001$), the effect size is negligible with $r^2 \approx 0.006$, explaining less than 1% of variance. Bootstrap resampling ($10,000$ samples) yields a 95% confidence interval of $[0.049, 0.116]$, confirming

the correlation is consistently small across resamples rather than arising from sampling variability.

Feature importance differences are statistically significant based on Wilcoxon rank-sum tests with Bonferroni correction ($\alpha = 0.05$ for 6 comparisons, adjusted $\alpha = 0.0083$). Temperature shows significantly higher sensitivity than Current ($p < 0.001$) and Voltage ($p < 0.001$). All 6 pairwise comparisons are statistically significant at the corrected alpha level.

IV. Discussion

The NASA PCoE dataset is a widely used benchmark for battery prognostics, but it remains a small four-unit dataset and therefore does not support strong claims of generalization across chemistries, operating conditions, or deployment settings. Consequently, we position the present study as a controlled benchmark evaluation using chronological splitting to prevent future-cycle information leakage rather than a claim of universal battery prognostics performance.

While the MVIP-Trans reference architecture achieves competitive performance on pre-processed statistical features, its large parameter count (over 5 million as reported in [3]) raises concerns about computational efficiency and interpretability. Literature reference implementations often report substantially better apparent performance under random splits, consistent with temporal leakage inflating accuracy by allowing models to observe future degradation patterns during training. In our random-split LNN-TS ablation, the same held-out-unit LOOCV test protocol was retained while the train/validation cycles were shuffled. This ablation produced RMSE 0.0865 with higher robustness degradation (0.0617) than the chronological LNN-TS variant in Table II. Thus, randomization did not provide a reliable advantage when test units remained fully held out, reinforcing the decision to treat literature random-split results as protocol reference points rather than direct baselines.

Hybrid architectures like CfC-F2RN can leverage temporal dynamics effectively through hard-attention mechanisms and dual-latent fusion, but the increased architectural complexity adds challenges for analysis and explainability. The learned tau dynamics comparison indicates that battery SOH signals are physically concentrated in late discharge characteristics rather than distributed uniformly across cycles, suggesting the model prioritizes end-of-discharge signals across both compact- and expanded-tau regimes. Hyperparameter analysis indicates that weak monotonicity regularization is helpful but stronger regularization can overconstrain the model, while positional embeddings improve accuracy despite their parameter overhead.

The tau-based confidence estimation results, when confronted with the miscalibration, flat confidence distribution, and insignificant Spearman correlation ($\rho = 0.0003$, $p = 0.99$), indicate that tau variance does not serve as an effective uncertainty indicator for this prognostics task even though a weak but statistically significant negative Pearson correlation suggests a trend where higher confidence (lower tau variance) is weakly associated with lower prediction errors.

The failure of both tau-based and hidden-state-based confidence estimation suggests that extracting reliable prediction confidence from LNN internal dynamics remains an open problem. Unlike the PCA trajectory analysis, which suggested that LNNs learn physically plausible latent representations (mean absolute PC1-target correlation across units $|r| = 0.926$), this representational structure does not directly translate to self-assessment of prediction reliability.

PCA-based trajectory analysis reveals that LNNs learn latent representations that track capacity/SOH progression (mean absolute PC1-target correlation across units $|r| = 0.926$). Unlike LSTM encoder latent spaces that show collapsed representations with PC1 capturing nearly all variance, LNN continuous-time dynamics produce more distributed representations across principal components. However, the LNN–LSTM difference in absolute PC1-target correlation is descriptive rather than statistically significant on four batteries.

Phase-wise importance analysis shows late-phase dominance aligns with physical understanding, as end-of-discharge voltage sag provides the primary signal for capacity degradation, consistent with the explanation that the model learns physically plausible behavior rather than discovering novel degradation signals.

The fixed-sparsity NCP analysis demonstrates that NCP topology enables controlled trade-offs between accuracy, structural simplicity, and noise sensitivity. Dense connectivity provides the best accuracy, while high sparsity provides the lowest noise degradation but substantially worse RMSE. Because masks are fixed rather than learned in this implementation, the NCP results should not be interpreted as learned causal pathway discovery. Instead, they show how imposed sparsity changes model behavior. Sensor noise mitigation remains essential for industrial deployment, especially for the most accurate dense models.

Counterfactual perturbation analysis identifies Temperature and Relative Time as dominant features, consistent with attention-based approaches [9], [11]. These findings align with well-established domain knowledge, indicating that the model learns physically plausible decision strategies where thermal signals and temporal progression dominate over electrical parameters. Convergence across complementary interpretability methods (latent space interpretation, temporal position analysis, perturbation-based importance) suggests that the LNN learns a stable representation consistent with physically plausible patterns.

Against the eight interpretability criteria of Zhang et al. [17], LNNs and NCPs strongly satisfy four criteria (distinguishable representations, standardization, information preservation, transparency/clarity through explicit τ parameters and compact structure), partially satisfy two (temporal patterns identified but uncertainty signals unreliable with ECE 0.656–0.78), and have two framework-incompatible criteria (membership functions, belief distributions). Intrinsic uncertainty quantification

through tau variance and hidden state stability did not perform effectively, suggesting alternative mechanisms are needed.

The LNN literature identifies five key advantages for time series modeling [6]. Our experiments provide empirical evidence for some claims while leaving others unverified. Computational efficiency is supported for the main LNN-TS configuration, which trains 4–20× faster than conventional raw time-series baselines while achieving competitive accuracy with under 10K parameters. Enhanced interpretability is partially confirmed. PCA trajectory analysis shows representations consistent with well-known cell degradation patterns (mean absolute PC1-target correlation across units $|r| = 0.926$), and phase-wise importance aligns with electrochemical domain knowledge. However, intrinsic uncertainty estimation through tau variance and hidden state stability yielded poorly calibrated confidence signals (ECE 0.656–0.78), indicating that interpretable structure does not guarantee reliable self-assessment. Exceptional robustness is partially confirmed. An accuracy-robustness trade-off emerged, since the most accurate raw-capacity models exhibit 4.6–5.5% noise degradation. While NCP sparsity provides controllable robustness (0.13% degradation at high sparsity), this comes at substantial accuracy cost. Small negative robustness values observed for PINN-Raw-16 represent statistical fluctuations from zero-mean noise rather than genuine noise regularization, as models trained without noise augmentation do not learn noise-robust representations. Claims of strong out-of-distribution robustness remain untested. Adaptive dynamics for time-varying environments and balanced handling of short-term versus long-term dependencies remain unverified. Our battery dataset exhibits gradual degradation rather than abrupt distribution shifts, and we did not conduct explicit comparisons between short-term and long-term dependency modeling against LSTM or Transformer baselines. These limitations identify targeted experiments required to validate the remaining claims.

V. Conclusion

This study evaluated liquid neural networks (LNNs) and neural circuit policies (NCPs) for capacity-based battery prognostics under chronological protocols designed to prevent future-cycle information leakage. The benchmark considered accuracy, efficiency, robustness, and interpretability rather than predictive error alone.

On the shared raw-capacity target, compact continuous-time architectures produced competitive results under this protocol: NCP-TS-Dense and LNN-TS obtained RMSE below 0.08 Ah with under 10K parameters, comparable to MVIP-Raw-32 and other baselines which have 5–550× more parameters. The main LNN-TS configuration trained 4–20× faster than conventional raw time-series baselines. These results support compact continuous-time models as efficient candidates for this controlled benchmark, not as evidence of broad cross-chemistry or deployment generalization.

The evaluation also exposed important limitations. An accuracy–robustness trade-off was observed, with the most accurate raw-capacity models showing 4.6–5.5% noise degradation, while sparse NCP variants reduced noise sensitivity at the cost of prediction quality. Because the NCP masks were fixed rather than learned, these sparsity results should be interpreted as architectural trade-offs rather than learned causal pathways.

Intrinsic uncertainty estimation through LNN internal dynamics remains unresolved. Neither tau variability nor hidden-state stability produced well-calibrated confidence signals, with expected calibration errors ranging from 0.656 to 0.78. Thus, although LNN internal states can support diagnostic analysis, they did not provide reliable self-assessment of predictive error in this study.

The interpretability experiments were consistent with physically plausible degradation behavior. Hidden-state trajectories had high absolute correlation with capacity/SOH progression, phase-wise gradients emphasized late discharge stages, and counterfactual perturbations identified temperature and temporal progression as influential features. These convergent results suggest representations correlated with established degradation mechanisms, but they do not establish causal explanations.

Limitations include evaluation on a four-unit dataset, limited hyperparameter exploration for chronological transformer and PINN baselines, and reference comparisons under non-chronological splits. Future work should extend evaluation across larger multi-chemistry datasets, implement fully chronological transformer baselines, and explore calibrated uncertainty mechanisms for continuous-time prognostics models.

Overall, under the evaluated NASA PCoE benchmark protocol, compact continuous-time architectures offer a favorable accuracy–efficiency–interpretability trade-off. Their small parameter counts and explicit time-scale dynamics make LNNs and related sparse continuous-time networks useful candidates for prognostics settings where transparent analysis, robustness, and efficiency must be considered jointly.

References

[1] J. Zhao, X. Feng, Q. Pang, J. Wang, Y. Lian, M. Ouyang, and A. F. Burke, "Battery prognostics and health management from a machine learning perspective," *Journal of Power Sources*, vol. 581, p. 233474, 2023.

[2] F. Wang, Z. Zhai, Z. Zhao, Y. Di, and X. Chen, "Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis," *Nature Communications*, vol. 15, 2024.

[3] T. Bai and H. Wang, "Convolutional transformer-based multiview information perception framework for lithium-ion battery state-of-health estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[4] X. Li, Y. Hu, H. Wang, Y. Liu, X. Liu, and H. Lu, "A closed-form continuous-depth neural-based hybrid difference features re-representation network for rul prediction," *Reliability Engineering and System Safety*, vol. 253, p. 110540, 2025.

[5] R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu, "Liquid time-constant networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 7657–7666, may 2021.

[6] J. Leng, T. Zhu, J. Li, B. Wang, B. Yang, X. Li, Q. Liu, X. Chen, and L. Wang, "Liquid neural network in smart manufacturing: A new opportunity," *Advanced Engineering Informatics*, vol. 70, p. 104222, 2026.

[7] A. Barrantes, J. J. Montero-Jiménez, and R. Vingerheods, "A multi-metric trade-off framework for ai model selection for diagnosis and prognosis in predictive maintenance," *Tecnología en Marcha*, 2026.

[8] G. Kannan, J. Chen, L. Zhou, Y. Zheng, and J. Leng, "When smaller wins: Dual-stage distillation and pareto-guided compression of liquid neural networks for edge battery prognostics," *arXiv preprint arXiv:2601.06227*, 2026.
[9] Z. Lei, Y. Su, K. Feng, and G. Wen, "Interpretable operational condition attention-informed domain adaptation network for remaining useful life prediction under variable operational conditions," *Computers & Industrial Engineering*, vol. 195, 2024.
[10] B. Ersöz, S. Oyucu, A. Aksöz, Ş. Sağıroğlu, and E. Biçer, "Interpreting cnn-rnn hybrid model-based ensemble learning with explainable artificial intelligence to predict the performance of li-ion batteries in drone flights," *Applied Sciences*, vol. 14, no. 23, 2024.
[11] F. Rahmanian, R. M. Lee, D. Linzner, K. Michel, L. Merker, B. B. Berkes, L. Nuss, and H. S. Stein, "Attention towards chemistry agnostic and explainable battery lifetime prediction," *npj Computational Materials*, vol. 10, no. 1, 2024.
[12] B. Saha and K. Goebel, "Battery data set." NASA Prognostics Data Repository, NASA Ames Research Center, 2007.
[13] M. Lechner, R. Hasani, A. Amini, and D. Rus, "Neural circuit policies enabling auditable autonomy," *Nature Machine Intelligence*, vol. 2, no. 10, pp. 642–652, 2020.
[14] M. Lechner, R. Hasani, and A. Amini, "Neuronal circuit policies," *arXiv preprint arXiv:1803.08554*, 2018.
[15] P. Kakati, D. Dandotiya, and R. R. Singh, "Estimating battery state of health using dconvblstm and modified particle filter under complex noise," *Journal of Energy Storage*, vol. 108, p. 115036, 2025.
[16] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
[17] Z. Zhang, Y. Qu, Y. Liu, M. Li, H. Li, and W. He, "A new interpretable state-of-health assessment model for lithium-ion batteries with multi-dimensional adaptability optimization," *Energy Science & Engineering*, vol. 12, no. 12, pp. 5647–5664, 2024.
[18] M. L. Baptista, M. Mishra, E. Henriques, and H. Prendinger, "Using explainable artificial intelligence to interpret remaining useful life estimation with gated recurrent unit," in *Proceedings of the Annual Conference of the PHM Society*, vol. 16, p. 4124, 2024.