Journal of Electronic Research and Application, 2026, Volume 10, Issue 8
http://ojs.bbwpublisher.com/index.php/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

LSTM-Based Operating-State Forecasting and Risk Warning for Small-Scale Battery Energy Storage Systems Using Multi-Source Aging Data

Qian Wang[1]*, Xing Wan[2], Lihua Luo[3]

[1]School of Artificial Intelligence, Leshan Vocational and Technical College, Leshan, China
[2]School of Intelligent Manufacturing, Leshan Vocational and Technical College, Leshan, China
[3]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia

***Corresponding author:** Qian Wang, wqian8603@gmail.com

Abstract: Senseless computing resources are needed to achieve reliable forecasting in small-scale battery energy storage. In this research, the authors present an LSTM-based framework to normalize NASA and CALCE aging data, separate the data for battery-level prediction, predict 5-cycle-ahead state of health (SOH), and transform residual and health signals to warnings. In this work, LSTM was tested against the NASA and CALCE datasets using a common seven-model protocol and obtained competitive cross-cell performance (NASA: MAE 0.01219, RMSE 0.01619; CALCE: MAE 0.01149, RMSE 0.02331). Controlled anomaly tests detected all severe abrupt drops, and low-SOH risk was separated from model unexplained deviations by replay on CALCE CS2_38. Results of a rerun with an independent CPU were consistent with the original rankings of models and the statistical interpretation. The result is that the framework offers a simple and repeatable foundation for battery-state monitoring, but validation of the field-fault is still required.

Keywords: Battery energy storage system; Long short-term memory; State of health; Time-series forecasting; Risk warning

Online publication: Sept 3, 2026

1. Introduction

Battery energy storage systems (BESSs) support renewable-energy integration, yet many small-scale systems lack regular capacity tests and advanced sensing. Routine cycling records can therefore provide early evidence of degradation through state-of-health (SOH) forecasting. SOH estimation methods include experimental, model-based, and data-driven approaches [1]. LSTM networks are designed to capture long-term dependencies, and recent battery studies have combined LSTM with engineered health indicators, hybrid feature extraction, and transfer learning [2–5]. However, reported comparisons are often weakened by

inconsistent datasets, warning rules that conflate low SOH with model error, and cycle-level leakage across train-test splits.

In this study, the new recurrent unit is not developed, but a compact evidence chain is built instead. The software standardizes NASA and CALCE data under explicit quality checks, separates batteries for training, validation, and testing, compares seven models under training, validation, and testing with the same windows, horizons, and repeats, links forecasts to residual and/or health-based warnings, and verifies run-level reproducibility on an independent CPU computer. All models were implemented in PyTorch [6].

2. Materials and methods

This section describes the data sources, preprocessing procedure, forecasting model, and risk-warning method used in this study. It first introduces the multi-source battery aging data and cycle-level feature construction, followed by the LSTM forecasting framework, baseline models, evaluation metrics, and residual-based warning rules.

2.1. Data and preprocessing

The NASA dataset comprised B0005, B0006, B0007, and B0018 (636 processed discharge cycles), and the CALCE CS2 set comprised CS2_35–CS2_38 (3,915 processed cycles) [7,8]. The CALCE CS2 dataset has been widely adopted for battery prognostics and SOH estimation studies [9–11]. The cycle-level data were mapped to a common schema containing capacity, voltage, current, time, and available temperature fields.

Processing included field and unit checks, discharge-cycle aggregation, duplicate removal, continuity checks, missingness screening, and physical-range audits. SOH was calculated as $SOH_t = C_t / C_0$, where C_0 is the first valid capacity. Cell-level splitting prevented adjacent observations from the same cell from appearing in both training and test sets.

2.2. Forecasting and warning framework

The input consists of the cycle-level feature vectors from the previous W cycles, and the prediction target is the SOH at cycle t+H. The main experiments used W = 12 and H = 5. The LSTM estimates SOH change and then generates the future SOH value using Equation (1). The model was trained using mean squared error and the Adam optimizer. Early stopping was applied according to the validation loss, and the checkpoint with the lowest validation loss was retained. The baselines were persistence, SVR, random forest, vanilla RNN, GRU, and TCN.

$$\widehat{SOH}_{t+H} = SOH_t + \widehat{\Delta SOH}_{t+H}$$

$$e_{t+H} = \widehat{SOH}_{t+H} - SOH_{t+H} \tag{1}$$

The mean squared error, Adam, validation set early stopping, and the best checkpoint were employed for training. The validation set was used to calibrate the residual thresholds. The forecast residual was defined as $e_{t+H} = \widehat{SOH}_{t+H} - SOH_{t+H}$. Moderate and strict thresholds were applied to the absolute residual. $|e_{t+H}|$ to define the attention and warning states, respectively. A persistent warning was triggered when the strict threshold was exceeded repeatedly. An additional rule, $SOH \le 0.80$, was used to identify low battery health independently of unexplained model deviations. Thus, low SOH and abnormal forecast residuals were retained as two distinct

risk sources. See **Figure 1**.

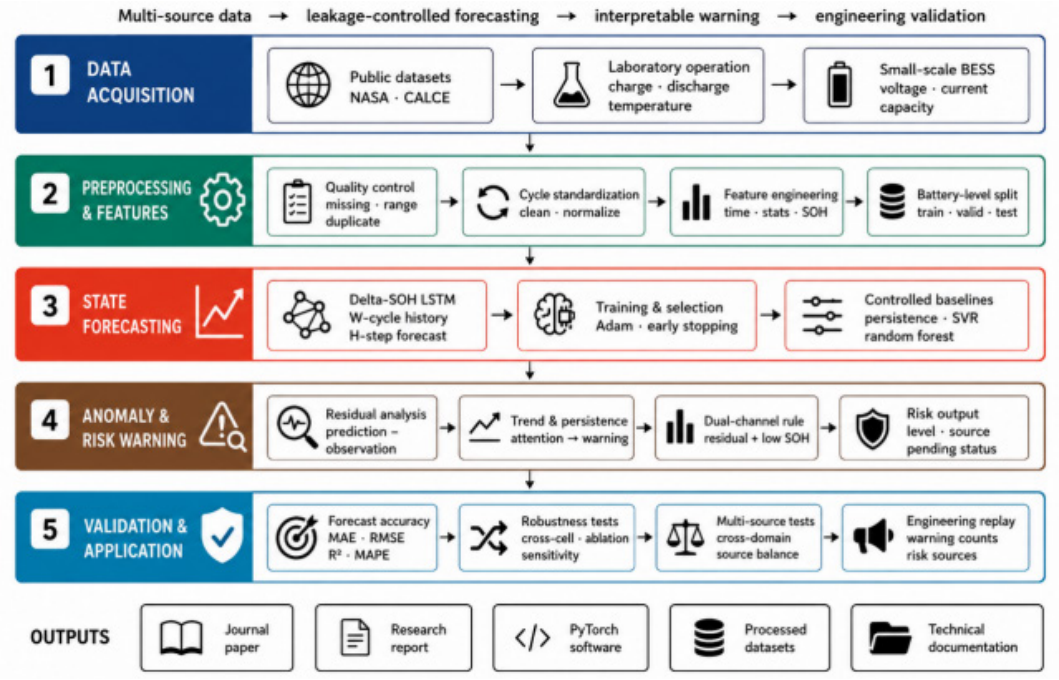

Figure 1. Workflow for multi-source preprocessing, leakage-controlled forecasting, warning calibration, and validation.

3. Experimental design

NASA held-out training was on B0005 / B0006, validation was B0007, and test was B0018. Cross-cell robustness was based on 4 folds and 5 seeds per fold. Ten seeds of CS2_35/36 were used for training, while a different ten seeds of CS2_37 were used for validation, and a different ten seeds of CS2_38 were used for testing for all CALCE testing. NASA-only and joint-source training involved the same CALCE test cell. The metrics used were MAE, RMSE, R^2, and MAPE. Holm correction at $\alpha = 0.05$ was used for paired tests.

Since the datasets do not have known operational errors, abrupt decreases, progressive degradation, and transients were only introduced to evaluate algorithmic response. There was a total of 210 runs in the full seven-model benchmark. The data used here were processed identically; the battery split was the same, and the seeds and settings were the same; an independent CPU rerun of the data resulted in 202 runs within 10^{-6}, and eight TCN runs had small differences that did not affect any ranking or significance conclusion. No sampling-level predictions were kept, and hence only run-level reproducibility is claimed.

4. Results and discussion

This section presents and discusses forecasting and risk-warning results. It first evaluates the predictive performance of the LSTM against the baseline models and then examines the effects of different data-source configurations, cross-domain generalization, and residual-based warning performance.

4.1. Forecasting performance

LSTM showed the best performance on all metrics in the fixed NASA holdout (MAE 0.01489; RMSE

0.02002; R^2 0.9185), with a 10.9% reduction in RMSE compared to the persistence method. When analyzing both data sets within the same repeated protocol (**Table 1**), no model was found to dominate both. SVR achieved minimum CALCE MAE, GRU achieved minimum CALCE RMSE, and NASA MAE and random forest achieved minimum NASA RMSE. LSTM was always in the lead, and it trained faster than the other deep sequence models, having 29,729 parameters and a state dictionary of 119.6 KiB. So, the evidence suggests that LSTM is a viable, small core forecasting system; it's not best all the way. See **Table 1**.

Table 1. Principal unified-protocol results (mean ± SD where repeated)

Scenario	Model	MAE	RMSE	R^2
CALCE	LSTM	0.01149 ± 0.00016	0.02331 ± 0.00012	0.9832
CALCE	GRU	0.01142 ± 0.00023	0.02324 ± 0.00004	0.9833
CALCE	RNN	0.01165 ± 0.00021	0.02336 ± 0.00013	0.9831
CALCE	TCN	0.01153 ± 0.00020	0.02329 ± 0.00010	0.9832
CALCE	SVR	0.01125	0.02328	0.9833
NASA	LSTM	0.01219 ± 0.00413	0.01619 ± 0.00464	0.9614
NASA	GRU	0.01189 ± 0.00391	0.01597 ± 0.00432	0.9624
NASA	RNN	0.01250 ± 0.00420	0.01644 ± 0.00453	0.9594
NASA	TCN	0.01223 ± 0.00408	0.01616 ± 0.00448	0.9615
NASA	Random forest	0.01238	0.01579	0.9640

It was more important to have domain shift than to have an architecture choice. On the CALCE test cell, CALCE-only training achieved MAE 0.01149 and RMSE 0.02331, compared with 0.01758 and 0.03258 for NASA-only training (Holm-adjusted $p = 0.0234$). Joint training showed very close in-domain performance (MAE 0.01174; RMSE 0.02317); unfortunately, not all metrics improved. However, source diversity does not necessarily transfer automatically from pooled data and will need explicit validation.4.2.

4.2. Warning response and reply

The strict detection was 100% at 6% severity, and higher for abrupt deviations than for progressive and transient deviations, but the attention channel enhanced sensitivity (**Figure 2**). In the 1,010-cycle CS2_38 replay, forecasting achieved MAE 0.01151, RMSE 0.02325, and R^2 0.98329. 371 critical labels were activated by SOH ≤ 0.80. 18 strict residual exceedances – 15 were below the low SOH, and three were above. This decomposition allows a battery with potentially degraded performance but one having a known characteristic to be distinguished from an unexpected model error. The anomaly results quantify only controlled sensitivity, and not real-fault precision or recall. See **Table 2**.

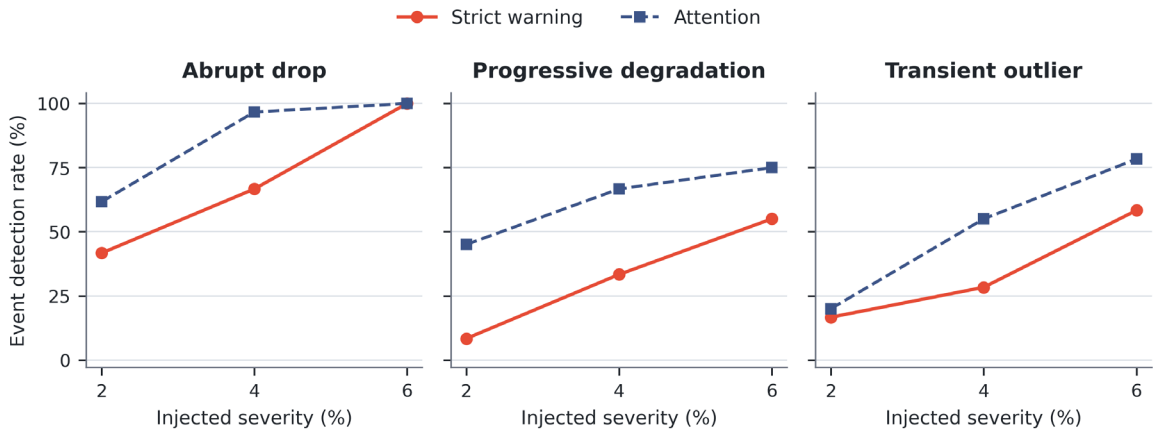

Figure 2. Strict-warning and attention-level event detection by simulated anomaly type and severity.

Table 2. Warning response and CS2_38 replay summary

Condition	Strict	Attention/share	Interpretation
Abrupt drop, 6%	100.0	100.0	Controlled event
Progressive, 6%	55.0	75.0	Controlled event
Transient, 6%	58.3	78.3	Controlled event
CS2_38 low SOH	371 cycles	36.73%	Health rule
CS2_38 strict residual	18 cycles	1.78%	Model deviation

Its major drawbacks are lab aging data, simulated and not proven fault, replaying offline single cells, and lack of saved sample-level prediction for independent metric recalculation. Field deployment should include a record of predictions, incorporate BMS data and maintenance records, recalibrate thresholds, and document expert-confirmed event recall, false-alert rates, and warning lead time.

4.3. Practical implications

The evidence comes together and suggests three implications. For one, when the data boundary of a compact neural model is not explicit, it is helpful. The cross-domain experiments reveal that the variations in cell design, cycling protocol, measuring instruments, as well as features can be as large as the performance disparities between LSTM, GRU, TCN, and conventional regressors. A deployment pipeline should thus be able to determine where each new record comes from, be able to detect feature drift, and initiate recalibration if the target system is outside the validation domain. Second, the semantics of warning should be maintainable. Split reporting of SOH and residual exceedance will enable the operator to distinguish between normal aging and an unexpected trajectory, and attention and strict thresholds will give different levels of sensitivity in one joint alarm.

Thirdly, reproducibility should be considered as an operational control. The independent CPU rerun validated that the central ranking and statistical interpretation were independent of any one computer, but there were minor differences in TCN. Each forecast should be stored to a field implementation with its model version, scaler, threshold, input window, and decision about the event. This would enable reconstruction, post-event review, and controlled updates to models at the sample level. In this respect, these practices are suitable for small storage installations and provide a more valuable monitoring component for the proposed framework than an accuracy demonstration on its own.

5. Conclusion

A small LSTM framework was integrated with multi-source data quality, leakage-controlled SOH forecasting, fair model comparison, and interpretable warning rules. Results at NASA and CALCE indicate competitive forecasting, high sensitivity to extreme sharp turnaround errors, and good separation between low health and unexplained residual risk. The main conclusions were maintained in the rerun by an independent CPU. However, operational claims must wait for prospective validation with prospective BMS and fault records.

Data and code

NASA and CALCE public archives [7,8]. The source code will be made publicly available in a permanent repository upon acceptance.

Author contributions

Q.W. conducted the experiments and drafted the manuscript. X.W. conceived and designed the study. L.H.L. analyzed the data and contributed to manuscript preparation. All authors reviewed and approved the final manuscript.

Funding

Leshan Engineering Technology Research Center for Energy Storage Materials and Systems through the project “Research on Operation State Perception and Risk Early Warning Methods for Small-Scale Energy Storage Systems” (Project No.: CNY202607)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Yang K, Zhang L, Zhang Z, et al., 2023, Battery State of Health Estimate Strategies: From Data Analysis to End-Cloud Collaborative Framework. Batteries, 9(7): 351.

[2] Hochreiter S, Schmidhuber J, 1997, Long Short-Term Memory. Neural Computation, 9(8): 1735–1780.

[3] Zhao S, Luo L, Jiang S, et al., 2023, Lithium-Ion Battery State-of-Health Estimation Method Using Isobaric Energy Analysis and PSO-LSTM. Journal of Electrical and Computer Engineering, 2023: 5566965.

[4] Feng S, Song M, Lin Y, et al., 2024, Convolutional Neural Network-Long Short-Term Memory-Based State of Health Estimation for Li-Ion Batteries Under Multiple Working Conditions. Energy Technology, 12(2): 2301039.

[5] Lin M, Hu D, Meng J, et al., 2025, Transfer Learning-Based Lithium-Ion Battery State of Health Estimation with Electrochemical Impedance Spectroscopy. IEEE Transactions on Transportation Electrification, 11(3): 7910–7920.

[6] Paszke A, Gross S, Massa F, et al., 2019, PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems, 32.

[7] NASA 2026, Li-Ion Battery Aging Datasets, visited on 29 Jul 2026, https://data.nasa.gov/dataset/li-ion-battery-

aging-datasets

[8] Center for Advanced Life Cycle Engineering, n.d., CALCE Battery Data: CS2 Battery, University of Maryland, visited on 29 Jul 2026, https://calce.umd.edu/battery-data.

[9] He W, Williard N, Osterman M, et al., 2011, Prognostics of Lithium-Ion Batteries Based on Dempster-Shafer Theory and the Bayesian Monte Carlo Method. Journal of Power Sources, 196(23): 10314–10321.

[10] Xing Y, Ma E, Tsui K, et al., 2013, An Ensemble Model for Predicting the Remaining Useful Performance of Lithium-Ion Batteries. Microelectronics Reliability, 53(6): 811–820.

[11] Williard N, He W, Osterman M, et al., 2013, Comparative Analysis of Features for Determining State of Health in Lithium-Ion Batteries. International Journal of Prognostics and Health Management, 4(1): 1–7.