Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD37070

When Labels Shift: Few-Shot Fine-Tuning Versus Unsupervised Domain Adaptation for Cross-Dataset Battery Cycle-Life Prediction

Pengcheng Zhao

School of Computer Science and Technology, Beijing Jiaotong University, Weihai, China
Zpc0704@outlook.com

Abstract. Lithium-ion batteries are widely deployed in electric vehicles and energy-storage systems, and accurate early prediction of cycle life is critical for safety, warranty cost, and second-life planning. Models trained on one dataset, however, rarely transfer to new chemistries, test protocols, or laboratories, and re-collecting full-life labels for every new scenario is prohibitively slow. This study investigates which transfer paradigm is effective for cross-dataset battery cycle-life prediction under such distribution shift. A CNN-LSTM regressor and two public datasets, Stanford (41 NMC532 cells) and Tongji (130 NCA/NCM cells), are used to run a six-experiment ladder covering in-domain baselines, direct transfer, few-shot fine-tuning, multi-kernel MMD alignment, MMD pre-training plus fine-tuning, and adversarial DANN, in both transfer directions, with strict isolation between adaptation and test cells and from-scratch controls for every supervised variant. Direct transfer collapses (MAE 517 cycles), driven mainly by a threefold gap in mean cycle life (220 versus 685 cycles). Fine-tuning with only five target cells reduces MAE by 81% to about 98 cycles, while unsupervised alignment methods (MMD 394, DANN 474) clearly underperform, and MMD pre-training adds no gain. The underperformance of feature alignment is attributed to label shift between the two fleets, and a pre-train-then-fine-tune recipe is recommended for transferable battery prognostics. Mixed-domain random splits are also shown to overstate model quality through domain leakage, so per-domain evaluation is advised.

Keywords: lithium-ion battery, cycle-life prediction, CNN-LSTM, transfer learning, label shift

1. Introduction

Lithium-ion batteries power electric vehicles and stationary energy storage. Their degradation is strongly nonlinear and sensitive to temperature and charge or discharge rates, so early prediction of full cycle life from the first few dozen cycles has become a central problem in battery health management. Data-driven models have made clear progress on this task, from early-prediction benchmarks on large fast-charging datasets [1] to closed-loop charging optimization [2] and health-estimation pipelines [3]. Most of these results, however, are obtained within a single dataset. The

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.LD37070

economic incentive for early prediction is substantial: cell manufacturers spend months on aging channels before a design can be qualified, and fleet operators must triage thousands of cells for warranty and second-life use long before full-life labels exist.

Deployment rarely respects that boundary. A model trained on one cell type or protocol often fails on batteries from another laboratory, and collecting new full-life labels takes months per cell. Transfer learning is the natural remedy [4], but the right paradigm is unclear. Unsupervised domain adaptation methods, including discrepancy-based alignment [5, 6] and adversarial training [7, 8], assume that feature distributions differ while label distributions stay comparable. Battery fleets violate this assumption: mean cycle life can differ threefold, a case of label shift [9].

This study therefore asks which transfer paradigm remains effective under label shift. A six-experiment ladder is conducted on two public datasets with a shared CNN-LSTM regressor, covering baselines, direct transfer, few-shot fine-tuning, two unsupervised alignment methods, and a combined variant. Both transfer directions are examined so that the conclusions do not rest on a favorable source-target pairing. The findings are expected to clarify how small labeling budgets should be spent and to inform the deployment of transferable prognostic models in battery management systems.

2. Data and methodology

2.1. Datasets

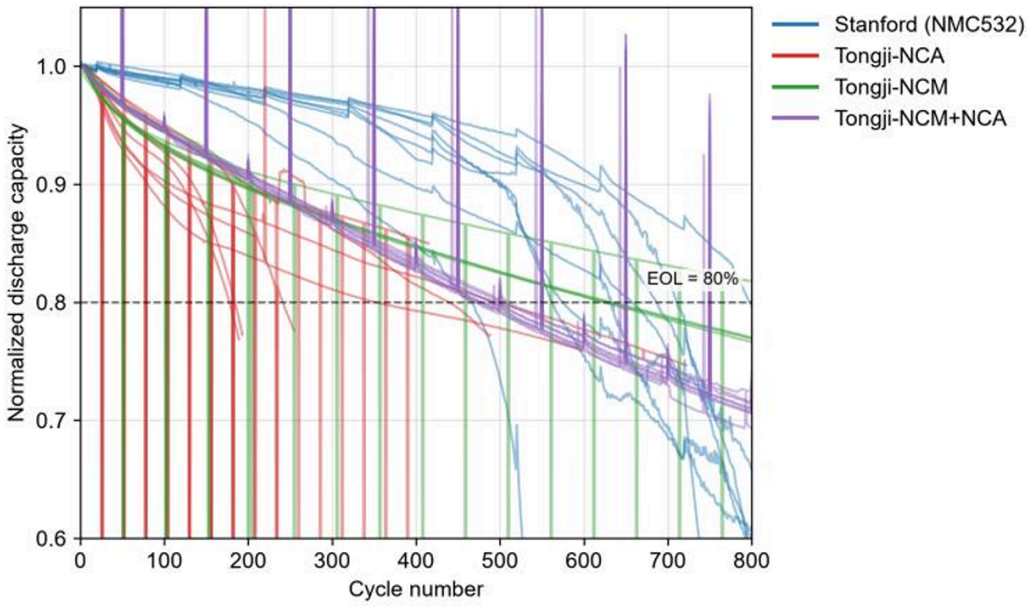

Figure 1. Normalized capacity-fade curves of representative cells, eight per subset

Two public cycle-life datasets are used. The Stanford set contains 41 NMC532 cells cycled under various protocols, with a mean life of about 685 cycles and a tight distribution (standard deviation about 121). The Tongji set [10] contains 130 commercial 18650 cells of three cathode systems (NCA, 66 cells, NCM, 55, and NCM+NCA blend, 9), tested at 25-55 degrees Celsius and multiple current rates, with a lower and more dispersed mean life of about 220 cycles. Representative normalized capacity-fade curves are shown in Figure 1. The Tongji fleet is also internally heterogeneous: the NCA subset is long-tailed, the NCM subset is broadly scattered, and the small blend subset clusters near 487 cycles with almost no spread. The two domains overlap in feature

DOI: 10.54254/2755-2721/2026.LD37070

space while their probability mass over the label range differs systematically. The threefold gap in mean label scale, together with differing protocols, creates both feature shift and label shift between the domains.

2.2. Cycle-level features and task definition

Raw voltage-current-temperature traces are aggregated per cycle into 11 summary features: charge and discharge capacity, voltage mean, minimum and maximum, current mean, minimum and maximum, temperature mean and maximum, and cycle duration. Cycle life is defined as the first cycle where discharge capacity falls to 80% of its initial value, taking the median of the first five cycles to suppress noise, a convention that follows common practice in public battery benchmarks. The model input is the sequence of the first 20 cycles, and the output is the scalar total cycle life. Features are standardized with a scaler fitted on the training split only. Because the Tongji records contain no temperature columns, cross-dataset experiments drop the two temperature features and use the remaining nine shared features, avoiding an artificial domain gap. The 20-cycle window corresponds to roughly the first 3-10% of the total life of a cell, so the prediction task stays genuinely early-stage. The chosen feature set keeps the degradation signal of capacity while retaining operating-condition information, and it is cheap to compute in an online setting.

2.3. CNN-LSTM model

The regressor stacks two Conv1d layers with batch normalization and ReLU activation, which extract local degradation patterns along the cycle axis, an LSTM [11] with hidden size 128 that models temporal dependencies, and an MLP head on the last-step hidden state that outputs the life estimate. The hidden state also serves as the alignment hook for the domain-adaptation methods, so that every transfer strategy compared here operates on exactly the same representation, and differences in outcome can be attributed to the adaptation paradigm rather than to the architecture. The network has about 123k parameters, lightweight enough for embedded battery-management deployment. Training uses the AdamW optimizer with a SmoothL1 loss for robustness to long-lived outliers, and the checkpoint with the best validation error is kept. Performance is reported as MAE, RMSE and R-squared in cycles.

3. Experiments and results

3.1. In-domain baseline and a leakage caveat

With mixed-domain random splits of 64/16/20 over the 160 usable cells, the 20-cycle window reaches MAE 63-74 cycles and R-squared of 0.83-0.86, while a 100-cycle window is markedly worse (MAE 129, R-squared 0.51). Longer windows inflate the input dimension faster than they add signal at this sample size, and the first 20 cycles already contain the fade slope and early voltage drift that correlate with final life. This result also cautions against the common assumption that more history is always better in battery prognostics. The strong mixed-domain score is also partly an artifact: the model can recognize which dataset a cell comes from and exploit that prior. Single-source validation yields R-squared near zero, which exposes domain leakage in the mixed protocol and motivates studying genuine cross-domain generalization.

DOI: 10.54254/2755-2721/2026.LD37070

3.2. Direct transfer: quantifying domain shift

Training on one dataset and testing on the other without adaptation fails dramatically. Tongji-to-Stanford reaches MAE 517 cycles (R-squared -18.84), worse than predicting the target mean (465 cycles), and Stanford-to-Tongji reaches MAE 280 cycles (R-squared -1.20). The mechanism is clear: the model internalizes the source prior of about 220 cycles and systematically underestimates the long-lived Stanford cells. The naive baseline, which predicts the mean life of the target training cells for every test cell, marks the floor that any useful model must beat, and direct transfer falls below it. Direct transfer therefore quantifies the domain gap rather than solving it.

3.3. Few-shot fine-tuning

Target cells are ordered by a fixed seed, the first k cells are used for fine-tuning and all remaining cells for testing, with strict isolation. A from-scratch control trained on the same k cells separates the transfer benefit from the value of the target data itself. Three adaptation budgets are compared, k = 5, 10 and 20 target cells, corresponding to about 12%, 24% and 49% of the Stanford fleet, the reverse direction uses k up to 40 because the Tongji fleet is larger. Fine-tuning updates all parameters at a learning rate of 5e-4 for 150 epochs. Five labeled Stanford cells already reduce MAE from 517 to about 98 cycles, a drop of 81%, with saturation at k of 10-20 (MAE 99-106). This saturation pattern indicates that the marginal value of additional target labels drops quickly after the first few cells. The controls collapse completely (MAE 534-587, R-squared down to -25), which proves that the gain comes from source-domain priors rather than the target data alone. Fine-tune strength matters: a weaker schedule (1e-4, 50 epochs) reaches only MAE 205, indicating that the output layer needs enough training to re-calibrate to the threefold longer life scale. Reverse transfer is far weaker: with 40 target cells the error reaches MAE 131, while smaller k values collapse, and in some rows the from-scratch control wins. This source-target asymmetry is itself a useful finding.

3.4. Unsupervised domain adaptation

The first adaptation experiment applies multi-kernel maximum mean discrepancy on the hidden state, with the total loss defined in Equation (1), using target features but no target labels.

$$L = Lpred + \lambda \cdot LMMD \tag{1}$$

The weighting coefficient is swept over 0.05, 0.1 and 0.5, and the adversarial weight follows an increasing schedule up to 0.5. Only $\lambda = 0.1$ improves over direct transfer (MAE 394), $\lambda = 0.05$ or 0.5 degrades beyond 665 cycles, so the method is highly sensitive and non-monotonic. This sensitivity means the coefficient must be re-tuned for every dataset pair, which limits the practical value of the method when no target labels are available for validation. The combined variant adds MMD pre-training before fine-tuning and reaches MAE 100-113, statistically indistinguishable from plain fine-tuning, so the pre-training stage gives no measurable gain. The adversarial variant trains a domain-adversarial network with a gradient-reversal layer and reaches MAE 474, barely better than the naive baseline. Table 1 and Figure 2 summarize the comparison: few-shot fine-tuning is clearly best, MMD is a distant second, and the adversarial method is barely better than the naive baseline.

DOI: 10.54254/2755-2721/2026.LD37070

Table 1. Transfer method comparison, Tongji to Stanford

Method	Target labels	Target MAE (cycles)	vs direct (517)
Naive-mean baseline	-	464.9	-
Direct transfer	no	517.0	baseline
MMD (lambda = 0.1)	no	394.0	-24%
DANN	no	474.2	-8%
MMD pretrain + fine-tune (k = 5)	5 cells	100.4	-81%
Fine-tune (k = 5)	5 cells	98.3	-81%

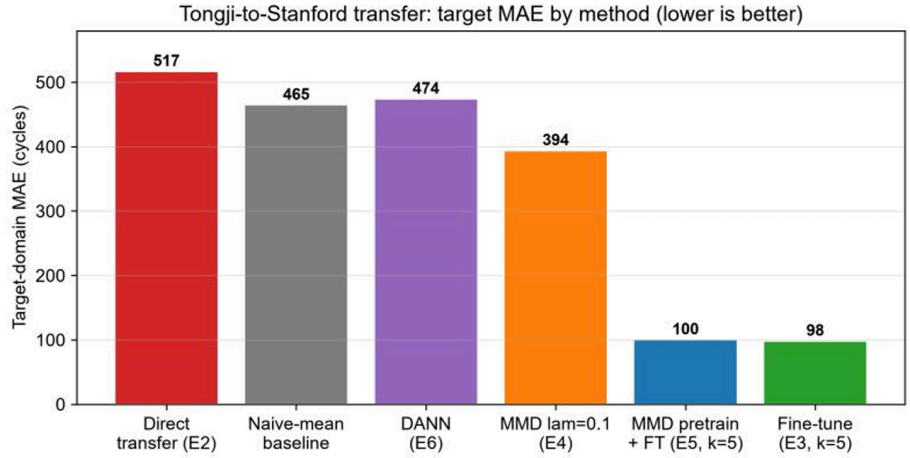

Figure 2. Target-domain MAE by transfer method, lower is better

4. Discussion

Label shift explains why unsupervised alignment underperforms. MMD and adversarial training align the feature distribution, but the dominant difference between the two fleets lies in the label distribution, with mean lives of 220 versus 685 cycles. Aligning features carries no information about this scale gap and can even distort features that are useful for life discrimination [9]. Where label shift dominates, a handful of target labels is worth more than any feature-alignment machinery: five cells, about 12% of the target fleet, delivered an 81% error reduction. From a deployment perspective, such a labeling budget is modest: a single formation channel occupied for a few weeks already provides enough labeled cells.

The tension between MAE and R-squared needs careful handling. Stanford lives cluster tightly (mean 685, standard deviation about 121), so predicting the mean is already a strong baseline, and MAE 98 leaves R-squared slightly negative even though the relative error of about 14% is usable in engineering practice. On concentrated targets, MAE and relative error should be reported alongside R-squared to avoid misleading conclusions, and all experiments here follow this rule. Evaluation protocols matter as well: mixed-domain random splits systematically overstate model quality, and battery studies should report per-domain or group-wise splits, ideally with the dataset identity held out entirely, so that reported scores reflect generalization rather than domain recognition.

DOI: 10.54254/2755-2721/2026.LD37070

Two further observations inform practice. First, outliers persist: one Tongji-NCA cell under condition CY45-05_1 lived 565 cycles while its 27 siblings died near 50, and the model misses it by roughly a factor of ten, although the SmoothL1 loss keeps aggregate metrics robust. Second, the transfer asymmetry has a practical reading: transfer works best from a large, diverse, short-lived fleet toward a small long-lived one, which is fortunately the common industrial case, since historical laboratory data are abundant while the new deployment is the scarce-label target. Where the reverse situation is unavoidable, more target labels, a lower fine-tuning rate, and a frozen convolutional front-end are advisable. Several limitations remain. Both datasets come from laboratory aging rather than field operation, cycle-level summaries discard curve-shape information, the Stanford fleet is small, and the label-shift diagnosis is supported by quantitative but indirect evidence, which a target-domain reweighting experiment could test directly.

5. Conclusion

This study set out to determine which transfer-learning paradigm is effective for cross-dataset battery cycle-life prediction when the label distributions of the source and target fleets differ substantially. A CNN-LSTM regressor was trained and evaluated through a six-experiment ladder on two public datasets, Stanford and Tongji, with strict isolation between adaptation and test cells and from-scratch controls for every supervised variant.

The results support three conclusions. First, direct transfer is not a viable option: the source prior of about 220 cycles produces a systematic underestimate on the long-lived target fleet, and the error exceeds even a naive mean baseline. Second, unsupervised domain adaptation recovers only part of the gap. Multi-kernel MMD improves over direct transfer only within a narrow range of its weighting coefficient, adversarial training barely beats the naive baseline, and adding MMD pre-training before fine-tuning gives no measurable gain. The common explanation, supported by the threefold difference in mean cycle life, is label shift: feature alignment cannot transmit information about a label scale it never sees. Third, few-shot fine-tuning is the clear winner. Five labeled target cells reduce the error by 81%, the from-scratch controls confirm that the gain originates in the source-domain priors, and the effect saturates quickly, so large labeling budgets are unnecessary.

For practitioners, the recipe is simple: pre-train on a large and diverse fleet, then fine-tune briefly on a handful of target cells, rather than investing in label-free alignment. Future work will test cross-chemistry transfer within the Tongji subsets to separate chemistry from protocol effects, add attention over cycles, incorporate impedance-spectroscopy features, and attach uncertainty estimates so that the approach can be trusted in deployed battery management systems.

References

[1] Severson, K. A., Attia, P. M., Jin, N., et al. (2019). Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy, 4*, 383–391.

[2] Attia, P. M., Grover, A., Jin, N., et al. (2020). Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature, 578*, 397–402.

[3] Roman, D., Saxena, S., Robu, V., et al. (2021). Machine learning pipeline for battery state-of-health estimation. *Nature Machine Intelligence, 3*, 447–456.

[4] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*, 1345–1359.

[5] Long, M., Cao, Y., Wang, J., et al. (2015). Learning transferable features with deep adaptation networks. *Proceedings of the 32nd International Conference on Machine Learning, 37*, 97–105.

[6] Gretton, A., Borgwardt, K. M., Rasch, M. J., et al. (2012). A kernel two-sample test. *Journal of Machine Learning Research, 13*, 723–773.

DOI: 10.54254/2755-2721/2026.LD37070

[7] Ben-David, S., Blitzer, J., Crammer, K., et al. (2010). A theory of learning from different domains. *Machine Learning, 79*, 151–175.
[8] Ganin, Y., Ustinova, E., Ajakan, H., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research, 17*, 1–35.
[9] Lipton, Z. C., Wang, Y. X., Smola, A. (2018). Detecting and correcting for label shift with black box predictors. *Proceedings of the 35th International Conference on Machine Learning, 80*, 3122–3130.
[10] Zhu, J., Wang, Y., Huang, Y., et al. (2022). Data-driven capacity estimation of commercial lithium-ion batteries from voltage relaxation. *Nature Communications, 13*, 2261.
[11] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.