International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/



(RESEARCH ARTICLE)

Check for updates

EXPLANATION PRESERVING KNOWLEDGE DISTILLATION FOR LIGHTWEIGHT BATTERY STATE OF HEALTH ESTIMATION ON EDGE DEVICES

JongMyoung Kim *

Department of Artificial Intelligence and Big Data, Sehan University.
* Corresponding Author

ORCID Details
JongMyoung Kim: https://orcid.org/0009-0009-4909-0637

International Journal of Science and Research Archive, 2026, 21(02), 162–168

Publication history: Received on 27 September 2026; revised on 05 November 2026; accepted on 07 November 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.21.2.1972

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Deploying battery state of health (SOH) estimators on edge devices requires model compression, but similar predictions do not necessarily imply similar explanations. This paper proposes an explanation preserving knowledge distillation framework that transfers both regression outputs and input attributions from a temporal convolutional teacher to a smaller student. The framework separates supervised estimation, prediction distillation, and signed attribution alignment, followed by integer quantization and a second explanation audit. A reproducible validation protocol is specified for public lithium ion battery aging data, with cell separated testing, chronological evaluation, and training only preprocessing. Evaluation combines SOH error in percentage points, attribution agreement, perturbation based faithfulness, model size, peak memory, and batch one latency on a Raspberry Pi. Analytical examination shows that output matching alone cannot constrain local input sensitivities and that attribution normalization can conceal changes in explanation magnitude. Consequently, the protocol includes signed and magnitude sensitive metrics, a method independent perturbation audit, and explicit checks for nearly zero attribution maps. No model training or hardware measurements are reported. The contribution is a testable method and evaluation design rather than an experimentally established improvement. The framework supports a constrained investigation of whether explanation alignment survives compression and quantization without unacceptable loss of estimation accuracy or edge efficiency.

Keywords: Battery state of health; Knowledge distillation; Explainable artificial intelligence; Edge computing; Model quantization; Attribution consistency

1 INTRODUCTION

Battery health estimation is a practical component of battery management because capacity loss changes usable energy and maintenance decisions. An earlier review outlined an edge artificial intelligence architecture for battery management, including sensing, local processing, and cloud supported updates [1]. The present paper narrows that architectural discussion to one technical problem: preserving the explanatory behavior of a compressed SOH estimator. It does not address remaining useful life prediction, anomaly detection, or closed loop battery control.

Knowledge distillation transfers information from a larger teacher to a smaller student [2]. For SOH regression, the teacher output can supplement capacity based labels. However, the original classification formulation cannot simply be imported without adaptation: a scalar regression output does not provide a class probability distribution. Output distillation is therefore formulated here as squared prediction disagreement. A student may reproduce the teacher on observed cycles while responding differently to voltage, current, or temperature changes outside those observations.

Attention transfer provides a precedent for matching internal or gradient based representations [3]. Such transfer does not establish that explanations are physically correct. Input attribution methods, including integrated gradients, make the selected reference input and attribution procedure explicit [4]. Sanity checks further demonstrate the need to test whether explanation maps depend meaningfully on learned model parameters [5]. These considerations motivate separate assessments of agreement, faithfulness, and stability.

The research question is whether signed attribution alignment can preserve teacher explanation behavior after compression and integer quantization while retaining acceptable SOH accuracy and edge efficiency. Three hypotheses are specified for later testing. First, explanation aligned distillation will improve attribution agreement relative to output only distillation. Second, any agreement gain will persist under an independent perturbation audit. Third, the quantized student will retain the gain under batch one deployment. None of these hypotheses is treated as established.

The contribution is a mathematically specified learning objective and a leakage aware evaluation protocol. The work extends the previous design framework through an operational research question, rather than reproducing its general discussion. A claim of priority or superiority would require a broader systematic comparison and completed experiments.

2 MATERIAL AND METHODS

2.1 Data sources and scope

The primary planned source is the NASA Battery Data Set attributed to Saha and Goebel [6]. The exact archive, cell identifiers, operating conditions, usable cycles, and exclusion reasons must be recorded before training. The design does not assume that every cell has identical charging segments or temperature coverage. Files must be checked against the source documentation, and checksums must accompany the eventual release manifest. The NASA turbofan simulation dataset is outside this study because its target system is not a battery.

The CALCE archive is a candidate external replication source [7]. It is not pooled with NASA by default. A cross source experiment will be included only if comparable capacity labels and a shared observation window can be established. If CALCE data are used, the archive's required dataset and associated experimental publications must be added to the final references. An experiment using only NASA must be described as single source validation.

2.2 Target and observable information

For cell c and reference capacity measurement k, the fractional target is $y(c,k) = Q(c,k) / Qref(c)$. $Qref(c)$ is the documented nominal capacity when reliably available; otherwise it is a fixed initial reference measurement acquired before the evaluated prediction period. The normalization choice must be identical for all model variants. A complete held out lifetime must never be used to determine $Qref(c)$.

The model receives voltage, current, temperature, and elapsed time from a fixed prefix of a charging segment available before the corresponding capacity check. The observation duration is selected using training and validation data, then frozen. A cycle is excluded if that prefix is unavailable; exclusion counts are reported by cell. Completed discharge capacity, future cycles, and retrospectively selected end of life indicators are excluded from the inputs. The timestamp relation between the prefix and the reference capacity measurement is retained in the sample manifest.

2.3 Preprocessing and partitioning

The primary protocol uses an outer held out cell and an inner validation cell drawn from the remaining cells. All samples belonging to either cell remain isolated from training. Missing value rules, channel means, scale factors, interpolation settings, and clipping bounds are fitted on training cells only. The fixed observation prefix is resampled to 128 points; interpolation uses only observations within that prefix. Long missing segments are excluded using a threshold fixed before testing. Test samples are processed with frozen transforms.

2.4 Chronological and external evaluation

A separate chronological protocol evaluates later cycles after fitting on earlier cycles, with a purge interval preventing overlap of any multi cycle input history. This protocol measures within cell progression and is reported separately from unseen cell performance. Results from the two settings must not be merged into a single generalization claim. If too few cells support independent inner validation, hyperparameters are fixed from a documented development subset, and the resulting limitation is stated explicitly.

2.5 Teacher and student estimators

The proposed input tensor has shape 128 by 4. The teacher contains three residual temporal convolutional blocks with 64 channels, kernel size 3, and dilations 1, 2, and 4. Each block contains two causal convolutions with smooth softplus activations. Global average pooling and a 32 unit dense layer precede a scalar linear output. The student uses two blocks with 16 channels, dilations 1 and 2, and a scalar output after global average pooling. Projection layers match residual dimensions. Exact padding, dropout, bias, and parameter counts must be exported with the implementation.

The teacher is trained on measured SOH labels and frozen before student training. All student comparisons use the same architecture, input processing, maximum epochs, and seed set. The initial optimization plan uses Adam, learning rate 0.001, batch size 32, a maximum of 200 epochs, and early stopping after 20 epochs without validation improvement. These are planned settings, not records of an executed run. Final settings and software versions must be reported after implementation.

2.6 Regression and prediction distillation

Let fT(x) and fS(x) denote the frozen teacher and trainable student outputs in fractional SOH units. The supervised and distillation terms are defined over a training batch of B examples as follows:

$$L_{sup} = (1/B) \sum_i (f_S(x_i) - y_i)^2 \quad (1)$$

$$L_{pred} = (1/B) \sum_i (f_S(x_i) - \text{stopgrad}(f_T(x_i)))^2 \quad (2)$$

No softmax temperature is assigned to the scalar output. The supervised term remains necessary because an inaccurate teacher can transfer its errors. Teacher performance is therefore evaluated on the same held out cells as student performance, and the student trained without distillation remains a required baseline.

2.7 Signed attribution alignment

For differentiable floating point models, a(x) is the flattened integrated gradients vector relative to a shared reference x0 [4]. The primary reference is the pointwise training set mean prefix in the standardized input space. The same fixed reference, feature order, and quadrature rule are used for teacher and student. The reference need not represent a physically realizable charging cycle; this limitation is examined with alternative references chosen entirely from training data.

$$a_j(x) = (x_j - x0_j) \int_0^1 [df(x0 + t(x - x0))/dx_j] \, dt \quad (3)$$

Attributions are approximated initially with 32 midpoint integration steps. A 64 step audit on a fixed validation subset checks numerical sensitivity. The attribution target is computed with respect to the scalar fractional SOH output. Define u(a) = a / (norm_2(a) + epsilon), with epsilon = 0.000001. The alignment objective uses signed vectors, not absolute values:

$$L_{attr} = \text{mean}_i \, \text{norm}_2(u(aS_i) - u(\text{stopgrad}(aT_i)))^2 \quad (4)$$

$$L_{total} = L_{sup} + \alpha L_{pred} + \beta L_{attr} \quad (5)$$

The teacher attribution is detached, but differentiation through the student's attribution remains enabled. This requires higher order automatic differentiation and increases training cost. Smooth activations are selected to support this operation. If the chosen implementation cannot differentiate the attribution objective, substituting an attention map changes the method and must be reported as a separate variant.

Alpha is selected from 0.1, 0.5, and 1.0; beta from 0, 0.01, 0.1, and 1.0 using inner validation only. Selection first restricts candidates to those within 0.2 percentage points of the output only student's validation MAE, then selects the highest signed attribution agreement. The tolerance is a proposed experimental decision rule, not a clinically or industrially established acceptance threshold. Failure to find a candidate within tolerance is reported rather than hidden by retuning on test cells.

2.8 Integer quantization

The selected student is converted to an INT8 deployment artifact using representative training inputs. Quantization aware training is a contingency if post training conversion violates the validation error tolerance [8]. Calibration never uses held out cells. Mixed precision operators, floating point fallbacks, and unsupported activations must be disclosed. Reduced storage alone does not establish integer only inference or a latency improvement.

3 EVALUATION PROTOCOL

3.1 Required comparisons and accuracy

The comparison set comprises the teacher, a directly supervised student, an output distilled student, an attribution aligned student, and the INT8 versions of each student. An attribution only ablation sets alpha to zero, and a shuffled teacher attribution control tests whether the alignment target contains useful information. Model selection budgets remain equal across applicable variants. The proposed seed set is 11, 23, 37, 51, and 71.

MAE_pp = (100/N) sum_i abs(yhat_i - y_i) (6)

RMSE_pp = 100 sqrt((1/N) sum_i (yhat_i - y_i)^2) (7)

Both errors are expressed in SOH percentage points, not relative percentage error. Metrics are first computed per held out cell and then averaged with equal cell weight. Per cell results, seed variability, and paired model differences are retained. Confidence intervals resample whole cells when enough independent cells exist. A small cell count limits inferential strength; treating correlated cycles as independent replicates is not justified.

3.2 Agreement and faithfulness

Explanation agreement uses signed cosine similarity, rank correlation of attribution magnitudes, top 10 percent overlap, and the ratio of attribution norms. Nearly zero maps are flagged separately, and rank correlations with constant inputs are reported as undefined. Time bin and channel level summaries are both retained. Magnitude agreement is necessary because normalization can make unequal sensitivity scales appear similar.

A method independent occlusion audit removes contiguous groups of time points using replacements sampled from training prefixes with similar operating conditions. For each group, the change in output is compared with the summed attribution for that group. Group sizes, replacement seeds, and conditional matching rules are fixed on validation data. Faithfulness is reported as association between attribution and output change, alongside the perturbation distribution. Such perturbations assess model behavior, not electrochemical causality.

INT8 attribution is evaluated using perturbation based explanations on the actual deployed artifact. A floating point shadow model is not accepted as proof of integer model explanation preservation. Both floating point and INT8 artifacts also receive the same occlusion audit, enabling comparison on a shared explanation mechanism. The gradient based agreement metric and the deployed perturbation metric are reported separately.

3.3 Robustness and sanity checks

Robustness tests add channel specific noise at 0.5, 1, and 2 percent of the training channel standard deviation, and remove contiguous segments covering 5 or 10 percent of the observation prefix. These perturbations are sensitivity tests rather

than claims about a particular sensor specification. The same mask and noise realizations are used across models. Clean and perturbed errors, attribution stability, and exclusion or abstention rates are reported separately.

Parameter randomization and label permutation controls test whether the attribution procedure responds to learned parameters and meaningful labels [5]. Sensitivity to three training derived references is reported for integrated gradients. High teacher student agreement is insufficient if either model fails the faithfulness or randomization audit. Physical interpretation is limited to plausible associations unless supported by an independent mechanistic experiment.

3.4 Edge deployment measurements

The planned target is one Raspberry Pi 5 configuration, with exact RAM, operating system, runtime, CPU governor, thread count, cooling, and model operator support recorded before benchmarking. Timing uses batch size one and a monotonic clock, with at least 100 warm up calls followed by 1,000 measured calls per artifact. Model loading, preprocessing, inference, and explanation generation are timed separately. Median and 95th percentile latency, model file bytes, parameter count, peak resident memory, and any thermal throttling are reported.

All variants use the same runtime configuration. Memory is measured in isolated processes with an idle baseline and a documented peak sampling method. Explanation generation is evaluated as a separate diagnostic workload, because repeated perturbation calls may dominate cost even when a single SOH prediction is fast. No energy saving is inferred from model size or timing; energy requires a separate calibrated measurement that is outside this initial study.

A deployment budget must follow the intended SOH update interval. The protocol does not borrow a protection loop deadline for a health estimation task. Meeting a measured batch one latency budget on a Raspberry Pi establishes performance only for that configuration and workload. It does not establish hard real time behavior, microcontroller suitability, or battery protection certification.

3.5 Reproducibility records

The eventual release should include source archive hashes, cell and cycle manifests, exclusion logs, split files, preprocessing statistics, all selected hyperparameters, training seeds, checkpoints, conversion logs, artifact hashes, raw per call timing, and explanation settings. Training and edge timing hardware must be distinguished. These materials are prospective requirements; no completed code or data deposit is claimed here.

4 RESULTS AND DISCUSSION

4.1 Analytical result on prediction agreement

Output agreement on sampled inputs does not imply gradient agreement. In one dimension, consider $fT(x) = x$ and $fS(x) = x + a \sin(\text{omega } x)$. At observations $x = n \pi / \text{omega}$, the two outputs agree exactly, while $fS'(x) = 1 + a \cdot \text{omega} \cos(\text{omega } x)$ may differ substantially from $fT'(x) = 1$. This counterexample establishes the logical insufficiency of sampled output matching; it is not a battery experiment or a claim about trained networks.

The example also clarifies the role of a separate explanation objective. Prediction fidelity and sensitivity fidelity constrain different properties. Integrated gradients has its own limitations: in this one dimensional example, matching endpoint outputs and a common baseline can yield matching integrated attribution even when local derivatives differ. The proposed audit therefore includes perturbation responses rather than treating any one explanation method as a complete description of model behavior.

4.2 Analytical result on normalized attribution

If $aS = k\, aT$ for positive k, normalized vectors are approximately equal whenever their norms are well above epsilon. Thus, the normalized alignment loss can be small while attribution magnitude differs. Signed cosine agreement has the same limitation. Reporting the norm ratio and signed perturbation effects prevents a scale change from being concealed by direction only metrics. Near zero teacher maps are a separate degeneracy that must be counted explicitly.

4.3 Interpretation of prospective empirical outcomes

No dataset processing, model training, conversion, or Raspberry Pi benchmarking has been performed for this manuscript. Accordingly, no empirical result table, accuracy claim, speedup, or significance test is presented. The planned comparisons can support explanation preservation only if the aligned student improves agreement over output distillation, remains faithful under independent perturbations, and retains the benefit after conversion within the declared error tolerance.

Alternative outcomes are informative. Improved agreement with reduced faithfulness would indicate imitation of a potentially misleading teacher explanation. Better floating point agreement that disappears in INT8 would identify quantization sensitivity. Reduced file size without lower latency would suggest runtime or operator overhead. Failure under cell separation despite strong chronological performance would indicate limited transfer across cells. These interpretations should guide reporting without changing the hypotheses after inspecting test results.

4.4 Limitations and research boundary

The principal limitation is the absence of empirical validation. The method cannot currently support a claim of improved estimation, trusted explanation, or deployment readiness. Public aging datasets also offer limited cell diversity and may differ from field charging, thermal conditions, and sensor behavior. Even successful held out cell testing would not establish fleet level reliability across chemistries or pack architectures.

Teacher errors and biases may be transferred along with useful sensitivities. Integrated gradients depends on reference selection and numerical integration, while perturbation explanations depend on replacement rules. The shared audits reduce but do not eliminate these dependencies. Agreement should be interpreted as behavioral consistency under specified tests, not as proof of a true causal degradation mechanism.

A completed study should compare recent battery specific compression and explanation preservation methods before asserting novelty. The selected foundational references motivate the formulation but do not constitute a systematic literature review. The single edge platform boundary is intentional: extending to multiple hardware families, remaining useful life estimation, or closed loop safety requires additional experiments.

5 CONCLUSION

This paper defines a framework for preserving the explanatory behavior of a lightweight battery SOH estimator through regression distillation, signed attribution alignment, and post conversion auditing. Its analytical contribution is to identify why sampled output agreement and normalized attribution similarity are individually insufficient evidence of preserved model behavior. The proposed evaluation combines cell separated accuracy, explanation agreement, perturbation faithfulness, and measured edge resource use. Whether the method improves the accuracy explanation efficiency tradeoff remains an empirical question. Completed experiments and confirmed author declarations are required before presenting the manuscript as a validated research article.

REFERENCES

[1] Kim J. Edge AI-based real-time battery management systems: Technology trends and design framework. International Journal of Science and Research Archive. 2026;21(1):644-651. doi:10.30574/ijsra.2026.21.1.1871.

[2] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv [Preprint]. 2015. arXiv:1503.02531. Available from: https://arxiv.org/abs/1503.02531

[3] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations; 2017. Available from: https://arxiv.org/abs/1612.03928

[4] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research. 2017;70:3319-3328. Available from: https://proceedings.mlr.press/v70/sundararajan17a.html

[5] Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems 31; 2018. Available from: https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[6] Saha B, Goebel K. Battery Data Set [Internet]. Moffett Field (CA): NASA Ames Research Center; 2007 [cited 2026 Nov 2]. Available from: https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/

[7] Center for Advanced Life Cycle Engineering. CALCE battery data archive [Internet]. College Park (MD): University of Maryland; [cited 2026 Nov 2]. Available from: https://web.calce.umd.edu/batteries/data/

[8] Jacob B, Kligys S, Chen B, Zhu M, Tang M, Howard A, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 2704-2713. doi:10.1109/CVPR.2018.00286.