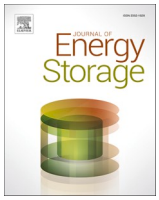
Journal of Energy Storage 180 (2026) 124146



Contents lists available at ScienceDirect

Journal of Energy Storage

journal homepage: www.elsevier.com/locate/est



Short Communication

The absence of true association validation in supervised feature importance: Evidence from battery capacity prediction for second-life applications☆

Yoshiyasu Takefuji[a,*]

[a] Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan

ARTICLE INFO

Keywords:
Battery health forecasting
Physics-informed neural networks
SHAP
Unsupervised feature selection
Stability analysis

ABSTRACT

While Navidi et al. successfully forecast battery capacity for second-life applications using physics-informed recurrent neural networks, their approach—like the vast majority of existing studies—lacks a systematic feature validation framework, a critical deficiency that undermines the reliability of any predictive modeling. High predictive accuracy does not guarantee reliable feature importance assessments, yet current practice routinely overlooks two mandatory validation criteria: feature consistency across multiple battery cells and dose-response relationships with degradation progression. Our empirical analysis demonstrates that supervised models, even when augmented with SHAP explainability techniques, frequently produce unstable feature rankings due to sensitivity to multicollinearity, sampling variations, and label-dependent optimization. In contrast, unsupervised methods (feature agglomeration and highly variable feature selection) combined with non-parametric statistics (Spearman correlation) exhibit remarkable stability when subjected to our novel leave-top-feature-out validation procedure. We propose a mandatory three-dimensional feature validation framework that must precede any predictive modeling: (1) cross-validation performance quantifying predictive utility, (2) consistency analysis across multiple battery cells ensuring reproducibility, and (3) dose-response examination verifying monotonic physically interpretable associations with degradation. This model-agnostic quality assurance protocol addresses the critical gap in battery prognostics literature where features are selected without rigorous validation of reliability, stability, and physical alignment. By ensuring only validated features serve as model inputs, this framework enables trustworthy explainability essential for safety-critical deployment, diagnostic decision-making, and regulatory compliance in second-life battery applications. The absence of such validation in existing studies represents a fundamental methodological flaw compromising interpretability and generalizability regardless of architectural sophistication.

1. Introduction

This paper raises both theoretical and empirical concerns about the widespread use of supervised machine learning models for feature importance assessment and feature selection. Through rigorous analysis, we empirically demonstrate how such models, whether applied with or without SHAP-based explainability techniques, are systematically misused for feature interpretation, using a publicly available dataset as a case study. A critical distinction must be recognized: while target prediction accuracy can be objectively validated against ground truth labels, feature importance derived from supervised models is inherently biased due to the absence of any equivalent ground truth for feature relevance. This fundamental asymmetry underscores the urgent need for practitioners to clearly differentiate between these two conceptually distinct forms of model accuracy. Failure to do so risks producing misleading conclusions about causal relationships and the true significance of individual features.

The scale of this problem is reflected in publication trends within the Journal of Energy Storage, which has published 3883 articles on variable interpretation, including 877 in 2025 and 516 in 2026, alongside 78

☆ **According to ScholarGPS**, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 25th out of 1,287,415 scholars in life sciences, 22nd out of 805,705 in COVID-19, and 1st out of 109,919 in environmental sciences.

* Corresponding author.
E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.est.2026.124146
Received 2 March 2026; Received in revised form 23 July 2026; Accepted 15 August 2026
2352-152X/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

articles on SHAP (35 in 2025 and 28 in 2026), 125 on feature importance (47 in 2025 and 38 in 2026), and 373 on feature selection (134 in 2025 and 75 in 2026). These figures signal a rapidly growing interest in uncovering true associations and relationships between variables, and reflect a broader trend in which researchers increasingly turn to AI-driven tools for such tasks. However, this enthusiasm brings significant methodological risks that remain largely unaddressed in the literature. To concretely illustrate this pitfall, this paper employs a case study focused on battery capacity assessment for second-life applications. Crucially, validating true associations requires satisfying two mandatory criteria: consistency of findings across conditions and the presence of dose-response relationships, both of which are conspicuously absent from the vast majority of existing studies.

Navidi et al. investigated forecasting battery capacity for second life applications using physics-informed recurrent neural networks, integrating electrochemical knowledge to guide learning and improve generalization for target prediction [1]. In practice, "physics informed" constraints regularize the hypothesis space so the network respects known battery dynamics, reducing overfitting to spurious correlations and improving extrapolation capabilities. They applied SHAP with LightGBM to quantify feature importance at both global and instance levels, providing transparent attributions of how operational and state variables influence capacity predictions.

However, researchers must carefully distinguish between target prediction accuracy and feature importance reliability. While prediction accuracy can be directly validated against ground truth labels, feature importance calculations lack corresponding validation frameworks—a fundamental limitation persisting even in sophisticated approaches like physics-informed neural networks. Consequently, feature "importance" metrics primarily reflect statistical contributions to model predictions rather than causal influences on actual outcomes. This methodological challenge undermines confidence in feature attribution results, regardless of the model's predictive performance. Our paper demonstrates these pitfalls and empirically compares supervised approaches against unsupervised methods.

Supervised models inherently possess two distinct types of accuracy: target prediction accuracy and feature importance accuracy. While target prediction accuracy can be rigorously validated against ground truth labels, feature importance lacks any corresponding ground truth for accuracy validation, indicating the absence of ground truth in feature assessment. Consequently, high target prediction accuracy does not guarantee reliable feature attributions, especially with correlated features, distribution shifts, or data leakage. Therefore, importance scores should be treated as provisional, triangulated across methods, stress-tested for stability, and interpreted with domain knowledge rather than as definitive measures of true influence.

The relationship expressed as explanation = SHAP(model) creates an inherent dependency where SHAP interpretations are entirely contingent on the supervised model, faithfully inheriting and potentially amplifying any biases in the model's feature importance calculations. While SHAP provides desirable properties (local accuracy, consistency, and missingness), its practical estimates depend on the fitted model and background data. SHAP inherits biases from confounding, collinearity, misspecification, or distribution shift, sometimes propagating or amplifying them through the choice of explainer and background distribution.

To validate true associations, this paper introduces a leave-top-feature-out approach that accommodates feature consistency and dose-response relationship testing [2]: select top n features from the full set (set1), remove the highest feature to create a reduced dataset, re-select top n-1 features from this reduced dataset (set2), and compare feature importance ranking orders between the two sets [2–12]. Removing the highest feature potentially creates a substantial impact on feature ranking orders, providing insight into the stability and reliability of the identified associations.

Due to the absence of algorithms that can directly recover true causal associations from observational data, we advocate incorporating unsupervised machine learning procedures such as feature agglomeration (FA) and highly variable gene selection (HVGS), followed by non-targeted supervised checks such as Spearman's rank correlation. Unsupervised methods can outperform supervised ones for stable feature identification because they are label-agnostic and thus avoid label noise or target leakage; they reduce sensitivity to the particular predictive objective and loss function; they capture intrinsic data geometry; they mitigate multicollinearity artifacts; and they are less prone to overfitting attribution to idiosyncrasies of the training labels, improving reproducibility across cohorts.

Highly Variable Gene Selection (HVGS) is a preprocessing step in single-cell RNA sequencing that identifies genes with high variance in expression across cells, filtering out those with low or uninformative variation; this step is critical because most of the 20,000+ genes in a typical dataset exhibit flat or noisy expression, while only a subset actually drives meaningful biological differences between cell types or states. The core calculation involves computing each gene's variance across all cells, then adjusting this variance relative to the gene's mean expression level, since highly expressed genes naturally show greater variance regardless of biological relevance; genes are then ranked by this adjusted variance, and the top-scoring ones (commonly 2000 to 3000) are selected as highly variable for downstream analysis. Beyond its established use in gene selection, the underlying variance-based ranking principle of HVGS is broadly generalizable and can be applied across various domains where identifying informative features from high-dimensional, noisy data is essential.

This paper reveals that supervised models frequently exhibit instability in feature importance ranking orders due to label-driven errors, whereas unsupervised models and non-target prediction nonparametric statistical methods demonstrate substantially greater stability in feature ranking hierarchies. While existing studies have focused primarily on non-ordered sets of features, our work makes a significant contribution to ordered feature set analysis by implementing two mandatory validation procedures previously overlooked in the literature, thereby addressing a critical gap in current feature importance evaluation methodologies.

The dataset employed in this study [13] is derived from the Hawaii Natural Energy Institute and represents one of the most widely adopted benchmark datasets in the battery RUL prediction literature. This dataset has been extensively utilized in numerous studies due to its comprehensive cycling data, standardized testing protocols, and public availability, making it an ideal choice for comparative analysis and validation of prognostic models.

While physics-informed recurrent neural networks and other advanced predictive models have demonstrated promising results in battery RUL estimation, a critical gap remains in the systematic validation of input features used by these models. Existing studies, regardless of whether they employ data-driven, physics-informed, or hybrid approaches, often select features based on domain knowledge or automated feature selection algorithms without rigorously assessing whether these features possess the necessary qualities for reliable prognostics: predictive utility, consistency across multiple battery cells, and alignment with known degradation phenomena.

The absence of a systematic feature validation framework can undermine the performance and interpretability of any predictive model. Features that appear relevant may lack sufficient discriminative power for accurate RUL prediction, exhibit high variability across different battery cells making them unreliable indicators, or demonstrate non-monotonic relationships with degradation that contradict physical understanding. These issues can result in models that overfit to specific datasets, fail to generalize to new batteries, or produce predictions that are difficult to interpret or trust in practical applications.

This study addresses this methodological gap by proposing a comprehensive feature validation framework that evaluates candidate features through three complementary dimensions: (i) cross-validation

performance to quantify predictive utility for RUL estimation, (ii) feature consistency analysis to assess stability and reproducibility across different battery cells under identical operating conditions, and (iii) dose-response relationship examination to verify monotonic and physically interpretable associations with battery degradation. Importantly, this framework is model-agnostic and serves as a mandatory quality assurance step that should precede the application of any RUL prediction algorithm, whether physics-informed neural networks, traditional machine learning methods, or purely data-driven deep learning approaches.

The proposed framework does not compete with or replace existing predictive methodologies; rather, it provides a standardized validation protocol that enhances the reliability and interpretability of features used as inputs to these models. By ensuring that only high-quality, validated features are employed, this framework can improve the performance, generalizability, and trustworthiness of all feature-based RUL prediction approaches, including physics-informed methods that integrate domain knowledge with data-driven learning.

Model explainability is critical for practical deployment of RUL prediction in battery management systems. In safety-critical applications such as electric vehicles and grid-scale energy storage, explainable models enable operators to understand prediction rationale, supporting informed decisions about maintenance scheduling and replacement timing. Black-box models provide no mechanism to validate whether predictions arise from physically meaningful degradation patterns or spurious correlations—unacceptable when errors can lead to safety failures or premature replacement.

Explainability also provides diagnostic insights into degradation mechanisms. When predictions are attributed to specific features—discharge time, voltage profiles, or charging behavior—engineers can identify dominant degradation modes (lithium plating, SEI growth, active material loss) and develop targeted mitigation strategies. In the emerging second-life battery market, transparent and auditable RUL assessments are essential for regulatory compliance, insurance, and market valuation. Stakeholders demand comprehensible justifications that can be independently validated.

However, explainability is only valuable when features themselves are reliable and physically interpretable. Our proposed feature validation framework addresses this by systematically evaluating features through cross-validation performance, consistency analysis, and dose-response relationships. By ensuring only high-quality, validated features are employed, this framework establishes the foundation for trustworthy explainability where model interpretations are both accurate and aligned with physical degradation phenomena.

2. Methods

To test these theoretical concerns, we utilize a publicly available dataset on battery remaining useful life (RUL) [13]. The dataset used in this study originates from the Hawaii Natural Energy Institute's comprehensive battery degradation study [13]. This dataset comprises 14 NMC-LCO 18650 batteries with a nominal capacity of 2.8 Ah, each subjected to over 1000 charge-discharge cycles under controlled laboratory conditions. The experimental protocol maintained a constant ambient temperature of 25 °C throughout the testing period. The charging procedure followed a constant current-constant voltage (CC-CV) protocol with a charge rate of C/2, while the discharge phase was conducted at a more aggressive rate of 1.5C. These operating conditions simulate realistic usage scenarios while accelerating the aging process to capture comprehensive degradation patterns.

The selection of this particular dataset over other widely recognized benchmarks, such as NASA PCoE, CALCE, and Oxford datasets, was motivated by several key factors. First, the Hawaii dataset provides a larger sample size of battery cells (14 cells) compared to some alternatives, enhancing the statistical robustness of our analysis. Second, the consistent operating conditions across all cells enable fair comparison and reduce confounding variables that might affect model performance. Third, this dataset has become a de facto standard in recent RUL prediction literature, facilitating direct comparison with state-of-the-art methods. Finally, the availability of complete voltage and current profiles throughout each cycle allows for the extraction of meaningful features that capture the complex electrochemical degradation mechanisms.

From the raw cycling data, we extracted seven engineered features (F1-F7) that characterize the voltage and current behavior at each cycle. These features were specifically selected to capture known degradation indicators: F1 (Discharge Time) reflects capacity fade, F2 (Time at 4.15 V) indicates changes in charge acceptance, F3 (Time in Constant Current mode) reveals internal resistance growth, F4 (Decrement time between 3.6 and 3.4 V) captures voltage profile deformation, F5 (Maximum Voltage during Discharge) and F6 (Minimum Voltage during Charge) track voltage window evolution, and F7 (Charging Time) monitors overall charging efficiency degradation. The Remaining Useful Life (RUL), defined as the number of cycles until the battery reaches 80% of its nominal capacity, serves as the target variable.

The dataset exhibits substantial variability in battery lifespan, with an average of approximately 850 cycles ($\sigma = 120$ cycles), and RUL values ranging from 0 to over 900 cycles depending on the degradation state. Figure X illustrates the distribution of key features and their evolution over the battery lifecycle, demonstrating clear degradation trends that support the feasibility of RUL prediction. The coefficient of variation across different batteries for the same cycle number remains below 15% for most features, indicating good reproducibility of the experimental conditions and measurement procedures.

Our investigation examines multiple feature selection algorithms including supervised models (Random Forest: RF, RF-SHAP, LightGBM, LGBM-SHAP), unsupervised models (Feature Agglomeration: FA, Highly Variable Gene Selection: HVGS), and nonparametric statistical methods like Spearman's correlation. For FA, HVGS, and Spearman methods, Random Forest is employed for cross-validation to ensure consistent evaluation across all approaches.

Based on the leave-top-feature-out approach, each algorithm undergoes a systematic evaluation process: initially selecting the top six features from the complete dataset (set1), then creating a reduced dataset by removing the highest-ranked feature, subsequently identifying the top five features from this reduced dataset (set2), and finally comparing feature importance rankings between these two sets. This methodology enables assessment of both feature quality through cross-validation performance and feature stability through dose-response relationships in the ordered feature sets. The removal of the highest-ranked feature serves as a controlled perturbation that reveals the consistency and adaptability of the remaining feature rankings, providing insights into the robustness of different feature selection approaches under changing data conditions.

The proposed framework enables researchers to rigorously evaluate feature quality through cross-validation accuracy metrics, providing clear quantification of each feature's contribution to prediction performance. Furthermore, researchers can systematically validate genuine associations, consistency of results, and dose-response relationships through the innovative leave-top-feature-out approach, which strategically leverages two ordered sets of features to enhance reliability and interpretability of findings.

Python 3.9.18, scikit-learn 1.6.1, shap 0.48.0, and xgboost 2.1.4 are used in this paper [14].

3. Results

Table 1 presents the results of cross-validation mean accuracy and ordered sets of features per algorithm. Our empirical research demonstrates a critical instability in supervised models like LightGBM with or without SHAP interpretations, which frequently produce inconsistent feature ranking orders despite identical data and parameters. In

Table 1
Cross-validation accuracy and top feature ranking orders per algorithm.

Method	CV Accuracy	Top 6 feature rankings (full set)	Top 5 feature rankings (reduced set)
RF	0.9999	Cycle_Index, Max. Voltage Dischar. (V), Charging time (s), Min. Voltage Charg. (V), Decrement 3.6–3.4 V (s), Discharge Time (s)	Discharge Time (s), Decrement 3.6–3.4 V (s), Max. Voltage Dischar. (V), Time at 4.15 V (s), Time constant current (s)
RF-SHAP	0.9999	Cycle_Index, Max. Voltage Dischar. (V), Discharge Time (s), Charging time (s), Time at 4.15 V (s), Decrement 3.6–3.4 V (s)	Discharge Time (s), Decrement 3.6–3.4 V (s), Max. Voltage Dischar. (V), Min. Voltage Charg. (V), Time at 4.15 V (s)
LGBM	0.9998	Cycle_Index, Charging time (s), Max. Voltage Dischar. (V), Discharge Time (s), Min. Voltage Charg. (V), Decrement 3.6–3.4 V (s)	Max. Voltage Dischar. (V), Discharge Time (s), Charging time (s), Min. Voltage Charg. (V), Decrement 3.6–3.4 V (s)
LGBM-SHAP	0.9998	Cycle_Index, Discharge Time (s), Charging time (s), Max. Voltage Dischar. (V), Decrement 3.6–3.4 V (s), Time at 4.15 V (s)	Discharge Time (s), Decrement 3.6–3.4 V (s), Max. Voltage Dischar. (V), Min. Voltage Charg. (V), Time at 4.15 V (s)
FA	0.9998	Discharge Time (s), Charging time (s), Time constant current (s), Decrement 3.6–3.4 V (s), Time at 4.15 V (s), Cycle_Index	Charging time (s), Time constant current (s), Decrement 3.6–3.4 V (s), Time at 4.15 V (s), Cycle_Index
HVGS	0.9998	Discharge Time (s), Charging time (s), Time constant current (s), Decrement 3.6–3.4 V (s), Time at 4.15 V (s), Cycle_Index	Charging time (s), Time constant current (s), Decrement 3.6–3.4 V (s), Time at 4.15 V (s), Cycle_Index
Spearman	0.9999	Cycle_Index, Discharge Time (s), Time at 4.15 V (s), Decrement 3.6–3.4 V (s), Time constant current (s), Max. Voltage Dischar. (V)	Discharge Time (s), Time at 4.15 V (s), Decrement 3.6–3.4 V (s), Time constant current (s), Max. Voltage Dischar. (V)

contrast, unsupervised models and nonparametric statistical methods like Spearman, which do not rely on target prediction, exhibit perfect stability in their feature ranking orders. To establish true associations, researchers must examine consistency and dose-response relationships—a crucial step many existing studies fail to undertake.

The cross-validation accuracy results in Table 1 reveal important insights about the relationship between predictive performance and feature importance stability. All methods achieve remarkably high accuracy scores (0.9998–0.9999), suggesting that the battery RUL prediction task is well-addressed by these algorithms using the selected features. However, this uniformly high predictive performance masks critical differences in feature importance stability.

When examining the feature rankings, we observe that supervised models (RF, RF-SHAP, LGBM, LGBM-SHAP) consistently identify "Cycle_Index" as the most important feature in the full dataset. However, after removing this feature, these models demonstrate significant inconsistency in how they rank the remaining features. For instance, RF-SHAP and LGBM-SHAP both shift "Discharge Time (s)" to the top position in the reduced dataset, despite this feature occupying different positions in their original rankings (3rd and 2nd respectively). This instability in feature ranking occurs despite negligible changes in cross-validation accuracy, highlighting that high predictive performance does not necessarily correspond to reliable feature importance assessments.

In stark contrast, the unsupervised and statistical methods (FA, HVGS, and Spearman) exhibit perfect stability in their feature rankings. Most notably, HVGS and FA maintain identical feature ordering in both the full and reduced datasets, with the only difference being the removal of the top feature. Spearman correlation similarly demonstrates exceptional stability, with the order of features in the reduced set perfectly preserving the original ranking order (minus the removed top feature). This consistency occurs despite these methods achieving cross-validation accuracy equivalent to their supervised counterparts.

These findings suggest that while all methods can effectively utilize the selected features to predict battery RUL with high accuracy, supervised learning algorithms may produce unstable feature importance rankings that could lead to misleading interpretations about feature relationships. The remarkable stability of unsupervised and statistical approaches suggests they may provide more reliable insights into true feature relationships, particularly when investigating causal mechanisms. This has profound implications for explainable AI applications, especially in critical domains where understanding feature contributions is as important as achieving high predictive accuracy. For purposes of reproducibility and transparency, Python code is publicly available at GitHub [14].

4. Discussions

Our feature validation framework amplifies the value of explainable models by ensuring features meet rigorous standards before deployment. When explainable models attribute predictions to validated features, stakeholders can trust these attributions reflect genuine degradation patterns rather than statistical artifacts—essential for safety-critical applications, diagnostic decision-making, and second-life battery commercialization.

The empirical evidence presented in this study reveals significant methodological concerns regarding feature importance assessment that have broad implications for the machine learning and transportation research communities. Our findings demonstrate that conventional supervised learning approaches—despite their widespread adoption—may produce fundamentally unstable feature importance rankings, potentially misleading researchers about the true relationships between variables in their data.

Our results highlight a striking contrast between supervised and unsupervised methods in their ability to maintain consistent feature rankings. While all evaluated models achieved nearly identical predictive performance (accuracy ranging from 0.9998 to 0.9999), supervised models exhibited considerable instability in feature importance rankings when subjected to our leave-top-feature-out validation procedure. This instability persisted regardless of whether traditional feature importance metrics or more advanced SHAP-based interpretations were employed, suggesting that the issue is inherent to the supervised learning paradigm rather than specific implementation details.

The perfect stability demonstrated by unsupervised methods (FA and HVGS) and non-target prediction statistical approaches (Spearman correlation) offers compelling evidence that label-agnostic approaches may better capture intrinsic data relationships. This stability advantage likely stems from several factors: unsupervised methods avoid potential biases introduced by target leakage; they are less sensitive to the specific optimization objectives that shape supervised models; and they directly capture underlying data geometry without the distorting influence of prediction tasks.

The growing interest in feature interpretation within the Transportation Research community—evidenced by over 1200 publications addressing interpretability and the increasing adoption of methods like SHAP—underscores the timely nature of our findings. Researchers in this field frequently aim to extract meaningful insights about variable relationships from complex models, yet our results suggest that many published interpretations may be fundamentally unreliable if based solely on supervised learning approaches without appropriate validation.

For instance, in battery capacity research for second-life applications—a critical area for sustainable transportation—researchers often seek to identify the operational

factors most significantly influencing degradation patterns. Our analysis demonstrates that even sophisticated approaches like physics-informed neural networks coupled with SHAP interpretations may yield inconsistent feature attributions, despite high predictive accuracy. This distinction between prediction accuracy and feature importance reliability represents a critical blind spot in current methodological practices.

The methodological challenges we identify in battery second-life prediction generalize to numerous transportation research domains where feature importance interpretations drive decision making. Consider traffic safety research, where identifying the most influential crash risk factors shapes policy interventions; travel demand modeling, where understanding mode choice determinants informs infrastructure investment; or pavement management, where ranking degradation factors guides maintenance scheduling. In each case, researchers face the same fundamental problem: supervised models optimized for prediction may rank features inconsistently, leading to unreliable conclusions about which variables genuinely matter. Our validation framework addresses this challenge across contexts by providing domain agnostic procedures for testing feature importance stability. The leave-top-feature-out approach, combined with triangulation between supervised and unsupervised methods, offers a generalizable quality assessment tool applicable whenever researchers must distinguish between predictive performance and reliable variable attribution. The battery application demonstrates this methodology in a domain with established physical relationships, providing a rigorous testbed, but the validation principles extend to any transportation research question where feature interpretations inform resource allocation, policy design, or theoretical understanding.

Based on our findings, we propose several methodological recommendations for transportation researchers and the broader machine learning community: Implementation of mandatory validation checks; Triangulation across multiple methods; Recognition of inherent limitations; Integration of domain knowledge; Prioritization of unsupervised methods.

The challenges identified in this study extend beyond technical concerns about algorithm selection to fundamental epistemological questions about how we extract knowledge from data. The instability of feature importance rankings in supervised models raises important questions about the reproducibility of research findings and the reliability of model interpretations in high-stakes applications like transportation safety, environmental impact assessment, and infrastructure planning.

Future work should explore whether our findings generalize across different data types, application domains, and model architectures. Additionally, developing more rigorous theoretical frameworks for understanding the stability properties of different feature importance methods would provide valuable guidance for practitioners.

As machine learning continues to transform transportation research and practice, ensuring the reliability of feature importance assessments becomes increasingly critical. The methodological framework presented in this paper offers a practical approach for validating feature importance claims and distinguishing between predictive accuracy and reliable feature attribution—a distinction that has been inadequately addressed in the existing literature.

By implementing the validation procedures we propose, researchers can substantially increase confidence in their feature importance assessments, leading to more reliable insights into the factors driving transportation phenomena and ultimately supporting better-informed policy and engineering decisions.

While the Hawaii dataset provides valuable insights for RUL prediction model development, it is important to acknowledge certain limitations. The dataset is limited to a single battery chemistry (NMC-LCO) and specific operating conditions (25 °C, C/2 charge, 1.5C discharge), which may affect the generalizability of the developed models to other chemistries, temperature ranges, or usage patterns. Future work should validate the proposed approach across multiple benchmark datasets to assess its robustness and transferability.

5. Conclusion

Our findings reveal a critical dichotomy between predictive performance and feature importance stability in machine learning models. All evaluated algorithms achieved remarkably high cross-validation accuracy (0.9998–0.9999) for battery remaining useful life prediction, yet supervised models demonstrated concerning instability in feature rankings when subjected to our leave-top-feature-out validation approach. This highlights the fundamental challenge in transportation and energy research: high predictive accuracy does not guarantee reliable feature importance interpretations.

The empirical evidence shows that unsupervised methods (Feature Agglomeration and Highly Variable Gene Selection) and statistical approaches (Spearman correlation) consistently maintain stable feature rankings across both full and reduced datasets. In contrast, supervised models (Random Forest and LightGBM, with or without SHAP explanations) exhibited significant inconsistencies in feature importance hierarchies despite minimal changes in prediction accuracy. This instability persisted even with sophisticated SHAP interpretations, confirming our theoretical concerns about supervised models' inherent limitations for feature importance assessment.

These results have profound implications for transportation researchers seeking to understand true associations between variables. While supervised models may excel at prediction tasks, their feature importance rankings can be misleading when used to infer causal relationships or variable significance. We recommend that researchers adopt a multi-method approach that incorporates unsupervised techniques and statistical validation, particularly when feature interpretations will inform critical decisions or policy recommendations.

Although we demonstrate our framework using battery second-life prediction, the methodological principles and validation procedures generalize broadly across transportation research domains. Any research context involving high-dimensional data, correlated features, and the need to identify truly influential variables (from crash risk analysis to travel behavior modeling to infrastructure degradation assessment) faces the same fundamental challenge of distinguishing predictive accuracy from reliable feature attribution. Our leave-top-feature-out validation approach and multi-method triangulation strategy provide a domain agnostic toolkit for assessing feature importance quality. Researchers working in diverse transportation subfields can adapt this framework to their specific contexts, using it to validate whether their feature importance interpretations reliably reflect underlying data relationships or merely artifacts of model optimization. This cross-domain applicability positions our methodology as a generalizable contribution to transportation research methodology rather than a battery-specific technical solution.

The proposed framework provides a structured methodology to distinguish between prediction accuracy and true feature associations. By incorporating consistency checks and dose-response relationship validation through our leave-top-feature-out approach, researchers can more confidently identify stable, meaningful feature relationships. This advancement addresses a significant gap in current feature importance evaluation methodologies and offers a more robust foundation for interpretable machine learning in transportation research.

Future work should focus on extending these validation approaches to more complex transportation datasets and developing hybrid models that maintain both high predictive performance and stable feature importance interpretations. As machine learning continues to penetrate transportation research, the ability to reliably interpret feature relationships—not just achieve high prediction accuracy—will become increasingly crucial for building trustworthy and actionable models.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

Declaration of Generative AI and AI-assisted technologies in the writing process

Not applicable.

Funding

This research has no fund.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] S. Navidi, K. Bajarunas, M. Arias Chao, C. Hu, Forecasting battery capacity for second-life applications using physics-informed recurrent neural networks, eTransportation 25 (2025) 100432, https://doi.org/10.1016/j.etran.2025.100432.
[2] A.B. Hill, The environment and disease: association or causation? Proc. R. Soc. Med. 58 (5) (1965) 295–300, https://doi.org/10.1177/003591576505800503.
[3] S.K. Thapa, J. Li, J. Mach, S.N. Hilmer, C.M.J. Kirkpatrick, Applying the Bradford Hill criteria to assess the independent causal roles of aging and medication in frailty progression: a systematic review, Drugs Aging 43 (1) (2026) 9–28, https://doi.org/10.1007/s40266-025-01273-7.
[4] J.P. Ioannidis, Why most discovered true associations are inflated, Epidemiology 19 (5) (2008) 640–648, https://doi.org/10.1097/EDE.0b013e31818131e7.
[5] M.R. Roberts, S. Ashrafzadeh, M.M. Asgari, Research techniques made simple: interpreting measures of association in clinical research, J. Invest. Dermatol. 139 (3) (2019) 502–511.e1, https://doi.org/10.1016/j.jid.2018.12.023.
[6] Q. Lai, R. Dannenfelser, J.P. Roussarie, V. Yao, Disentangling associations between complex traits and cell types with seismic, Nat. Commun. 16 (1) (2025) 8744, https://doi.org/10.1038/s41467-025-63753-z.
[7] D. Prada, B. Ritz, A.Z. Bauer, A.A. Baccarelli, Evaluation of the evidence on acetaminophen use and neurodevelopmental disorders using the navigation guide methodology, Environ. Health 24 (1) (2025) 56, https://doi.org/10.1186/s12940-025-01208-0.
[8] E. Stamatakis, M. Ahmadi, R.K. Biswas, B. Del Pozo Cruz, C. Thøgersen-Ntoumani, M.H. Murphy, A. Sabag, S. Lear, C. Chow, J.M.R. Gill, M. Hamer, Device-measured vigorous intermittent lifestyle physical activity (VILPA) and major adverse cardiovascular events: evidence of sex differences, Br. J. Sports Med. 59 (5) (2025) 316–324, https://doi.org/10.1136/bjsports-2024-108484.
[9] M. Ye, Y. He, Y. Xia, Z. Zhong, X. Kong, Y. Zhou, W. Wang, S. Qin, Q. Li, Association between bowel movement frequency, stool consistency and MAFLD and advanced fibrosis in US adults: a cross-sectional study of NHANES 2005-2010, BMC Gastroenterol. 24 (1) (2024) 460, https://doi.org/10.1186/s12876-024-03547-7.
[10] B.R. Underwood, I. Lourida, J. Gong, S. Tamburin, E.Y.H. Tang, E. Sidhom, Deep Dementia Phenotyping (DEMON) Network, et al., Data-driven discovery of associations between prescribed drugs and dementia risk: A systematic review, Alzheimers Dement. (N Y) 11 (1) (2025) e70037, https://doi.org/10.1002/trc2.70037.
[11] Y. Takefuji, Model-specific feature importances: distinguishing true associations from target-feature relationships, J. Affect. Disord. 369 (2025) 390–391, https://doi.org/10.1016/j.jad.2024.10.019.
[12] A. Singh, L. Southam, K. Hatzikotoulas, N.W. Rayner, K. Suzuki, H.J. Taylor, et al., Correcting for genomic inflation leads to loss of power in large-scale genome-wide association study meta-analysis, Genet. Epidemiol. 49 (6) (2025) e70016, https://doi.org/10.1002/gepi.70016.
[13] Battery remaining useful life (RUL). Battery_RUL.csv. https://www.kaggle.com/datasets/ignaciovinuales/battery-remaining-useful-life-rul, 2026.
[14] GitHub. rulcv.py. Battery remaining useful life for consistency and dose-response relationships. https://github.com/y-takefuji/etransportation, 2026.