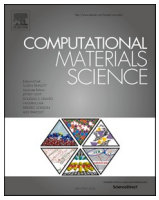
Computational Materials Science 267 (2026) 114609

Contents lists available at ScienceDirect

Computational Materials Science

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



Beyond prediction: Assessing stability in feature selection methods for materials science applications

Yoshiyasu Takefuji *

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

ARTICLE INFO

Keywords:
Machine learning interpretability
Feature importance validation
Model-specific bias
Unsupervised feature selection
Materials informatics

ABSTRACT

This study examines the reliability of feature selection methods in materials science, where machine learning applications have surged despite widespread misapplications stemming from limited understanding of interpretability constraints. We compare supervised models (XGBoost, Random Forest), unsupervised techniques (Feature Agglomeration, HVGS), and statistical methods (Spearman's correlation) through a novel stability testing framework using a public materials dataset. Our results reveal that despite high predictive accuracy ($R^2 > 0.95$), supervised models produce unstable feature rankings when the highest-ranked feature is removed—a critical flaw when identifying structure-property relationships. Common misapplications include over-reliance on black-box models for scientific interpretation, insufficient cross-validation procedures, and failure to test feature importance stability. In contrast, unsupervised methods and Spearman's correlation demonstrate perfect ranking stability while maintaining competitive performance. This highlights a fundamental distinction between prediction accuracy and feature importance reliability. We recommend that materials researchers supplement supervised learning with model-agnostic approaches to avoid misinterpretation of material-property relationships and ensure scientifically robust conclusions about causal mechanisms in materials development.

1. Introduction

With the advent of AI analysis tools, Computational Materials Science has published an extensive collection of articles employing machine learning methodologies: 1046 articles utilizing machine learning (with 228 in 2025 and 9 in 2026), 153 articles focusing on feature selection (29 in 2025), 76 articles implementing XGBoost (24 in 2025), and 50 articles incorporating SHAP analysis (19 in 2025) as of October 20, 2025. This publication trend clearly demonstrates the materials science community's growing interest in AI applications. However, despite this enthusiasm, insufficient understanding of machine learning fundamentals has led to prevalent misapplications in research. This paper systematically identifies common pitfalls in supervised machine learning for feature selection and rigorously evaluates the comparative effectiveness of supervised machine learning models, unsupervised approaches, and non-target prediction methods through comprehensive cross-validation using a publicly available dataset. The findings underscore that feature selection plays a crucial role in material data-driven analysis, particularly for discovering new composite alloys with diverse properties that meet specialized industrial requirements.

A critical challenge in this domain is that many researchers lack familiarity with the reliability constraints of supervised models. Supervised algorithms such as eXtreme Gradient Boosting (XGBoost) and Random Forest exhibit two distinct types of accuracy that are often conflated: target prediction accuracy and feature importance accuracy. While target prediction accuracy can be systematically validated against ground truth label values, feature importances lack corresponding ground truth references for accuracy validation. Consequently, different models generate distinct feature importance rankings, leading to model-specific interpretations that may prove erroneous in practice. This distinction is particularly important because feature importance in supervised models reflects contributions to prediction performance rather than true causal or correlational associations with the target variable. Even models with high target prediction accuracy can produce unreliable feature importance rankings due to the absence of objective validation metrics, as these importance scores merely quantify contributions to the prediction mechanism rather than intrinsic relationships.

The practical implications of these limitations can be observed in recent research. For example, Hou et al. conducted a pioneering investigation into artificial intelligence applications for microstructure

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

https://doi.org/10.1016/j.commatsci.2026.114609
Received 23 November 2025; Received in revised form 8 February 2026; Accepted 21 February 2026
Available online 25 February 2026

prediction in aluminum alloy castings [1]. Their methodical evaluation of seven distinct AI algorithms revealed that XGBoost demonstrated superior performance in accurately predicting microstructural characteristics. To address the interpretability challenge, the researchers implemented SHapley Additive exPlanations (SHAP) analysis to illuminate the complex relationships between specific alloy compositions, processing parameters, and resulting microstructural features. While this analytical framework attempted to bridge the gap between sophisticated machine learning techniques and fundamental physical metallurgical principles, our current work demonstrates that such approaches may still be subject to the inherent limitations of supervised learning for feature selection, potentially leading to misinterpretation of the actual driving factors in microstructure formation mechanisms.

Despite Hou et al.s remarkable achievement in prediction accuracy, their approach raises fundamental concerns regarding the reliability of XGBoost with SHAP interpretation due to the model-specific nature of supervised machine learning algorithms. While supervised learning models like XGBoost benefit from ground truth values for validating target prediction accuracy, the feature importance rankings they generate lack equivalent ground truth metrics for validation. Hou et al.s own research demonstrates that different models produced substantially different feature importance hierarchies—a critical inconsistency stemming from the absence of objective validation mechanisms.

A significant knowledge gap exists among materials researchers, including Hou et al., regarding three critical methodological pitfalls: the violation of fundamental assumptions underlying data analysis tools, the inherent challenges in validating model interpretations against ground truth, and preprocessing artifacts such as normalization and transformation techniques that can generate misleading results. An extensive body of literature—comprising over 300 peer-reviewed articles—has systematically documented fundamental limitations in feature importance metrics derived from all contemporary supervised machine learning models, including XGBoost [2–7]. The prevailing misconception in the field is that improving predictive accuracy will necessarily enhance the interpretation of variable relationships, when in fact, prediction accuracy and feature importance reliability represent distinct and often orthogonal challenges in machine learning applications [8–12]. This distinction is particularly critical in materials science applications where causal understanding, not merely predictive power, is essential for advancing fundamental materials knowledge.

The functional relationship expressed as explain = SHAP(model) highlights that SHAP interpretations inherently inherit and potentially amplify biases present in the underlying models feature importance calculations [13–18]. Although SHAP has gained widespread adoption as an interpretability tool, its explanations remain fundamentally constrained by model-specific biases and assumptions. Consequently, feature importance metrics primarily reflect contributions to prediction outcomes rather than representing true causal relationships between variables, meaning that high predictive accuracy does not necessarily translate to reliable feature importance rankings.

In the absence of accurately calculating true associations between variables, this paper advocates for the use of multifaceted approaches using unsupervised machine learning models that avoid many of the interpretational pitfalls of supervised methods. Specifically, feature agglomeration techniques—which hierarchically cluster related features based on their intrinsic similarity rather than their predictive power—can reveal natural groupings of materials parameters that interact in physically meaningful ways. This approach identifies correlated feature clusters without imposing model-specific biases about their relationship to target variables. Similarly, highly variable gene selection methods, originally developed for bioinformatics applications, can be adapted to materials science to identify the most informative compositional and processing parameters through variance-based filtering rather than model-dependent importance metrics. These techniques detect features with significant information content across the dataset independent of their relationship to any specific target variable. When these unsupervised approaches are combined with nonlinear nonparametric statistical methods such as Spearman's correlation, researchers can establish robust, model-agnostic associations between materials parameters and properties that better reflect underlying physical relationships rather than algorithmic artifacts. This comprehensive analytical framework provides a more rigorous foundation for materials knowledge discovery than relying solely on model-specific interpretability tools like SHAP.

It is crucial that researchers leverage domain knowledge when applying machine learning to materials science. Three fundamental contradictions must be addressed: between high-dimensional feature spaces and limited sample sizes, between model accuracy and practical usability, and between algorithmic learning results and established domain knowledge [21]. In this work, we mitigate these challenges by embedding materials science principles throughout our modeling pipeline. Our feature selection process incorporates physicochemical understanding of material properties, while we validate model predictions against known materials behavior patterns to ensure physical plausibility. Furthermore, we evaluate consistency and dose-response relationships between input features and predicted outcomes, providing mechanistic interpretability that aligns with materials science theory. This domain-knowledge-embedded approach follows recommendations from recent advances in materials informatics that stress the importance of domain expertise throughout the machine learning modeling process.

Data quality and quantity critically shape the reliability and generalizability of machine learning outcomes in materials research, and governance frameworks embedded with domain knowledge offer lifecycle strategies for evaluating and improving datasets to support high-quality, appropriate-quantity data acquisition and model deployment. Building on these principles, our study emphasizes true association assessment—an aspect often overlooked—by explicitly examining consistency and dose-response relationships. Specifically, we propose a leave-top1-out procedure that removes the highest-impact feature and reassesses feature ranking orders and model performance to identify potential dominance-driven artifacts, strengthen robustness against spurious correlations, and ensure that learned relationships persist under perturbations of the feature space. This approach complements data governance efforts by providing association-level validation that aligns machine learning outputs with materials domain knowledge and enhances the credibility of the resulting insights.

2. Methods

Due to the unavailability of the original datasets from Hou et al., this study employs a comprehensive public dataset of aluminum alloy with 1154 instances and 31 features [19] to systematically evaluate feature selection methodologies. The dataset contains information on the composition and processing conditions of aluminum alloys. The mechanical properties included are yield strength, tensile strength, and elongation. Additionally, the dataset provides information about the class to which each alloy belongs. Our investigation encompasses multiple analytical approaches: supervised models including XGBoost and Random Forest; unsupervised techniques such as Feature Agglomeration (FA) and Highly Variable Gene Selection (HVGS); and non-target prediction nonlinear nonparametric methods exemplified by Spearman's correlation with corresponding p-values. This diverse methodological spectrum enables robust comparative analysis across fundamentally different feature selection paradigms.

To rigorously assess true feature associations and algorithmic reliability, we implement a novel consistency-stability testing framework. First, we identify the top five features from the complete feature set and cross-validate their predictive accuracy. Subsequently, we eliminate the highest-ranked feature from the full dataset to create a strategically reduced dataset. We then re-select the top four features from this reduced set and meticulously compare the feature importance ranking orders between the original and reduced datasets. This methodical

approach provides critical insights into ranking stability, a key indicator of reliable feature selection that is often overlooked in materials science research employing machine learning techniques.

3. Results

For purposes of reproducibility and transparency, Python code, alloy.py is publicly available at GitHub [20]. As shown in Table 1, the cross-validation results reveal striking differences in both predictive performance and feature ranking stability across different algorithm categories. Random Forest achieved the highest initial 5-fold cross-validation R^2 score (0.9801) when using the top five features, with "Yield Strength (MPa)" identified as the most important feature. However, when this top feature was removed, the R^2 score decreased to 0.9191, and notably, the feature ranking order changed dramatically, with Cu rising to the top position and a new feature (Zn) appearing while Mg disappeared completely, indicating significant instability in feature importance determination.

XGBoost demonstrated similar instability with an even more pronounced performance drop, from an initial R^2 score of 0.956 to 0.8332 after removing the top feature. The feature importance rankings showed radical reorganization, with only Cu maintaining a presence in both sets while completely different features populated the remaining positions, highlighting the unreliability of supervised models for consistent feature importance interpretation.

In remarkable contrast, the unsupervised models exhibited perfect stability in their feature rankings. Feature Agglomeration (FA) maintained identical ranking positions for Al, Cu, Elongation (%), and class between the full and reduced feature sets, with only a moderate decrease in R^2 score from 0.9783 to 0.9259. Similarly, HVGS preserved the exact same feature order (Elongation (%), class, Processing_Solutionised + Artificially peak aged, Processing_Solutionised + Artificially over aged) despite experiencing a substantial performance drop from 0.9664 to 0.5554.

Most impressively, Spearman's correlation method demonstrated both perfect ranking stability and robust performance maintenance, with R^2 scores of 0.9725 and 0.9472. The method preserved the exact ordering of Cu, Eu, Yield Strength (MPa), and Zn between the two feature sets. This exceptional stability in both performance and feature importance rankings confirms that unsupervised models and non-parametric correlation approaches offer significantly more reliable feature selection frameworks for materials science applications where understanding true feature relationships is critical, rather than merely optimizing prediction accuracy.

4. Discussion

Our findings reveal a fundamental disconnect between prediction accuracy and feature importance stability in materials science applications of machine learning. While supervised models with/without SHAP achieved impressive R^2 scores (0.956–0.9802), they exhibited alarming instability in feature rankings following removal of the highest-ranked feature. This instability manifests as complete reorganization of importance hierarchies, with features appearing or disappearing entirely from the top rankings. Such volatility undermines the scientific reliability of conclusions drawn about material-property relationships.

In stark contrast, unsupervised methods and Spearman's correlation demonstrated perfect ranking stability despite feature reduction. This stability suggests these approaches may better capture intrinsic data relationships rather than merely optimizing prediction functions. The consistent performance of Spearman's correlation ($R^2 = 0.9725 \rightarrow 0.9472$) with identical feature ordering is particularly noteworthy, as it combines robust prediction with interpretational reliability.

A key factor contributing to the superior stability of unsupervised methods lies in their independence from target variables. Supervised models inherently optimize for prediction accuracy against labels, which can introduce and propagate target-specific biases throughout the feature selection process. These biases create artificial dependencies between features and target variables that may reflect statistical artifacts rather than physical relationships. In contrast, unsupervised methods evaluate feature importance based solely on intrinsic data characteristics—such as variance structures and natural clustering patterns—without being influenced by potentially noisy or incomplete target variables. This fundamental difference enables unsupervised approaches to identify more stable feature relationships that persist regardless of prediction objectives, making them particularly valuable for scientific discovery in materials science where understanding underlying physical mechanisms is paramount.

These results highlight that materials researchers should prioritize methodological approaches that maintain feature importance stability across different analytical conditions. We recommend implementing multiple model-agnostic methods alongside supervised learning to triangulate genuine relationships between materials parameters and properties. Furthermore, stability testing should become standard practice to validate feature selection before drawing scientific conclusions about material-property relationships.

Despite the promising results, this study has several limitations. First, our analysis relies on a single, albeit comprehensive, public dataset; future work should extend this stability testing framework across multiple materials science datasets spanning diverse applications to verify the generalizability of our findings. Second, we examined only a limited subset of available machine learning algorithms; expanding the comparison to include deep learning architectures, ensemble methods, and other emerging techniques would provide a more comprehensive assessment of feature selection stability.

Additionally, while our stability testing approach provides valuable insights, it represents only one dimension of feature importance validation. Future studies should investigate complementary validation strategies, including physics-informed feature selection that incorporates domain knowledge constraints, sensitivity analysis under different preprocessing conditions, and experimental validation of identified key features through targeted materials synthesis and characterization.

Finally, developing hybrid approaches that combine the predictive power of supervised methods with the stability of unsupervised feature selection represents a promising research direction. Such integrated

Table 1
cross-validation and feature importance ranking orders by algorithm.

model	CV5	[top 5 feature ranking order]	CV4	[Top 4 feature ranking order]
RF	0.9802	[Yield Strength (MPa), Al, Cu, Elongation (%), Mg]	0.9802	[Cu, Al, class, Zn]
XGBoost	0.9567	[Yield Strength (MPa), Processing_Naturally aged, Cu, Ti, Co]	0.9567	[Cu, class, Zn, Processing_Solutionised + Artificially peak aged]
FA	0.9784	[Yield Strength (MPa), Al, Cu, Elongation (%), class]	0.9784	[Al, Cu, Elongation (%), class]
HVGS	0.9664	[Yield Strength (MPa), Elongation (%), class, Processing_Solutionised + Artificially peak aged, Processing_Solutionised + Artificially over aged]	0.9664	[Elongation (%), class, Processing_Solutionised + Artificially peak aged, Processing_Solutionised + Artificially over aged]
Spearman	0.9724	[Al, Cu, Eu, Yield Strength (MPa), Zn]	0.9724	[Cu, Eu, Yield Strength (MPa), Zn]
RF-SHAP	0.9802	[Yield Strength (MPa), Cu, Al, Elongation (%), Mg]	0.9802	[Cu, Al, Zn, class]
XGB-SHAP	0.9802	[Yield Strength (MPa), Cu, Al, Elongation (%), Mg]	0.9802	[Cu, Al, class, Zn]

frameworks could potentially leverage the strengths of both paradigms while mitigating their respective limitations, ultimately advancing the scientific reliability of machine learning applications in materials science and engineering.

CRediT authorship contribution statement

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

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.commatsci.2026.114609.

Data availability

The authors do not have permission to share data.

References

[1] Q. Hou, X. Wu, Z. Li, et al., Artificial intelligence enabled microstructure prediction in Al alloy castings, in: J. Mater. Sci. Technol. 241, 2026, pp. 21–34, https://doi.org/10.1016/j.jmst.2025.02.093.
[2] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: learning a variables importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (2019) 177.
[3] L.H. Nazer, R. Zatarah, S. Waldrip, J.X.C. Ke, M. Moukheiber, A.K. Khanna, et al., Bias in artificial intelligence algorithms and recommendations for mitigation, PLOS Digit Health 2 (6) (2023) e0000278, https://doi.org/10.1371/journal.pdig.0000278.
[4] J. Ugirumurera, E.A. Bensen, J. Severino, J. Sanyal, Addressing bias in bagging and boosting regression models, Sci. Rep. 14 (1) (2024) 18452, https://doi.org/10.1038/s41598-024-68907-5.
[5] P. Alaimo Di Loro, D. Scacciatelli, G. Tagliaferri, 2-step gradient boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy, Stat. Methods Appl. 32 (2023) 237–270, https://doi.org/10.1007/s10260-022-00643-4.
[6] A.I. Adler, A. Painsky, Feature importance in gradient boosting trees with cross-validation feature selection, Entropy (Basel) 24 (5) (2022) 687, https://doi.org/10.3390/e24050687.
[7] P.M. Steiner, Y. Kim, The mechanics of omitted variable Bias: Bias Amplification and cancellation of offsetting biases, J. Causal Infer. 4 (2) (2016) 20160009, https://doi.org/10.1515/jci-2016-0009.
[8] Z.C. Lipton, The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57, https://doi.org/10.1145/3236386.3241340.
[9] K. Lenhof, L. Eckhart, L.M. Rolli, H.P. Lenhof, Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer, Brief. Bioinform. 25 (5) (2024) bbae379, https://doi.org/10.1093/bib/bbae379.
[10] H. Mandler, B. Weigand, A review and benchmark of feature importance methods for neural networks, ACM Comput. Surv. 56 (12) (2024) 318, https://doi.org/10.1145/3679012.
[11] J.L. Potharlanka, N.B. M, Feature importance feedback with deep Q process in ensemble-based metaheuristic feature selection algorithms, Sci. Rep. 14 (2024) 2923, https://doi.org/10.1038/s41598-024-53141-w.
[12] D. Wood, T. Papamarkou, M. Benatan, et al., Model-agnostic variable importance for predictive uncertainty: an entropy-based approach, Data Min. Knowl. Disc. 38 (2024) 4184–4216, https://doi.org/10.1007/s10618-024-01070-7.
[13] B. Bilodeau, N. Jaques, P.W. Koh, B. Kim, Impossibility theorems for feature attribution, Proc. Natl. Acad. Sci. USA 121 (2) (2024) e2304406120, https://doi.org/10.1073/pnas.2304406120.
[14] X. Huang, J. Marques-Silva, On the failings of Shapley values for explainability, Int. J. Approx. Reason. 171 (2024) 109112, https://doi.org/10.1016/j.ijar.2023.109112.
[15] D. Hooshyar, Y. Yang, Problems with SHAP and LIME in interpretable AI for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction, IEEE Access. 12 (2024) 137472–137490, https://doi.org/10.1109/ACCESS.2024.3463948.
[16] M.A. Lones, Avoiding common machine learning pitfalls, Patterns 5 (10) (2024) 101046, https://doi.org/10.1016/j.patter.2024.101046.
[17] C. Molnar, et al., General pitfalls of model-agnostic interpretation methods for machine learning models, in: A. Holzinger, R. Goebel, R. Fong, T. Moon, K. R. Müller, W. Samek (Eds.), xxAI - beyond Explainable AI. xxAI 2020 13200, Lecture Notes in Computer Science. Springer, 2022, https://doi.org/10.1007/978-3-031-04083-2_4.
[18] I. Kumar, C. Scheidegger, S. Venkatasubramanian, S. Friedler, Shapley residuals: quantifying the limits of the shapley value for explanations, Adv. Neural Inf. Proces. Syst. 34 (2021) 26598–26608.
[19] N. Bhat, A. Barnard, N. Birbilis, Aluminium alloy dataset for supervised learning. Mendeley Data, 2023, https://doi.org/10.17632/b6br4yk6r3.1.
[20] GitHub. alloy.py and al_data.csv. https://github.com/y-takefuji/alloy.
[21] Y. Liu, B. Guo, X. Zou, Y. Li, S. Shi, Machine learning assisted materials design and discovery for rechargeable batteries, Energy Storage Mater. 31 (2020) 434–450, https://doi.org/10.1016/j.ensm.2020.06.033.