NEXRES 13 (2026) 102457

Contents lists available at ScienceDirect

Next Research

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

Research Paper

Establishing true associations in dielectric properties: The role of mandatory consistency and dose-response relationships

Yoshiyasu Takefuji

SciencePark Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa, 252-0029, Japan

ARTICLE INFO

Keywords:
Feature importance stability
Materials informatics
Supervised learning bias
Unsupervised feature selection
Structure-property relationships

ABSTRACT

Dielectric properties are fundamental to chemical engineering, enabling applications in separation processes, process monitoring, reaction engineering, material characterization, heat transfer, and sensing systems. The growing adoption of machine learning in materials science is reflected in Next Research's expanding publication record on dielectric properties analysis, feature importance, SHAP, LIME, Random Forest, and XGBoost. However, AI misapplications remain widespread due to a critical but overlooked distinction: while prediction accuracy can be validated against ground truth, feature importance lacks equivalent verification, making supervised model interpretations unreliable for causal inference. This paper demonstrates these pitfalls by comparing supervised models (Random Forest, XGBoost), unsupervised approaches (feature agglomeration, highly variable gene selection), and non-target-prediction methods (Spearman's correlation) on a publicly available dielectric ceramic dataset. Using a novel leave-top-feature-out approach evaluating consistency and dose-response relationships, results show that unsupervised methods produce significantly more stable feature rankings, while supervised models exhibit erratic instability, confirming that prediction accuracy cannot serve as a proxy for three dimensional feature importance.

Introduction

Dielectric properties play a significant role in various aspects of power sources. In energy storage systems such as batteries and supercapacitors, dielectric properties are critical for electrolyte characterization, determining ionic conductivity and permittivity of electrolyte solutions, as well as evaluating separator membranes to optimize charge transport and safety. In fuel cells, dielectric measurements help monitor membrane hydration levels in proton exchange membrane fuel cells, where moisture content directly impacts proton conductivity, and also assist in quality control of membrane electrode assemblies during manufacturing. For solar energy systems, dielectric properties are used to characterize encapsulant materials such as EVA films in photovoltaic applications, ensuring long-term durability and minimal energy loss, while also supporting fault detection and degradation monitoring in solar panels through capacitive sensing techniques. In power electronics and conditioning, dielectric properties are essential in capacitor design and selection for power inverters and converters, where the dielectric constant and loss factor determine energy storage efficiency and thermal performance. For renewable energy storage, dielectric heating is applied in the processing and drying of electrode materials used in next-generation battery manufacturing. The most commonly measured parameters across these applications are the dielectric constant and loss factor, which indicate how a material stores and dissipates electrical energy, both of which are fundamental to improving the efficiency, reliability, and performance of modern power source technologies.

The growing integration of machine learning interpretability methods in materials science is clearly reflected in the publication record of the Next Research, which has featured 161 articles on dielectric properties analysis, 56 articles on feature importance, 86 on feature selection, 43 on SHapley Additive exPlanations (SHAP), 8 on Local Interpretable Model-agnostic Explanations (LIME), 158 on Random Forest, and 70 on XGBoost, despite the complete absence of ground truth against which feature assessments can be validated. This accelerating publication trend reflects the materials science community's expanding interest in establishing true associations or reliable feature importance through data-driven approaches. This paper theoretically discusses the complete absence of ground truth in feature assessments derived from supervised models, whether or not SHAP or LIME is applied, thereby highlighting the need for external feature validation, and empirically demonstrates the quality of selected features across three dimensions. This stands in contrast to existing studies, which rely solely on cross-

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.nexres.2026.102457
Received 11 September 2026; Accepted 19 September 2026
Available online 19 September 2026
3050-4759/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

validation accuracy under the implicit and unproven assumption that high target-prediction accuracy guarantees reliable feature importance, notwithstanding the absence of any ground truth for feature assessment.

Despite this momentum, AI misapplications remain widespread, largely due to insufficient understanding of supervised learning principles. A critical gap exists between prediction and interpretation: while prediction accuracy can be validated against ground truth labels, feature importance lacks any equivalent validation reference, making it inherently unverifiable. As a result, SHAP-based and LIME-based interpretations, which depend entirely on the underlying model, inherit this unvalidated nature and should not be treated as objective measures of true feature relevance. Moreover, establishing causal associations requires systematic evaluation of consistency, plausibility, and dose-response relationships, criteria that existing studies frequently overlook. Without such validation, feature importances should be regarded as predictive correlations rather than confirmed causal drivers.

To address these concerns, this paper identifies key methodological deficiencies in applying supervised models for feature evaluation and selection. Central to this critique is the distinction between two fundamentally different types of accuracy: predicting target outcomes and identifying important features. While prediction accuracy can be validated against ground truth, feature importance lacks equivalent verification mechanisms. A widespread misconception assumes that high predictive performance guarantees reliable feature importance assessments. This assumption is fundamentally flawed, as no guaranteed relationship exists between these two forms of accuracy, a disconnect that can render feature importance determinations unreliable even when predictive performance appears excellent.

This disconnect manifests directly in the inconsistency of feature importance rankings across different algorithms. Without ground truth for validation, different supervised models inevitably generate varied and often conflicting feature importance rankings, leading to model-dependent interpretations that are inherently inconsistent and potentially misleading. More fundamentally, feature importance in supervised models measures a feature's contribution to prediction accuracy rather than its actual causal relationship with the target variable. As a result, even highly accurate predictive models may produce unreliable feature importance rankings [1–9], since model-derived importance scores are optimized for prediction performance rather than for identifying true underlying relationships [10–20].

The function call explainer = SHAP(model) illustrates this dependency explicitly: SHAP is mathematically constrained to operate on a given supervised model, meaning its attributions can never exceed the reliability of that underlying model's feature-target associations. If the base model encodes spurious or unstable relationships, SHAP does not correct for this—it faithfully propagates, and can even amplify, these biases while presenting them through visually compelling and seemingly rigorous outputs. The same logic applies to LIME, which locally approximates the same unvalidated model. This creates a false sense of methodological rigor: the mathematical sophistication of Shapley values or local surrogate models does not compensate for the fundamental absence of ground truth in feature importance. In other words, explainability tools can make an unreliable result look more trustworthy without making it more true. Consequently, whether feature importance is derived directly from a supervised model or indirectly through SHAP or LIME, all such outputs remain fundamentally unverifiable and should not be interpreted as evidence of genuine causal or mechanistic relationships. This underscores the necessity of external, model-independent validation—such as the three-dimensional feature quality.

Recognizing that no algorithm can flawlessly determine true causal associations, evaluation efforts must prioritize two essential criteria: consistency and dose-response relationships [21–30]. Consistency refers to associations that remain stable across diverse studies, settings, and populations, thereby reducing the likelihood that observed relationships are artifacts of a specific dataset or modeling choice. Dose-response relationships, in turn, reveal systematic and monotonic changes in outcomes corresponding to variations in exposure levels, providing mechanistic plausibility to observed associations. Despite their critical importance, these two criteria have been largely neglected in research representing a significant gap that undermines the scientific rigor of feature importance analyses in the field.

This paper introduces a novel "leave-top-feature-out" (LTFO) approach to address these issues. The method works by: first selecting the top n features from the complete dataset (set1), then removing the highest-ranked feature to create a reduced dataset, subsequently reselecting the top n-1 features from this reduced dataset (set2), and finally comparing feature importance ranking orders between the two sets. This approach reveals that removing the highest-ranked feature significantly impacts the ranking order of remaining features, providing valuable insights into the consistency and dose-response relationships among variables. LTFO has been validated successfully across diverse domains and disciplines employing different datasets to demonstrate its methodological robustness [30–39].

This paper demonstrates the fundamental pitfalls of using supervised models for feature importance or selection by systematically evaluating both consistency and dose-response relationships across multiple methodologies. The key contribution is a comprehensive comparative analysis between supervised models (Random Forest, XGBoost) with and without SHAP or LIME, unsupervised approaches (feature agglomeration, highly variable gene selection), and non-target-prediction methods (Spearman's correlation) using a publicly available dataset for dielectric ceramic design. Through this analysis, the paper reveals that supervised models frequently exhibit instability in feature importance rankings due to their inherent model-specific architecture and potential label-driven errors. In contrast, unsupervised models and non-target prediction methods demonstrate significantly greater stability in feature ranking orders, primarily because they operate without the influence of label-driven biases. This finding represents an important advancement for materials science researchers seeking reliable feature importance or true association assessments for material property prediction and design.

Methods

This paper employs a publicly available dataset of dielectric ceramic design from Mendeley Data [40], which contains 15,172 instances and 55 features. We utilized the "data.xlsx" file, focusing on the "Dielectric Constant" as our target variable and retaining only the relevant feature columns. The final dataset had a shape of (15,172, 55) with all features being either integers or real numbers. All rows with missing values (NaN) were removed. For a fair comparison, default settings are used for all feature selection algorithms. For SHAP and LIME, randomly selected 100 instances are employed rather than full instances.

We implemented multiple feature selection algorithms to identify the most significant predictors of dielectric properties: Random Forest Importance, where features were ranked by their contribution to variance reduction; XGBoost Importance using the gain-based importance metric from gradient boosted trees; Feature Agglomeration (FA), which applied hierarchical clustering of features with selection of top features across all clusters; Highly Variable Gene Selection (HVGS) ranking features by their variance; and Spearman's Correlation ranking features by their absolute correlation coefficient with the target variable. Since FA, HVGS, and Spearman correlation do not provide built-in predictive capability, Random Forest was employed as the downstream model for cross-validation in these cases. In addition, four feature selection approaches—RF-SHAP, XGB-SHAP, RF-LIME, and XGB-LIME—were implemented for comparison.

For model evaluation, we used R-squared as our accuracy metric. Based on the leave-top-feature-out approach, we selected the top 10 features from the full feature set to create a reduced dataset (set1). Additionally, we removed the highest-ranked feature from the full set and reselected the top 9 features to create a second reduced dataset (set2). Both sets were evaluated using 5-fold cross-validation with their

respective feature selection algorithms. For FA, HVGS, and Spearman's correlation methods, Random Forest was used as the underlying model for cross-validation. The results were compiled into a summary table saved as "result.csv" containing the method name, cross-validation accuracy for both feature sets, along with the top 5 features from set1 and top 4 features from set2. For purposes of reproducibility and transparency, Python code, dielectric.py is publicly available at GitHub [41].

Results

Table 1 summarizes cross-validation accuracies alongside the top-ranked features identified by each algorithm. Overall, the results reveal a clear dissociation between predictive performance and the consistency and stability of feature importance rankings.

The supervised learning models (RF, XGB, RF-SHAP, XGB-SHAP, RF-LIME, and XGB-LIME) achieved high target prediction accuracy, with cv10 mean values ranging from 0.931 to 0.988 for most tree-based methods (RF, RF-SHAP, and XGB-SHAP). However, despite this strong predictive performance, these models exhibited pronounced inconsistency and instability in their feature ranking orders, both with and without the application of post-hoc interpretability techniques such as SHAP or LIME. Specifically, the identity, order, and composition of the top 5 and top 4 ranked features varied substantially across RF, XGB, RF-SHAP, XGB-SHAP, RF-LIME, and XGB-LIME, indicating that the application of SHAP or LIME did not reconcile the underlying disagreement in feature attribution across algorithms while SHAP and LIME often improve target prediction accuracy. Notably, XGB alone showed markedly lower and unstable cross-validation accuracy (cv10 = 0.303, cv9 = 0.307) coupled with a feature ranking that diverged considerably from all other supervised variants, further underscoring the instability of feature importance estimation in this model class. Bolded entries in Table 1 highlight the specific features whose rankings fluctuated inconsistently across these supervised approaches.

In marked contrast, the unsupervised model (FA) and the Spearman correlation-based ranking exhibited perfect consistency and stability in their feature rankings, with identical top 5 and top 4 feature sets and orders reproduced across the cv10 and cv9 evaluation schemes. This stability, however, was accompanied by comparatively low and unstable predictive accuracy for FA (cv10 = 0.689, cv9 = 0.330) and consistently low accuracy for Spearman (cv10 = cv9 = 0.310), suggesting that the stability of these rankings does not necessarily translate into strong predictive capability.

Notably, HVGS occupied an intermediate and favorable position between these two extremes: it demonstrated relatively high and stable prediction accuracy (cv10 = 0.715, cv9 = 0.716) while simultaneously achieving perfect consistency in its top 5 and top 4 feature rankings across both cross-validation schemes. This combination indicates that HVGS was the only method to reconcile reasonably strong predictive performance with fully reproducible feature attribution, distinguishing it from both the high-accuracy-but-unstable supervised models and the stable-but-lower-accuracy unsupervised/correlation-based approaches.

Discussion

This study empirically demonstrates a critical insight for materials informatics: high prediction accuracy does not guarantee reliable feature importance rankings. Random Forest achieved an accuracy of 0.988 in CV10 but dropped sharply to 0.391 in CV9, despite strong feature inconsistency and instability between the two sets. This dramatic decline demonstrates that the single feature removed between CV10 and CV9 dominantly drives the target prediction accuracy and exhibits feature inconsistency and dose-response.

The overwhelming majority of existing studies have failed to recognize that feature importance derived from supervised models, whether through intrinsic model attributes or post-hoc explanation tools such as SHAP and LIME, lacks any ground truth against which its correctness can be verified, yet such outputs are routinely treated as credible causal insights without validating consistency or dose-response relationships. Critically, SHAP and LIME are explanation techniques layered on top of an underlying supervised model; they can only redistribute or approximate the importance signal already embedded in that model, and thus cannot resolve instability that originates from the model's own sensitivity to training data composition, hyperparameters, or label noise. In other words, no matter how sophisticated the explanation method, it cannot manufacture ground truth that does not exist in the first place; it can only describe, with greater or lesser fidelity, the same ungrounded importance signal produced by the model itself. Improving predictive accuracy through better model tuning, or improving explanation fidelity through SHAP or LIME, therefore addresses an entirely different problem than establishing which features are consistently and causally associated with the target property. This systemic omission undermines the scientific integrity of feature importance claims across the field and empirically refutes the widespread assumption that high predictive performance, or the availability of a sophisticated post-hoc explanation method, validates feature importance interpretations. It is precisely this complete absence of ground

Table 1
Cross-validation accuracy and feature ranking orders across algorithms.

method	cv10 mean	cv9 mean	Top 5 feature rankings	Top 4 feature rankings
RF	0.988	0.391	temperature, pellet_diameter, BT_content, BMT_content, upper_limits_of_sintering_temp	pellet_diameter, **sintering_container_alumina_crucible, BT_content, BMT_content**
XGB	0.303	0.307	sintering_container_alumina_crucible, pellet_diameter, raw_MgO, upper_limits_of_calcination_temp, NaNbO3_content	pellet_diameter, **upper_limits_of_calcination_temp, calcination_time, ectrode_gold**
RF-SHAP	0.988	0.396	temperature, pellet_diameter, sintering_container_alumina_crucible, BT_content, BMT_content	pellet_diameter, sintering_container_alumina_crucible, BT_content, BMT_content
XGB-SHAP	0.979	0.394	temperature, BT_content, pellet_diameter, frequency, Lower_limits_of_sintering_temp	**pellet_diameter, BT_content, sintering_container_alumina_crucible, NaNbO3_content**
RF-LIME	0.973	0.97	NaNbO3_content, sintering_container_alumina_crucible, pellet_diameter, temperature, BMT_content	sintering_container_alumina_crucible, pellet_diameter, temperature, **BT_content**
XGB-LIME	0.931	0.935	NaNbO3_content, BT_content, temperature, sintering_container_alumina_crucible, pellet_diameter	**raw_MgCO3_5H2O, BT_content, temperature, sintering_container_alumina_crucible**
FA	0.689	0.33	upper_limits_of_calcination_temp, NaNbO3_content, milling_speed, electrode_bonding_temp, Lower_limits_of_sintering_temp	NaNbO3_content, milling_speed, electrode_bonding_temp, Lower_limits_of_sintering_temp
HVGS	0.715	0.716	milling_speed, frequency, electrode_bonding_temp, binder_removal_firing_temp, temperature	frequency, electrode_bonding_temp, binder_removal_firing_temp, temperature
Spearman	0.31	0.31	upper_limits_of_calcination_temp, NaNbO3_content, post-calcination_grinding_time, sintering_ramp_rate_C_per_min, electrode_bonding_time_min	NaNbO3_content, post-calcination_grinding_time, sintering_ramp_rate_C_per_min, electrode_bonding_time_min

Bold entries indicate features with inconsistent and unstable rankings.

truth that necessitates an external, independent validation framework capable of assessing feature importance through criteria, such as consistency and dose-response behavior, that are not themselves derived from the model being evaluated.

While SHAP and LIME often improve model predictions, as demonstrated in Table 1, they cannot examine feature consistency or dose-response relationships on their own, since both methods rely solely on a given supervised model and therefore inherit its unvalidated and potentially unreliable feature importance estimates. External feature validation frameworks, such as LTFO, are consequently needed to independently verify these relationships.

Feature importances derived from supervised models, whether obtained directly from model-internal metrics or through post-hoc explainers such as SHAP and LIME, are inherently skewed because these models optimize for prediction accuracy rather than true causal associations. Much of this instability stems from label-driven errors, as supervised models are guided by target values that may contain measurement noise or experimental variability, and SHAP or LIME explanations inherit this same label-driven bias because they are computed relative to the predictions of the very model that was fit to those noisy labels. Existing studies have treated prediction-optimized importance scores—including SHAP values, which are often presented as offering greater theoretical rigor through their basis in cooperative game theory—as causally validated associations, a conflation that this paper both theoretically critiques and empirically exposes. When the top-ranked feature is removed under the leave-top-feature-out framework, supervised models such as Random Forest and XGBoost produce almost entirely different feature importance rankings, generating inconsistent and potentially misleading structure-property interpretations. This instability would be expected to propagate directly into any SHAP- or LIME-based explanation of these same models, since both techniques operate on the model's fitted decision boundaries rather than on any independent, label-free structure in the data. That existing studies have overwhelmingly failed to apply cross-algorithm consistency checks—and have similarly failed to test whether SHAP or LIME explanations remain stable when the model is retrained under a leave-top-feature-out design—represents a critical gap that this paper directly addresses.

The two validation criteria most consequentially neglected by existing studies are feature consistency and dose-response relationships. Consistency, defined as the stability of feature importance rankings across different algorithms and dataset configurations, has been almost entirely absent from published analyses, including those relying on SHAP and LIME, which are typically computed for a single fitted model on a single dataset split without any cross-algorithm or cross-configuration comparison. Dose-response relationships, which require that removing a truly important feature produces predictable and proportional changes in model performance, have been similarly overlooked despite representing a fundamental criterion for inferring causal associations; notably, SHAP and LIME are not designed to test this criterion at all, since they quantify each feature's contribution to a given prediction rather than the causal consequence of the feature's removal from the modeling process. The collective failure to enforce these criteria means that a potentially substantial proportion of reported structure-property relationships in the materials science literature—including many explanations attributed to SHAP or LIME—may reflect modeling artifacts rather than genuine physical mechanisms. This paper addresses this failure both theoretically and empirically, demonstrating that supervised models, and by extension the post-hoc explainers built upon them, fail these criteria while unsupervised and correlation-based methods satisfy them.

In contrast, unsupervised and correlation-based methods demonstrated substantially stronger empirical support for genuine structure-property relationships. Feature Agglomeration, Highly Variable Gene Selection, and Spearman's correlation all maintained identical or nearly identical feature rankings across both validation sets, satisfying the consistency criterion that existing studies—including those relying on SHAP or LIME as their principal interpretability tool—have routinely ignored. These methods achieve stability precisely because they operate independently of target labels, avoiding the label-driven biases that systematically distort supervised importance rankings and that are consequently inherited, rather than corrected, by SHAP and LIME. A particularly striking finding emerged from Feature Agglomeration, where removing the top-ranked feature actually decreased predictive accuracy from 0.689 to 0.330. This counterintuitive result reveals that the highest-ranked feature was in fact substantially responsible for the model's predictive performance, directly demonstrating that structural feature importance and predictive contribution are analytically distinct quantities that existing studies, including those relying on SHAP- or LIME-based rankings, have problematically conflated.

The leave-top-feature-out framework also enabled rigorous evaluation of dose-response relationships. Supervised models exhibited irregular and inconsistent performance changes upon feature removal, incompatible with genuine dose-response behavior; because SHAP and LIME explanations are derived from these same unstable models, any dose-response narrative constructed from SHAP or LIME values would be equally unreliable, even though such narratives are rarely tested empirically in the literature. In contrast, HVGS not only maintained near-perfect stability in feature rankings (0.715 versus 0.716) but also demonstrated consistent and interpretable accuracy changes, providing converging evidence that its identified features, including milling speed, frequency, electrode bonding temperature, and binder removal firing temperature, genuinely govern dielectric behavior. This convergence of consistency and dose-response evidence represents a substantially more rigorous validation standard than prediction accuracy alone—or than the explanatory sophistication offered by SHAP and LIME—a standard that supervised models, and the post-hoc explainers built upon them, systematically fail to meet.

These findings carry direct practical implications for the materials science community. Researchers employing SHAP values, LIME, or other supervised model-derived importance metrics to guide experimental design or identify processing-structure-property linkages must recognize that such interpretations are model-dependent and unvalidated against any causal ground truth. It is particularly important to emphasize that adopting SHAP or LIME does not, by itself, resolve the inconsistency and instability documented in this study: these tools may enhance the transparency or local interpretability of a given model's predictions, and in some cases modestly improve perceived explanatory plausibility, but they operate downstream of the same unstable model and therefore cannot correct for ranking instability that arises from the model-fitting process itself. Without complementary evaluation using consistency and dose-response criteria, reported feature importances—regardless of whether they are derived from raw model metrics or from SHAP and LIME—may reflect modeling artifacts rather than genuine material physics. These results therefore advocate for a paradigm shift in materials informatics practice, wherein feature importance validation is treated as a distinct and mandatory analytical step that is not conflated with predictive performance evaluation or with the use of increasingly popular post-hoc explanation tools such as SHAP and LIME.

A significant limitation of this study is its focus on a single dielectric ceramic dataset, which constrains immediate generalizability. Future work should systematically extend this leave-top-feature-out validation framework across diverse material classes, including piezoelectrics, thermoelectrics, and structural alloys, to determine whether the observed superiority of unsupervised methods generalizes beyond dielectric ceramics, and should explicitly test whether SHAP- and LIME-based rankings remain unstable under the same leave-top-feature-out protocol in these additional material systems. Additionally, developing hybrid frameworks that integrate the predictive power of supervised models—potentially augmented with SHAP or LIME for local interpretability—with the causal stability of unsupervised approaches represents a promising research direction, separating the two

analytically distinct tasks that are currently conflated in existing studies. Establishing standardized community protocols that explicitly mandate evaluation of consistency and dose-response relationships, rather than relying on the presence of a SHAP or LIME analysis as a proxy for interpretive rigor, will be essential for ensuring that machine learning-derived insights genuinely advance structure-property understanding and accelerate responsible materials discovery.

This paper aims to provide an external feature validation tool that examines feature consistency and dose–response relationships alongside cross-validation accuracy—that is, three dimensions of evaluation—whereas existing studies typically rely on cross-validation accuracy as a single dimension, which is insufficient to substantiate reliable feature importance. From a feature engineering perspective, the ideal outcome is one in which higher predictive accuracy coincides with a consistent, physically ordered set of features. Feature grouping strategies, such as separating process history from chemical composition, can be applied as a preprocessing step to narrow the hypothesis space of candidate features prior to applying this three-dimensional validation framework.

Funding

This research has no fund.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Code availability

Python code is publicly available on GitHub.

AI use

Not applicable.

According to scholargps (prior 5 years) on productivity

Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 25th out of 1287,415 scholars in life sciences, 22nd out of 805,705 in COVID-19, and 1st out of 109,919 in environmental sciences. https://scholargps.com/scholars/14204688896425/yoshiyasu-takefuji?.

CRediT authorship contribution statement

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

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 author has no permission to share data.

References

[1] T. Parr, J. Hamrick, J.D. Wilson, Nonparametric feature impact and importance, Inf. Sci. (Ny.) 653 (2024) 119563, https://doi.org/10.1016/j.ins.2023.119563.
[2] D.S. Watson, M.N. Wright, Testing conditional independence in supervised learning algorithms, Mach. Learn. 110 (8) (2021) 2107–2129, https://doi.org/10.1007/s10994-021-06030-6.
[3] 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.
[4] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (2019) 177.
[5] 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.
[6] H. Mandler, B. Weigand, A review and benchmark of feature importance methods for neural networks, ACM Comput. Surv. 56 (12) (2024) 1–30, https://doi.org/10.1145/3679012.
[7] J.L. Potharlanka, M.N. Bhat, Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms, Sci. Rep. 14 (1) (2024) 2923, https://doi.org/10.1038/s41598-024-53141-w.
[8] D. Wood, T. Papamarkou, M. Benatan, et al., Model-agnostic variable importance for predictive uncertainty: an entropy-based approach, Data Min. Knowl. Discov. 38 (2024) 4184–4216, https://doi.org/10.1007/s10618-024-01070-7.
[9] C. Molnar, G. König, J. Herbinger, 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: International Workshop, Held in Conjunction with ICML 2020, Vienna, AustriaCham, Springer International Publishing, 2022, pp. 39–68, https://doi.org/10.1007/978-3-031-04083-2_4. July 18, 2020Revised and Extended Papers.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] P.M. Steiner, Y. Kim, The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases, J. Causal. Inference 4 (2) (2016) 20160009, https://doi.org/10.1515/jci-2016-0009.
[15] M.L. Wallace, L. Mentch, B.J. Wheeler, A.L. Tapia, M. Richards, S. Zhou, et al., Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction, BMC Med. Res. Methodol. 23 (1) (2023) 144, https://doi.org/10.1186/s12874-023-01965-x.
[16] T. Salles, L. Rocha, M. Gonçalves, A bias-variance analysis of state-of-the-art random forest text classifiers, Adv. Data Anal. Classif. 15 (2021) 379–405, https://doi.org/10.1007/s11634-020-00409-4.
[17] R. Dunne, R. Reguant, P. Ramarao-Milne, et al., Thresholding Gini variable importance with a single-trained random forest: an empirical Bayes approach, Comput. Struct. Biotechnol. J. 21 (2023) 4354–4360, https://doi.org/10.1016/j.csbj.2023.08.033.
[18] M. Huti, T. Lee, E. Sawyer, A.P. King, An investigation into race bias in random forest models based on breast DCE-MRI derived radiomics features, in: Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging 14242, 2023, pp. 225–234, https://doi.org/10.1007/978-3-031-45249-9_22.
[19] T.T. Nguyen, J.Z. Huang, T.T. Nguyen, Unbiased feature selection in learning random forests for high-dimensional data, Sci. World J. 2015 (2015) 471371, https://doi.org/10.1155/2015/471371.
[20] C. Strobl, A.L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinform. 8 (2007) 25, https://doi.org/10.1186/1471-2105-8-25.
[21] J.P. Ioannidis, Why most discovered true associations are inflated, Epidemiology 19 (5) (2008) 640–648, https://doi.org/10.1097/EDE.0b013e31818131e7.
[22] V. Prasad, A.B. Jena, Prespecified falsification end points: can they validate true observational associations? JAMA 309 (3) (2013) 241, https://doi.org/10.1001/jama.2012.96867.
[23] J.P.A. Ioannidis, Genetic associations: false or true? Trends Mol. Med. 9 (4) (2003) 135–138, https://doi.org/10.1016/S1471-4914(03)00030-3.
[24] 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, e1.
[25] 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.
[26] 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.
[27] E. Stamatakis, M. Ahmadi, R.K. Biswas, P. Del, B. 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.
[28] 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.
[29] B.R. Underwood, I. Lourida, J. Gong, S. Tamburin, E.Y.H. Tang, E. Sidhom, et al., Deep Dementia Phenotyping (DEMON) Network. Data-driven discovery of associations between prescribed drugs and dementia risk: a systematic review, Alzh. Dement. (N Y) 11 (1) (2025) e70037, https://doi.org/10.1002/trc2.70037.
[30] 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.
[31] Y. Takefuji, Beyond XGBoost and SHAP: unveiling true feature importance, J. Hazard Mater. 488 (2025) 137382, https://doi.org/10.1016/j.jhazmat.2025.137382.
[32] Y. Takefuji, Stability of feature attribution: contrasting supervised and unsupervised selection for radiopathomics and TCGA outcomes, Cancer Lett. 636 (2026) 218133, https://doi.org/10.1016/j.canlet.2025.218133.
[33] Y. Takefuji, The reliability gap: why high predictive accuracy doesn't guarantee stable feature importance, Mar. Pollut. Bull. 226 (2026) 119398, https://doi.org/10.1016/j.marpolbul.2026.119398.
[34] Y. Takefuji, Evaluating linear parametric Poisson regression vs nonparametric unsupervised learning in ulcerative colitis data, Gastroenterology 171 (2026) 384–386, https://doi.org/10.1053/j.gastro.2026.04.011. Published online April.
[35] Y. Takefuji, When high accuracy misleads: stability limits of supervised feature importance in QSAR biodegradation, Chemosphere 396 (2026) 144846, https://doi.org/10.1016/j.chemosphere.2026.144846.
[36] Y. Takefuji, Beyond prediction: assessing stability in feature selection methods for materials science applications, Comput. Mater. Sci. 267 (2026) 114609, https://doi.org/10.1016/j.commatsci.2026.114609.
[37] Y. Takefuji, Ensuring reliable feature importance in food chemistry AI, Food Chem. 509 (2026) 148515, https://doi.org/10.1016/j.foodchem.2026.148515.
[38] Y. Takefuji, The absence of true association validation in supervised feature importance: evidence from battery capacity prediction for second-life applications, J. Energy Stor. 180 (2026) 124146, https://doi.org/10.1016/j.est.2026.124146.
[39] Y. Takefuji, A leave-top-feature-out framework for validating feature importance reliability in concrete compressive strength modeling, Civil Eng. Des. (2026), https://doi.org/10.1002/cend.70026. Published online.
[40] P.Q. Qin, Dataset and code for "A domain-informed automated machine learning framework for dielectric ceramic design: application to the $BaTiO_3$–$Bi(Mg_{1/2}Ti_{1/2})O_3$ system, Mend. Data (2025) V1, https://doi.org/10.17632/7gybvssrpf.1.
[41] GitHub, limeshap.Py, in: , 2026. https://github.com/y-takefuji/Dielectric.