Cancer Letters 636 (2026) 218133

Contents lists available at ScienceDirect

Cancer Letters

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

COR

Stability of feature attribution: Contrasting supervised and unsupervised selection for radiopathomics and TCGA outcomes

ARTICLE INFO

Keywords:
Interpretability
SHAP
Feature-selection stability
Supervised vs unsupervised
Radiopathomics

ABSTRACT

Interpretable machine learning is increasingly used in oncology, yet feature attributions from supervised models (e.g., Random Forest, XGBoost) can be unstable and bias-prone when grounded solely in SHAP explanations. We contrast target-prediction accuracy with the reliability of feature-importance estimates and assess stability via feature-elimination tests on TCGA data (705 samples, 1936 features). While supervised models achieved modest gains (Random Forest ROC–AUC: 0.8851 to 0.8865; XGBoost: 0.8681 to 0.8695), their feature-selection stability was low (6/10 and 3/10 retained, respectively). Unsupervised and non-target methods were markedly more stable: Feature Agglomeration, Highly Variable Gene Selection, and Spearman retained 10/10 features with unchanged performance (0.8823, 0.8823, 0.8766). We recommend combining unsupervised criteria with causal design and external validation to mitigate model-specific biases.

Huang et al. developed a multimodal radiopathomics signature to predict response to immunotherapy-based combination therapy in gastric cancer using interpretable machine learning [1]. They evaluated a range of classifiers, including Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), Naive Bayes (NB), Random Forest (RF), XGBoost (XGB), and Support Vector Machine (SVM). Model performance was reported using receiver operating characteristic–area under the curve (ROC–AUC), supplemented by standard metrics and confusion matrices. To enhance interpretability, they applied SHapley Additive exPlanations (SHAP) to quantify the contribution of individual radiopathomic features to predicted outcomes. In transparent machine learning, SHAP values provide instance-level feature attributions by estimating each feature's marginal contribution to a model's prediction across data points [1].

This paper is intended as a methodological warning against relying solely on supervised models with SHAP. Three methodological considerations are critical for interpreting such analyses. First, supervised learning entails two distinct notions of performance: target-prediction accuracy (which can be validated against ground-truth labels) and the reliability of feature-importance estimates (for which ground truth typically does not exist). Second, because feature importance lacks a ground-truth reference, high predictive accuracy does not ensure that inferred importances reflect true causal or associative relationships [2–10]; importances capture contributions to the model's predictions and can be biased by data artifacts, confounding, leakage, or model misspecification. Third, when explanations are computed as 'explain = SHAP(model)' for a supervised model, SHAP attributes importance relative to that model's decision function. Consequently, SHAP can inherit, propagate, and sometimes amplify a model's existing biases in feature attribution [11–20]. Robust causal inference strategies and external validation are therefore necessary to substantiate any biological or clinical interpretations derived from feature importances.

These points raise both theoretical and empirical concerns about using supervised models for feature attribution and, by extension, about relying on SHAP for scientific interpretation when explanations are anchored to a single model. While SHAP is a powerful and widely adopted diagnosis tool, exclusive dependence on any given supervised model can yield unreliable attributions. A practical way to assess stability is through feature-perturbation or elimination tests: iteratively remove top-ranked features from the full set and examine whether feature rankings and performance remain consistent. Supervised models often display instability under such tests because attributions are label-driven and sensitive to confounding, outcome prevalence, collinearity, and shifts in decision boundaries when correlated predictors are removed. By contrast, unsupervised methods can exhibit greater stability in rankings because they do not condition on labels; they prioritize structure intrinsic to the feature space (e.g., variance, clustering cohesion, manifold geometry), making them less susceptible to label noise, mis-specified loss functions, class imbalance, and label leakage. Moreover, unsupervised criteria (such as redundancy reduction and stability selection across bootstraps) tend to be more robust to multicollinearity and reweight correlated features more evenly, reducing the volatility seen in supervised attributions. Still, greater stability does not imply causal validity; feature importance quantifies contribution to prediction, not underlying mechanistic truth.

There is no single algorithm that can accurately recover "true" associations between variables from observational data. This paper therefore advocates a multifaceted strategy that combines unsupervised models with non-target supervised analyses, alongside rigorous sensitivity and stability checks. Approaches such as feature agglomeration (FA) and highly variable gene selection (HVGS) can identify stable, structure-driven feature sets. Rank-based association measures like Spearman's correlation with p-values offer a nonlinear nonparametric, monotonic, and robust alternative that does not rely on target labels. Together with external validation, causal design principles (e.g., negative controls, instrumental variables where appropriate), and cross-

https://doi.org/10.1016/j.canlet.2025.218133
Received 27 August 2025; Received in revised form 21 October 2025; Accepted 10 November 2025
Available online 12 November 2025

cohort replication, these methods can mitigate model-specific biases and support more reliable scientific interpretation.

In the absence of Huang et al.'s datasets, we evaluate feature-selection effectiveness on publicly available TCGA data (705 samples, 1936 features) [21] by selecting the Top 10 features from the full set and assessing predictive performance via cross-validated accuracy, where higher accuracy indicates better selection. We compare supervised models (Random Forest, XGBoost, Logistic Regression), unsupervised methods (Feature Agglomeration, Highly Variable Gene Selection), and a non-target supervised approach (Spearman's correlation). To probe stability, we remove the highest-ranked feature from the full set to form a reduced dataset, reselect the Top 9 features, and then compare ranking concordance between the two selections, implementing a perturbation-based protocol aligned with best practices for stability assessment. This procedure quantifies sensitivity of rankings to feature removal while benchmarking predictive utility under consistent cross-validation and supports more reliable interpretation alongside external validation.

Removing the highest-ranked feature represents a targeted perturbation specifically designed to test the underlying robustness of the feature importance hierarchy, not just set membership. This approach directly addresses whether features derive their importance from true underlying relationships or from correlations with the top feature.

While conventional stability metrics like Jaccard indices quantify set overlap, they fail to capture the critical ordering information in feature rankings. Our perturbation test specifically evaluates whether secondary features maintain their relative positions when the dominant feature is removed—a stronger criterion than set membership stability. If rankings dramatically shuffle after removing the top feature, this reveals potential collinearity effects or signal "borrowing" that conventional stability measures would miss.

Our targeted approach deliberately tests the worst-case perturbation scenario that most directly challenges the reliability of feature rankings for biological interpretation—precisely the use case we caution against in the manuscript.

Our tests reveal a clear contrast in stability between supervised and unsupervised approaches. Among the supervised models, Random Forest improves slightly from 0.8851 (Top 10) to 0.8865 (combined) but retains only 6/10 features in the stability test, while XGBoost increases from 0.8681 to 0.8695 yet shows the lowest stability at 3/10 features. In contrast, the unsupervised methods are markedly more stable: Feature Agglomeration holds steady at 0.8823 for both Top 10 and combined and achieves perfect stability with 10/10 features retained; Spearman's correlation also shows identical performance at 0.8766 and perfect stability at 10/10 features. Highly Variable Gene Selection (HVGS) maintains identical scores at 0.8823, reinforcing the overall pattern that unsupervised and non-target methods offer higher feature-selection stability, even when predictive performance remains unchanged. "combined" refers to a hybrid feature set created by taking the top 1 feature from the original dataset and joining it with the top 9 features selected from the reduced dataset (where the original top feature has been removed). For purposes of reproducibility and transparency, Python code, stability.py, is publicly available at GitHub [22].

To accurately calculate true associations [23–28], we must examine two key issues:

1. Consistency: True associations should replicate across different studies, settings, and populations. Our approach specifically addresses this by examining stability across feature subsets.
2. Dose-response relationship: True associations should demonstrate systematic changes in outcome with varying levels of exposure. Our leave-one-out approach explicitly tests this by measuring how prediction changes when specific features are removed or modified.

While unsupervised methods might achieve stability by ignoring the outcome entirely, and supervised methods might be unstable due to multiple equivalent predictive feature subsets, our method navigates this trade-off. By implementing consistency checks via stability assessment and examining dose-response relationships through systematic feature perturbation, we provide a more robust framework for distinguishing reliable feature associations from artifacts of model selection. Rather than merely accepting instability as an inherent property of having multiple equivalent models, our approach helps identify which feature associations persist across the space of near-equivalent models, offering stronger evidence for scientific interpretation.

Consent to participate

Not applicable.

Ethics approval

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Code availability

Not applicable.

AI use

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.

References

[1] W. Huang, X. Wang, R. Zhong, et al., Multimodal radiopathomics signature for prediction of response to immunotherapy-based combination therapy in gastric cancer using interpretable machine learning, Cancer Lett. 631 (2025) 217930, https://doi.org/10.1016/j.canlet.2025.217930.
[2] T. Parr, J. Hamrick, J.D. Wilson, Nonparametric feature impact and importance, Inf. Sci. 653 (2024) 119563, https://doi.org/10.1016/j.ins.2023.119563.
[3] 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.
[4] C. Molnar, G. König, J. Herbinger, et al., General Pitfalls of model-agnostic Interpretation Methods for Machine Learning Models, Springer International Publishing, 2022, https://doi.org/10.1007/978-3-031-04083-2_4.
[5] Z.C. Lipton, The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery, ACM Queue 16 (3) (2018) 31–57, https://doi.org/10.1145/3236386.3241340.
[6] 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.
[7] 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, Briefings Bioinf. 25 (5) (2024) bbae379, https://doi.org/10.1093/bib/bbae379.
[8] 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.

[9] 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.
[10] 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.
[11] L. Wu, A review of the transition from Shapley values and SHAP values to RGE, Statistics (2025) 1–23, https://doi.org/10.1080/02331888.2025.2487853.
[12] B. Bilodeau, N. Jaques, P.W. Koh, B. Kim, Impossibility theorems for feature attribution, Proc. Natl. Acad. Sci. U. S. A 121 (2) (2024) e2304406120, https://doi.org/10.1073/pnas.2304406120.
[13] 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.
[14] 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.
[15] M.A. Lones, Avoiding common machine learning pitfalls, Patterns 5 (10) (2024) 101046, https://doi.org/10.1016/j.patter.2024.101046.
[16] 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. Vol 13200. Lecture Notes in Computer Science, Springer, 2022, p. 4, https://doi.org/10.1007/978-3-031-04083-2_4.
[17] I. Kumar, C. Scheidegger, S. Venkatasubramanian, S. Friedler, Shapley residuals: quantifying the limits of the shapley value for explanations, Adv. Neural Inf. Process. Syst. 34 (2021) 26598–26608.
[18] O. Létoffé, X. Huang, J. Marques-Silva, Towards trustable SHAP scores, Proc. AAAI Conf. Artif. Intell. 39 (17) (2025) 18198–18208, https://doi.org/10.1609/aaai.v.
[19] A.V. Ponce-Bobadilla, V. Schmitt, C.S. Maier, S. Mensing, S. Stodtmann, Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development, Clin Transl Sci 17 (11) (2024) e70056, https://doi.org/10.1111/cts.70056.
[20] H. Coupland, N. Scheidwasser, A. Katsiferis, et al., Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis, BMC Public Health 25 (1) (2025) 1520, https://doi.org/10.1186/s12889-025-22705-4. Published 2025 Apr 24.
[21] G. Ciriello, M.L. Gatza, A.H. Beck, et al., Comprehensive molecular portraits of invasive lobular breast cancer, Cell 163 (2) (2015) 506–519, https://doi.org/10.1016/j.cell.2015.09.033.
[22] GitHub, Stability.py. https://github.com/y-takefuji/cell/blob/main/stability.py.
[23] J.P.A. Ioannidis, Why most discovered true associations are inflated, Epidemiology 19 (5) (2008) 640–648, https://doi.org/10.1097/EDE.0b013e31818131e7.
[24] 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.
[25] A. Singh, L. Southam, K. Hatzikotoulas, 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.
[26] V. Prasad, A.B. Jena, Prespecified falsification end points: can they validate true observational associations? JAMA 309 (3) (2013) 241–242, https://doi.org/10.1001/jama.2012.96867.
[27] 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.
[28] 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.

Yoshiyasu Takefuji[1]
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan
E-mail address: takefuji@keio.jp.

[1] **According to ScholarGPS,** Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 54th out of 395,884 scholars in neural networks (AI), 23rd out of 47,799 in parallel computing, and 14th out of 7222 in parallel algorithms. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.