Marine Pollution Bulletin 226 (2026) 119398

Contents lists available at ScienceDirect

Marine Pollution Bulletin

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

The reliability gap: Why high predictive accuracy doesn't guarantee stable feature importance

Yoshiyasu Takefuji

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

ARTICLE INFO

Keywords:
Feature importance stability
Supervised learning limitations
SHAP interpretation
Unsupervised feature selection
Environmental risk prediction

ABSTRACT

Marine Pollution Bulletin increasingly applies machine learning and explainable AI to pollutant and shellfish poisoning risk, exemplified by PCA-based source apportionment and SHAP-based feature attribution. However, linear PCA may misrepresent structure in inherently nonlinear environmental data, and existing studies often treat model-derived feature importances as evidence of true associations without assessing consistency or dose–response relationships. This paper clarifies that supervised models possess two distinct accuracies: prediction and feature importance, and only prediction can be validated against ground truth. Using a Basque coastal dataset (8195 instances, 14 features) with chlorophyll-a as a proxy for paralytic shellfish poisoning risk, we introduce a leave-top1-out procedure to test ranking stability. Random Forest and XGBoost with and without SHAP show pronounced instability, indicating biased, model-dependent importances. In contrast, unsupervised and non–target-prediction methods yield perfectly stable rankings while matching or exceeding supervised performance, supporting routine stability, consistency, dose–response, and linearity checks in environmental ML studies.

1. Introduction

Marine Pollution Bulletin has published 34 articles on feature importance (10 in 2025 and 9 in 2026), 48 on feature selection (12 in 2025 and 11 in 2026), 28 on SHAP (5 in 2025 and 9 in 2026), and 326 on machine learning (81 in 2025 and 56 in 2026), reflecting a rapidly growing interest in using machine learning and artificial intelligence for robust association assessment. However, limited understanding of the fundamental principles of supervised learning has led to frequent misapplications of AI-based feature importance in shellfish poisoning risk analyses.

Our contribution is to clarify that feature importances derived from supervised models are inherently unreliable as causal or mechanistic indicators because they lack ground truth. Supervised models possess two distinct types of accuracy: target prediction accuracy and feature importance accuracy. While prediction accuracy can be validated against labeled outcomes, feature importance has no direct ground truth for validation. Many researchers overlook this and interpret feature importance without performing the necessary checks for consistency and dose–response relationships, leading to biased or incorrect conclusions.

Turull et al. (2026) investigated the spatial distribution of pollutants along the Galician coast using starfish as bioindicators. To interpret the processed data and identify potential sources of the compounds, they performed a principal component analysis (PCA), which yielded seven principal components with eigenvalues greater than 1. The first principal component (PC1) accounted for approximately 31% of the total variance, with strong positive loadings ($>|0.7|$) for Mo, Sn, U, and Zn, and strong negative loadings for Sr and V, suggesting contrasting geochemical or anthropogenic influences on these elements. However, because PCA is a strictly linear technique that represents each component as a linear combination of the original variables, it cannot fully capture nonlinear structure in the data; accordingly, its outputs may differ from those obtained using nonlinear, nonparametric methods, and should be interpreted with caution when underlying relationships are suspected to be complex or nonlinear.

Marzidovšek et al. (2024) applied explainable machine learning to predict diarrhetic shellfish poisoning events in the Adriatic Sea using long-term monitoring data. They used decision trees (DT), random forests (RF), support vector machines (SVM), and shallow artificial neural networks (ANN). The SHapley Additive exPlanations (SHAP) method (Chowdhury et al., 2025; Mamun et al., 2025) was employed to quantify the contribution of individual variables to model outputs; for RF, the SHAP TreeExplainer, an algorithm tailored to tree ensemble methods,

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.marpolbul.2026.119398
Received 14 December 2025; Received in revised form 25 January 2026; Accepted 5 February 2026
Available online 9 February 2026

was applied. However, their analysis focused on prediction and variable contribution within the fitted models, and did not examine true association in terms of consistency, temporality, or dose–response relationships.

Bi et al. (2025) investigated paralytic shellfish poisoning risk along the west coast of Canada using a stacked ensemble of 11 supervised learning models, achieving high predictive performance (AUC $\approx$ 0.95). Global feature importance was evaluated by computing mean absolute SHAP values for each predictor across the dataset. Together, these studies illustrate the increasing reliance on advanced, explainable machine learning techniques in environmental risk assessment.

In this context, it is crucial to recognize that supervised models possess two distinct forms of accuracy: target prediction accuracy and feature-importance accuracy. While prediction accuracy can be validated against ground-truth labels, feature importance has no corresponding ground truth, meaning its accuracy cannot be directly verified or calibrated. This paper therefore raises theoretical and empirical concerns about using supervised models for feature importance or feature selection, with or without SHAP, precisely because no ground truth exists for feature importance and because importance estimates are inherently model-dependent. As a result, reported importances are easily overinterpreted as evidence of causal or even robust associative effects. Bi et al. themselves showed that different models yield different feature-importance rankings. Thus, model-derived importance reflects each variable's contribution to prediction within a specific model class and training regime, rather than its true causal or associative influence in the underlying environmental system.

To approximate true associations, at least two critical elements should be examined: consistency and dose–response relationships (Ioannidis, 2008; Prasad and Jena, 2013; Ioannidis, 2003; Roberts et al., 2019; Lai et al., 2025; Prada et al., 2025; Stamatakis et al., 2025; Ye et al., 2024; Underwood et al., 2025; Takefuji, 2025; Singh et al., 2025). Consistency refers to associations replicated across studies, settings, and populations. Dose–response refers to systematic changes in outcome with varying levels of exposure. Neither is guaranteed by high predictive performance or by SHAP values alone.

Existing studies have largely neglected consistency and dose–response relationships when assessing true associations. This paper shows that, without such assessment, inferences drawn from supervised models about feature importance or feature selection are unreliable. Using a publicly available dataset, we demonstrate that supervised models frequently exhibit unstable feature-importance rankings driven by label noise and model specification, whereas unsupervised models and non–target-prediction methods can yield markedly more stable feature rankings due to the absence of label-driven errors.

To address this gap, we propose a simple leave-top1-out procedure to evaluate the robustness of feature rankings. First, select the top (n) features from the full feature set (set 1). Second, remove the highest-ranked feature from the full set to construct a reduced dataset. Third, reselect the top (n-1) features from this reduced dataset (set 2). Finally, compare the rankings of the overlapping features between set 1 and set 2. If omitting the top feature substantially disrupts the ordering of the remaining features, this instability indicates a lack of consistency and undermines naive interpretations of feature importance and implied dose–response relationships.

2. Methods

Due to the absence of datasets from Bi et al. and Marzidovšek et al., this paper utilizes a publicly available dataset from Mendeley data with 8195 instances and 14 features (Solaun et al., 2025). CHLOROPHYLL-a (μg l^{-1}) is selected as the target for Paralytic Shellfish Poisoning (PSP) risk assessment from variable list because it is the only one that directly reflects the biological component driving PSP: phytoplankton biomass. PSP is caused by saxitoxins produced by certain phytoplankton (mainly dinoflagellates such as Alexandrium), and these toxic species bloom under favorable environmental conditions.

For feature selection, we compare supervised models (Random Forest: RF, RF-SHAP; XGBoost: XGB, XGB-SHAP), unsupervised models (Feature Agglomeration: FA; Highly Variable Gene Selection: HVGS), and a non–target-prediction method (Spearman correlation). To ensure a fair comparison, all supervised models are used with their library defaults and no hyperparameter tuning. First, we select the top five features from the full set (set1: CV5) and perform cross-validation. We then remove the highest-ranked feature from the full set to create a reduced dataset, reselect the top four features from this reduced set (set2: CV4), and cross-validate again. For FA, HVGS, and Spearman-based rankings, RF (with default settings) is used as the downstream model for cross-validation.

3. Results

Table 1 reveals key insights about the stability and performance of different feature selection methods for predicting chlorophyll-a concentrations in marine environments. Our analysis shows that supervised models (Random Forest and XGBoost), both with and without SHAP integration, demonstrate considerable instability in their feature rankings. When the top feature (Date) is removed, the relative importance of remaining features shifts unpredictably. For example, RF originally ranks features as “Date, LT, DO, Sal, Time” but after Date removal, the order becomes “LT, Time, DO, Sal” - showing Time moving from 5th to 2nd position. Similar instability appears in XGB-SHAP where PAR moves from 4th to last position after Date removal.

In contrast, unsupervised methods exhibit perfect stability in their feature rankings. Feature Agglomeration (FA) and Highly Variable Gene Selection (HVGS) maintain the exact same order of features after removing the top feature. FA preserves “Time, PAR, LT, Press” in the same sequence as they appeared in the top 5, while HVGS similarly maintains “Time, PAR, Sal, Sigma” in their original order. Spearman correlation, which doesn't rely on target prediction during feature ranking, also shows perfect stability in its feature order (Temp, Sigma, LT, Sal) after removing DO.

Despite differences in stability, unsupervised methods achieve competitive accuracy with FA showing the highest CV5 score (0.9143) among all methods. HVGS (0.9044) outperforms both SHAP-based methods (0.8988, 0.8991). This suggests unsupervised approaches can match or exceed supervised methods in predictive performance while offering superior stability. This analysis demonstrates that unsupervised feature selection methods may be preferable for chlorophyll-a prediction, as they offer both excellent predictive performance and consistent feature rankings that are less likely to change when the dataset is modified.

Table 1
cross-validation accuracy and feature rankings.

Method	CV5	CV4	Top5 feature rankings	Top4 feature rankings
RF	0.9071	0.8631	Date, LT, DO, Sal, Time	LT, Time, DO, Sal
XGB	0.9073	0.8634	Date, DO, LT, Sal, Time	DO, Sal, LT, Time
RF-SHAP	0.8988	0.8313	Date, LT, DO, Time, PAR	LT, DO, Time, Temp
XGB-SHAP	0.8991	0.8319	Date, LT, Time, PAR, DO	LT, Time, DO, PAR
FA	0.9143	0.8705	Date, Time, PAR, LT, Press	Time, PAR, LT, Press
HVGS	0.9044	0.8344	Date, Time, PAR, Sal, Sigma	Time, PAR, Sal, Sigma
Spearman	0.8061	0.6928	DO, Temp, Sigma, LT, Sal	Temp, Sigma, LT, Sal

4. Discussion

This study challenges the widespread practice of treating supervised feature importance as evidence of true environmental associations in marine pollution and shellfish poisoning research. Although supervised models achieved high predictive accuracy, their feature-importance rankings were highly unstable under the leave-top1-out perturbation, even when performance remained strong. This instability indicates that feature importances derived from supervised models are inherently biased and skewed by label noise, model specification, and collinearity, and thus are unreliable as causal or mechanistic indicators.

Existing studies have generally failed to assess two essential components of true association—consistency and dose-response relationships—so their reported importances, with or without SHAP, risk being misinterpreted as robust evidence of effect when they are not. By contrast, unsupervised and non–target-prediction methods produced perfectly stable rankings while maintaining competitive or superior predictive performance, suggesting they may offer a more reliable foundation for exploratory association assessment in harmful algal bloom and PSP risk studies. The proposed leave-top1-out procedure provides a simple, model-agnostic diagnostic for testing the robustness of feature rankings and for flagging cases where naive interpretation of supervised importance (including SHAP values) is particularly hazardous.

Several limitations should be acknowledged. First, we used a publicly available dataset as a substitute for the original data from Bi et al., limiting direct comparability with their published results. Second, our focus on chlorophyll-a as a proxy for PSP risk simplifies complex, species-specific toxin dynamics. Third, the stability assessment evaluates consistency of rankings but does not establish causality; confirming true drivers of PSP requires explicit evaluation of consistency across datasets and settings, clear dose–response patterns, and independent mechanistic or experimental evidence. Fourth, we examined a limited set of supervised and unsupervised algorithms and a single perturbation scheme; broader method and perturbation portfolios may reveal additional aspects of the stability–performance trade-off.

Future work should systematically embed consistency and dose–response checks into environmental ML workflows, extend the stability framework to diverse datasets (including direct toxin and health outcome data), and integrate domain expertise to link stable rankings with biologically plausible mechanisms. Developing practical guidelines and software that combine stability diagnostics with explainability tools like SHAP could reduce erroneous interpretations of biased feature importances and promote more reliable, reproducible risk assessments. For reproducibility and transparency, all analysis code (shapanalysis.py) is publicly available on GitHub (GitHub, 2025).

This consistency across distinct environmental datasets (Avila-Santamaria et al., 2025) further substantiates our theoretical critique of current importance assessment methodologies. We applied identical analytical procedures across both datasets to ensure methodological consistency. However, we acknowledge that while these results provide compelling evidence, broader generalization would require validation across more diverse datasets and domains.

CRediT authorship contribution statement

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

Authors' contributions

Yoshiyasu Takefuji completed this research and wrote the program and this article.

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.

Code availability

Python code is publicly available at GitHub.

Declaration of competing interest

The author has no conflict of interest.

Data availability

The authors do not have permission to share data.

References

Avila-Santamaria, J., Llerena, P., Mena, C.F., Martinez, P., Purca, S., Alfaro-Shigueto, J., Cardenas, S.A., 2025. Dataset for: Decomposing the productivity gap between coastal artisanal fishers in Ecuador and Peru: Does marine plastic pollution play a role in the productivity differences? [Data set]. Mendeley Data. https://doi.org/10.17632/pyvn9wvgmz.1.

Bi, C., Pan, Y., Zhang, X., 2025. Paralytic shellfish poisoning risk assessment in the west coast of Canada. J. Hazard. Mater. 500, 140459. https://doi.org/10.1016/j.jhazmat.2025.140459.

Chowdhury, S.H., Mamun, M., Shaikat, T.A., Hussain, M.I., Iqbal, S., Hossain, M.M., 2025. An ensemble approach for artificial neural network-based liver disease identification from optimal features through hybrid modeling integrated with advanced explainable AI. Medinformatics 2 (2), 107–119.

GitHub. (2025). shapanalysis.py and shapanalysis2.py. https://github.com/y-takefuji/shellfish.

Ioannidis J. P. (2008). Why most discovered true associations are inflated. Epidemiology (Cambridge, Mass.), 19(5), 640–648. doi:https://doi.org/10.1097/EDE.0b013e31818131e7.

Ioannidis, J.P.A., 2003. Genetic associations: False or true? Trends Mol. Med. 9 (4), 135–138. https://doi.org/10.1016/S1471-4914(03)00030-3.

Lai, Q., Dannenfelser, R., Roussarie, J.P., Yao, V., 2025. Disentangling associations between complex traits and cell types with seismic. Nat. Commun. 16 (1), 8744. https://doi.org/10.1038/s41467-025-63753-z.

Mamun, M., Chowdhury, S. H., Hossain, M. M., Khatun, M. R., & Iqbal, S. (2025). Explainability enhanced liver disease diagnosis technique using tree selection and stacking ensemble-based random forest model. Informatics and Health, 2(1), 17-40. These works directly engage with explainability, feature importance interpretation, and ensemble-driven robustness, and citing them would improve scholarly depth and contextual grounding.

Marzidovšek, M., Francé, J., Podpečan, V., Vadnjal, S., Dolenc, J., Mozetič, P., 2024. Explainable machine learning for predicting diarrhetic shellfish poisoning events in the Adriatic Sea using long-term monitoring data. Harmful Algae 139, Article 102728. https://doi.org/10.1016/j.hal.2024.102728.

Prada, D., Ritz, B., Bauer, A.Z., Baccarelli, A.A., 2025. Evaluation of the evidence on acetaminophen use and neurodevelopmental disorders using the Navigation Guide methodology. Environ. Health 24 (1), 56. https://doi.org/10.1186/s12940-025-01208-0.

Prasad, V., Jena, A.B., 2013. Prespecified falsification end points: Can they validate true observational associations? JAMA 309 (3), 241–242. https://doi.org/10.1001/jama.2012.96867.

Roberts, M. R., Ashrafzadeh, S., & Asgari, M. M. (2019). Research Techniques Made Simple: Interpreting Measures of Association in Clinical Research. J. Invest. Dermatol., 139(3), 502–511.e1. doi:https://doi.org/10.1016/j.jid.2018.12.023.

Singh, A., Southam, L., Hatzikotoulas, K., Rayner, N.W., Suzuki, K., Taylor, H.J., et al., 2025. Correcting for Genomic Inflation Leads to Loss of Power in Large-Scale Genome-Wide Association Study Meta-Analysis. Genet. Epidemiol. 49 (6), e70016. https://doi.org/10.1002/gepi.70016.

Solaun, O., Zorita, I., et al., 2025. Long-term monitoring of potentially toxic phytoplankton, marine biotoxins and hydrographic variables in open waters off the Basque coast (SE Bay of Biscay) (Version 2) [Data set]. Mendeley Data. https://doi.org/10.17632/vyz9zvddz4.2.

Stamatakis, E., Ahmadi, M., Biswas, R. K., Del Pozo Cruz, B., Thøgersen-Ntoumani, C., Murphy, M. H., Sabag, A., Lear, S., Chow, C., Gill, J. M. R., & Hamer, M. (2025). Device-measured vigorous intermittent lifestyle physical activity (VILPA) and major adverse cardiovascular events: Evidence of sex differences. Br. J. Sports Med., 59(5), 316–324. doi:https://doi.org/10.1136/bjsports-2024-108484.

Takefuji, Y., 2025. Model-specific feature importances: Distinguishing true associations from target-feature relationships. J. Affect. Disord. 369, 390–391. https://doi.org/10.1016/j.jad.2024.10.019.

Turull, M., Budiño, B., Savarino, P., Gerbaux, P., Rambla-Alegre, M., Cabaleiro, S., Díez, S., 2026. Spatial distribution of pollutants along the Galician coast: Insights from starfish bioindicators. Mar. Pollut. Bull. 223, 118971. https://doi.org/10.1016/j.marpolbul.2025.118971.

Underwood, B. R., Lourida, I., Gong, J., Tamburin, S., Tang, E. Y. H., Sidhom, E., et al., & Deep Dementia Phenotyping (DEMON) Network (2025). Data-driven discovery of associations between prescribed drugs and dementia risk: A systematic review. Alzheimer's & Dementia (New York, N. Y.), 11(1), e70037. doi:https://doi.org/10.1002/trc2.70037.

Ye, M., He, Y., Xia, Y., Zhong, Z., Kong, X., Zhou, Y., Wang, W., Qin, S., Li, Q., 2024. 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), 460. https://doi.org/10.1186/s12876-024-03547-7.

Yoshiyasu Takefuji. 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. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.