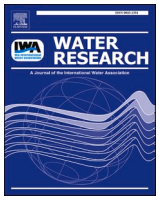
Water Research 288 (2026) 124766

Contents lists available at ScienceDirect

Water Research

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



Limitations of SHAP-based interpretations in environmental and membrane filtration applications

Yoshiyasu Takefuji

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

ARTICLE INFO

Keywords
Interpretable machine learning
SHAP
Microplastic filtration
Feature importance
Environmental modeling

ABSTRACT

Maliwan et al. (2025) identified key parameters in microplastic ultrafiltration using interpretable machine learning (SHAP), attributing 57.6-70.6 % feature importance to factors like transmembrane pressure. This paper critically examines their methodological approach, highlighting significant concerns regarding SHAP's application. SHAP values are inherently model-dependent and lack ground truth for validating feature importance accuracy, leading to potentially biased and erroneous conclusions; high prediction accuracy does not ensure reliable insights. SHAP's underlying assumptions, particularly feature independence, rarely hold in complex environmental systems characterized by multicollinearity, potentially misattributing variable importance. We advocate for a more robust analytical framework incorporating unsupervised machine learning (e.g., feature agglomeration) and nonlinear nonparametric statistical methods (e.g., Spearman's correlation) to provide more reliable insights into variable relationships, moving beyond model-dependent interpretations.

Maliwan et al. (2025) conducted a comprehensive study on the removal and release of microplastics through point-of-use ultrafiltration membranes. The research encompassed a year-long monitoring period and leveraged interpretable machine learning techniques to bolster its findings. Utilizing SHapley Additive exPlanations (SHAP), the authors identified several key filtration parameters—specifically, transmembrane pressure (TMP), filtration volume, permeability, and total resistance. These parameters collectively accounted for 57.6 % to 70.6 % of the feature importance in predicting microplastic (MP) concentration in the membrane permeate. Furthermore, their study proposed a diverse array of predictive models, including multiple linear regression (MLR), Ridge regression, Lasso regression, Bayesian regression, k-nearest neighbors (kNN), support vector machine (SVM), decision tree (DT), random forest (RF), extreme gradient boosting (XGB), and artificial neural networks (ANN) (Maliwan et al., 2025).

This paper strengthens the link to environmental and membrane filtration practice by emphasizing that accurate identification of the most influential features is essential for optimal membrane design and operation, particularly for key parameters such as transmembrane pressure (TMP), filtration volume, permeability, and total resistance. Although membrane filtration parameters were more significant than water quality and feed microplastics in categorical comparison, the establishment of setpoints, fouling control strategies, and performance targets in these applications depends on accurate parameter ranking and interpretation. Unreliable feature rankings may cause designers to incorrectly prioritize transmembrane pressure, volume, permeability, and resistance, thereby compromising optimal operating conditions and design decisions.

The fundamental challenge in feature importance analysis lies in the absence of definitive ground truth, rendering feature importance calculations from supervised models—whether enhanced by SHAP or not—inherently susceptible to bias. Maliwan et al. premise their evaluation on R-squared values as a proxy for prediction accuracy, operating under the assumption that higher predictive performance indicates superior feature selection. However, this approach conflates two distinct analytical objectives: predictive accuracy and feature importance reliability. A model may achieve impressive predictive performance while still misattributing importance to features based on spurious correlations or complex interactions rather than fundamental relationships. This disconnect highlights the critical need to evaluate feature importance methodologies through multiple complementary lenses rather than relying exclusively on predictive performance metrics, particularly when the goal extends beyond prediction to understanding underlying data relationships.

This paper raises significant theoretical and empirical concerns regarding the use of supervised machine learning models alongside SHAP, primarily due to the model-specific nature of SHAP interpretations, which can lead to erroneous conclusions. While

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.watres.2025.124766
Received 20 July 2025; Received in revised form 4 October 2025; Accepted 10 October 2025
Available online 11 October 2025

supervised machine learning models provide ground truth values that enable validation of target prediction accuracy, the feature importances derived from these models lack corresponding ground truth for their own accuracy validation. This absence of ground truth complicates the determination of true associations between variables, resulting in different models yielding distinct feature importances. Consequently, this can lead to biased feature importances. Importantly, high target prediction accuracy does not necessarily correlate with reliable feature importances, as the feature importances generated by models are inherently biased and can result in skewed interpretations (Fisher et al., 2019; Nazer et al., 2023; Ugirumurera et al., 2024; Alaimo Di Loro et al., 2023; Adler and Painsky, 2022; Steiner and Kim, 2016). Over 300 peer-reviewed articles have documented this non-negligible bias in feature importances derived from machine learning models.

While this paper acknowledges that SHAP is a powerful explanatory tool with robust methodologies that faithfully capture nonmonotonic relationships between variables within a given model, this paper specifically addresses a critical limitation: SHAP must ultimately rely on feature importances derived from supervised machine learning models, which themselves can be unreliable. It's important to distinguish between a model's two types of accuracy: target prediction accuracy and feature importance reliability. The former can be validated against known labels, while the latter lacks ground truth for validation. This paper does not dismiss SHAP's utility and mathematical strengths, but rather cautions against exclusively relying on it when the underlying model may prioritize prediction accuracy over accurate feature importance representation. A more balanced approach would combine SHAP's explanatory power with additional validation methods to ensure feature importances reflect true causal relationships in the data.

The implementation of `explain=SHAP(model)` represents a fundamental methodological limitation, as SHAP values are inherently model-dependent interpretations rather than objective measures of variable relationships. This dependency dictates that SHAP exclusively relies on the given model's architecture and assumptions, inevitably inheriting and potentially amplifying any biases already present in the model's feature importance calculations (Wu, 2025; Bilodeau et al., 2024; Huang and Marques-Silva, 2024; Kumar et al., 2021; Hooshyar and Yang, 2024; Lones, 2024; Molnar et al., 2022; Létoffé et al., 2025). Importantly, SHAP's mathematical underpinnings, founded on cooperative game theory principles, assume that features operate independently—a critical prerequisite rarely met in complex environmental systems where multicollinearity is common. This can lead to misleading attribution of importance to correlated variables. Furthermore, the absence of a reliable mechanism for accurately calculating true associations between variables poses a significant methodological challenge, preventing definitive verification of whether SHAP-derived insights reflect genuine physical processes or are merely statistical artifacts.

The expression explain = SHAP(model) indicates that SHAP operates by analyzing a specific trained model rather than functioning independently. SHAP generates explanations by calculating feature importance values derived directly from the model's structure and predictions. This makes SHAP fundamentally model-dependent, as its explanations reflect the particular relationships and patterns captured by the given model rather than operating from external assumptions or predefined feature relationships. The explanations therefore inherit both the strengths and limitations of the underlying model they interpret

The field currently lacks definitive methodologies for determining true causal relationships between variables, with existing supervised models primarily quantifying predictive contribution rather than uncovering genuine causal mechanisms. This paper proposes integrating unsupervised modeling approaches alongside traditional methods, leveraging their potential to reduce the inherent bias of label-driven analysis by identifying patterns without the constraint of predefined outcomes. This hybrid approach offers a more comprehensive analytical framework while preserving the integrity of established techniques like SHAP. Importantly, the unsupervised components serve as complementary tools rather than replacements, enriching rather than disrupting existing interpretability frameworks. However, substantial empirical validation remains necessary to establish the reliability, interpretability, and comparative advantages of feature importance measures across both supervised and unsupervised paradigms, particularly for identifying meaningful relationships in complex datasets.

To overcome these significant limitations, this paper advocates for a more comprehensive analytical framework. This framework incorporates unsupervised machine learning approaches such as feature agglomeration (FA) (to identify natural groupings of related variables) and highly variable gene selection techniques (HVGS) (adapted from genomics to identify truly influential parameters). These should be complemented by nonlinear nonparametric statistical methods like Spearman's correlation, which can capture complex monotonic relationships with robust p-value assessments, without imposing distributional assumptions. This multifaceted approach would provide more robust evidence of variable relationships than relying solely on SHAP (model) interpretations, especially in complex environmental systems where mechanistic understanding remains paramount.

Stability in feature ranking is critical for establishing reliable feature importance analysis across diverse analytical contexts. When evaluating this stability through progressive removal of top-ranked features among full features, supervised models frequently demonstrate inconsistent ranking patterns—a vulnerability stemming from their model-specific architectures and their focus on predictive contribution rather than capturing fundamental relationships within data. In contrast, unsupervised approaches such as FA, HVGS, and Spearman correlation demonstrate remarkable ranking stability across iterative feature removal scenarios. This stability advantage emerges from their focus on inherent data structures rather than optimizing for specific prediction tasks, suggesting these methods may more faithfully reflect genuine underlying variable associations rather than model-specific predictive utility. This distinction becomes particularly valuable when seeking robust feature importance metrics that remain consistent despite changes in feature composition.

This paper proposes a novel pipeline framework that integrates unsupervised models with non-target supervised Spearman's correlation to enhance SHAP approaches. The complete Python implementation is publicly accessible at GitHub (GitHub, 2025a, 2025b). The code repository provides robust solutions for feature importance analysis of omics data from The Cancer Genome Atlas (TCGA) (GitHub, 2025a) comprising 705 samples and 1936 features, and for MNIST data with 70,000 samples and 784 features (GitHub, 2025b). These datasets are comprehensively analyzed through diverse methodologies—supervised, unsupervised, and non-target supervised models—while ensuring the proposed methods complement rather than interfere with established SHAP approaches.Feature importance calculations should be performed on raw data rather than scaled, normalized, or transformed values to prevent artifactual results that misrepresent natural relationships in the data. While techniques like hyperparameter tuning effectively enhance predictive accuracy, they may inadvertently distort feature importance interpretations by optimizing for prediction at the expense of interpretability. Notably, unsupervised approaches and non-target supervised statistical methods such as Spearman's correlation offer computational efficiency advantages, as they directly measure relationships within the data without requiring the intensive training procedures associated with complex predictive models. This balance between interpretability and computational demands is particularly valuable when feature importance, rather than prediction, is the primary analytical goal.

Feature importance calculations should be performed on raw data rather than scaled, normalized, or transformed values to prevent artifactual results that misrepresent natural relationships in the data. Spearman's rank correlation effectively handles outliers while preserving the natural ordering of raw data values through ranking - a process distinct from scaling or normalization as it maintains relative

relationships without altering the data's distribution characteristics. Similarly, feature agglomeration can be implemented without prior scaling or transformation by using the fit method rather than fit_transform, allowing clusters to form based on inherent data structures. Both methods can thus respect the original data's relationships while mitigating the distortions that arbitrary scaling or normalization might introduce to feature importance calculations.

While techniques like hyperparameter tuning effectively enhance predictive accuracy, they may inadvertently distort feature importance interpretations by optimizing for prediction at the expense of interpretability. Notably, unsupervised approaches and non-target supervised statistical methods such as Spearman's correlation offer computational efficiency advantages, as they directly measure relationships within the data without requiring the intensive training procedures associated with complex predictive models. This balance between interpretability and computational demands is particularly valuable when feature importance, rather than prediction, is the primary analytical goal.

Superior predictive performance does not necessarily translate to reliable feature importance attribution, highlighting a critical disconnect in current interpretability approaches. This paper addresses this gap by proposing a complementary framework that enhances traditional SHAP-based and supervised model interpretations with unsupervised analytical techniques. While the unsupervised modeling landscape is diverse—offering numerous methodological options with distinct theoretical foundations and practical implications—this work establishes a foundation for their integration into feature importance analysis. Future research should systematically evaluate these varied unsupervised approaches across different data domains and complexity levels to determine their relative effectiveness in feature selection and importance attribution. This evaluation would ideally include stability analysis, cross-validation across diverse datasets, and comparison with ground truth in controlled scenarios where such truth is available, ultimately advancing toward more robust and trustworthy feature importance frameworks.

Funding

This research has no fund.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Code availability

Not applicable.

AI use

Not applicable.

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 7,222 in parallel algorithms. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original draft, Validation, Investigation, 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.

Data availability

No data was used for the research described in the article.

References

Adler, A.I., Painsky, A., 2022. Feature importance in gradient boosting trees with cross-validation feature selection. Entropy 24 (5), 687. https://doi.org/10.3390/e24050687.

Alaimo Di Loro, P., Scacciatelli, D., Tagliaferri, G., 2023. 2-step gradient boosting approach to selectivity bias correction in tax audit: An application to the VAT gap in Italy. Stat. Methods Appl. 32, 237–270. https://doi.org/10.1007/s10260-022-00643-4.

Bilodeau, B., Jaques, N., Koh, P.W., Kim, B., 2024. Impossibility theorems for feature attribution. Proc. Natl. Acad. Sci. 121 (2), e2304406120. https://doi.org/10.1073/pnas.2304406120.

Fisher, A., Rudin, C., Dominici, F., 2019. 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, 177.

GitHub. (2025a) stability.py. https://github.com/y-takefuji/cell.

GitHub. (2025b). mniststability.py. https://github.com/y-takefuji/mnist.

Hooshyar, D., Yang, Y., 2024. 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, 137472–137490. https://doi.org/10.1109/ACCESS.2024.3463948.

Huang, X., Marques-Silva, J., 2024. On the failings of Shapley values for explainability. Int. J. Approx. Reason. 171, 109112. https://doi.org/10.1016/j.ijar.2023.109112.

Kumar, I., Scheidegger, C., Venkatasubramanian, S., Friedler, S., 2021. Shapley residuals: Quantifying the limits of the Shapley value for explanations. Adv. Neural Inf. Process. Syst. 34, 26598–26608.

Létoffé, O., Huang, X., Marques-Silva, J., 2025. Towards trustable SHAP scores. In: Proceedings of the AAAI Conference on Artificial Intelligence, 39, pp. 18198–18208. https://doi.org/10.1609/aaai.v39i17.34002.

Lones, M.A., 2024. Avoiding common machine learning pitfalls. Patterns 5 (10), 101046. https://doi.org/10.1016/j.patter.2024.101046.

Maliwan, T., Zhang, T., Yeo, M.M.E., Lohwacharin, J., Hu, J., 2025. Revisiting microplastic removal and release by point-of-use ultrafiltration membranes: 1-year monitoring and interpretable machine learning. Water Res. 285, 124053. https://doi.org/10.1016/j.watres.2025.124053.

Molnar, C., et al., 2022. General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K. R., Samek, W. (Eds.), xxAI - Beyond Explainable AI, xxAI - Beyond Explainable AI, 13200. Springer, p. 4. https://doi.org/10.1007/978-3-031-04083-2_4.

Nazer, L.H., Zatarah, R., Waldrip, S., Ke, J.X.C., Moukheiber, M., Khanna, A.K., Choi, S. W., Holder, A.L., Churpek, M.M., 2023. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit. Health 2 (6), e0000278. https://doi.org/10.1371/journal.pdig.0000278.

Steiner, P.M., Kim, Y., 2016. The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. J. Causal Inference 4 (2), 20160009. https://doi.org/10.1515/jci-2016-0009.

Ugirumurera, J., Bensen, E.A., Severino, J., Sanyal, J., 2024. Addressing bias in bagging and boosting regression models. Sci. Rep. 14 (1), 18452. https://doi.org/10.1038/s41598-024-68907-5.

Wu, L., 2025. A review of the transition from Shapley values and SHAP values to RGE. Statistics 1–23. https://doi.org/10.1080/02331888.2025.2487853.