Journal of Environmental Chemical Engineering 13 (2025) 120094

Contents lists available at ScienceDirect

Journal of Environmental Chemical Engineering

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

Beyond XGBoost-SHAP: Strengthening THM mechanistic inference with consistency and dose-response validation

ARTICLE INFO

Keywords:
THMs
XGBoost
SHAP
Consistency
Dose–response

ABSTRACT

Tao et al. demonstrate an XGBoost–SHAP framework that predicts trihalomethanes (THMs) with high accuracy ($R^2 = 0.9506$; RMSE = 0.5340 μg/L) and elevates CODMn, TDS, and FCR as dominant drivers. While powerful for real-time monitoring, supervised feature attributions can be fragile—susceptible to collinearity, proxies, distributional shifts, and confounding—yielding overconfident mechanistic claims despite strong target-prediction accuracy. We argue that trustworthy inference requires independent validation emphasizing two pillars: consistency across studies, settings, and populations, and clear dose–response relationships. We outline a complementary workflow: unsupervised exploration (feature agglomeration, HVGS), nonparametric associations paired with explicit dose–response tests, and systematic perturbation with resampling and cross-site/temporal validation. Integrating these steps with supervised modeling can stabilize interpretations and advance reliable, actionable insights on THM formation.

Tao et al. introduced an XGBoost–SHAP framework that accurately predicts trihalomethanes (THMs) in drinking water while illuminating a three-stage regulatory mechanism [1]. Their XGBoost model delivered strong predictive performance ($R^2 = 0.9506$; RMSE = 0.5340 μg/L), and SHAP analyses highlighted the permanganate index (CODMn), total dissolved solids (TDS), and free chlorine residual (FCR) as the dominant drivers of THM formation. This work exemplifies how combining gradient-boosted trees with explainable AI can enhance real-time monitoring and guide targeted mitigation strategies in distribution systems [1].

However, this paper underscores the pitfalls of leaning solely on XGBoost–SHAP feature importances from a supervised model. Without independent validation, these attributions can be skewed by collinearity, proxy variables, distributional shifts, or hidden confounders, leading to overconfident or incorrect mechanistic claims. Put plainly, feature importances derived from supervised models are inherently fragile and prone to misinterpretation when taken at face value [2–10]. Over 300 peer-reviewed articles have documented non-negligible biases in feature importances derived from supervised models. The Journal of Environmental Chemical Engineering has already published ninety-five SHAP-related articles (seventy-nine in 2025 alone), reflecting rapid adoption of model interpretation tools and the urgent need for stronger validation protocols.

Supervised learning produces two distinct forms of accuracy: target-prediction accuracy, which can be directly benchmarked against observed labels (e.g., R^2 and RMSE), and feature-importance accuracy, for which no ground-truth reference exists. As a result, different algorithms or even repeated runs of the same algorithm under minor data perturbations can yield inconsistent importance rankings despite comparable predictive performance. In short, high target-prediction accuracy does not guarantee trustworthy feature importances [11–18]. Because SHAP explains the trained model as-is, it will faithfully propagate any biases or instabilities already present, potentially misleading researchers if explanations are interpreted as causal or mechanistic evidence [19–28].

To assess true associations, two key elements must be met: consistency and dose-response relationships [29–38]. Consistency means that an association is replicated across different studies, settings, and populations. Dose-response means systematic changes in the outcome with varying levels of exposure. To reduce related risks, we recommend a multifaceted workflow that goes beyond supervised feature attributions. First, unsupervised exploratory analyses such as feature agglomeration and highly variable gene selection (HVGS) can reveal stable clusters or latent structures without referencing the target, supporting consistency across subsets, resamples, sites, and time periods. Second, non-targeted, nonlinear nonparametric association measures (e.g., Spearman's rank correlation with p-values) should be paired with explicit dose-response assessments (e.g., monotonic trend tests, spline-based partial dependence, and stratified analyses) to confirm that relationships are consistent across strata and exhibit plausible monotonic or threshold-like dose-response behavior. Third, systematic perturbation and leave-one-out validation should be routine: remove the top-ranked feature, re-rank the remaining predictors, and assess the stability of their ordering to gauge robustness, complemented by bootstrapped resampling and cross-site or temporal validation to test consistency. By integrating these complementary strategies with supervised modeling, studies of THM formation can produce more reliable, consistent insights and avoid artifacts introduced by the learning algorithm while substantiating true dose-response relationships.

CRediT authorship contribution statement

takefuji yoshiyasu: Conceptualization, Investigation, Validation, Writing – original draft, Writing – review & editing.

https://doi.org/10.1016/j.jece.2025.120094
Received 30 October 2025; Accepted 1 November 2025
Available online 10 November 2025
2213-3437/© 2025 Elsevier Ltd. All rights reserved, including those for text and data mining, AI training, and similar technologies.

Ethics approval

Not applicable

Authors' contributions

Yoshiyasu Takefuji completed this research and wrote this article.

Declaration of Generative AI and AI-assisted technologies in the writing process

Not applicable

Consent to participate

Not applicable

Consent for publication

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 7222 in parallel algorithms. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

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.

Data availability

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

References

[1] Y. Tao, G. Fan, F. Liu, J. Luo, R. Zou, Y. Huang, Y. Cao, J. Long, K.-Q. Xu, Precision prediction of trihalomethanes in drinking water and three-stage regulation mechanism: a study based on XGBoost-SHAP framework, J. Environ. Chem. Eng. 13 (6) (2025) 119316, https://doi.org/10.1016/j.jece.2025.119316.

[2] 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.

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

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

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

[6] A.I. Adler, A. Painsky, Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection, Entropy 24 (5) (2022) 687, https://doi.org/10.3390/e24050687.

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

[8] 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. Method. 23 (1) (2023) 144, https://doi.org/10.1186/s12874-023-01965-x.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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. Bioinforma. 25 (5) (2024) bbae379, https://doi.org/10.1093/bib/bbae379.

[15] 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.

[16] J.L. Potharlanka, Bhat M N. 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.

[17] 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.

[18] Molnar C., König G., Herbinger J., et al. 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: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers. Cham: Springer International Publishing; 2022:39-68. doi:10.1007/978-3-031-04083-2_4.

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] O. Létoffé, X. Huang, J. Marques-SilvaTowards trustable SHAP Scores 17 39 Proc. AAAI Conf. Artif. Intell.2025, 181981820810.1609/aaai.v.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] E. Stamatakis, M. Ahmadi, R.K. Biswas, B. Del Pozo 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.
[36] 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.
[37] 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, Alzheimers Dement 11 (1) (2025) e70037, https://doi.org/10.1002/trc2.70037.
[38] Y. TakefujiModel-specific feature importances: distinguishing true associations from target-feature relationships 369 J. Affect Disord.2025, 39039110.1016/j.jad.2024.10.019.

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