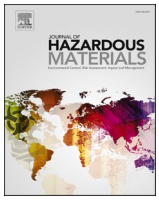
Journal of Hazardous Materials 493 (2025) 138366

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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



Letter to the Editor

Comments on "Dialogue between algorithms and soil: Machine learning unravels the mystery of phthalates pollution in soil" by Pan et al. (2025)

HIGHLIGHTS

- MLP excels in PAE pollution prediction despite model biases.
- ML models (XGBoost, MLP, SVR) have inherent feature biases.
- SHAP inherits and may amplify model-derived feature biases.
- Prediction accuracy varies from feature importance reliability.
- Robust statistics enhance feature reliability.

ARTICLE INFO

Keywords:
Phthalates pollution
Feature selection
Machine learning
SHapley Additive exPlanations
Biases

ABSTRACT

Pan et al. demonstrated the superior predictive performance of their machine learning ML models for soil phthalate PAE concentrations, highlighting the critical role of feature importance as assessed by SHapley Additive exPlanations (SHAP). Notably, the Multilayer Perceptron (MLP) model achieved the highest performance (R^2 = 0.8637), followed by SVR and XGBoost. However, concerns persist regarding the reliability of feature importance derived from these models and their SHAP interpretations. Specifically, predictive accuracy does not guarantee the validity of feature rankings due to the inherent biases present in tree-based, neural network, and kernel-based methods, which are further exacerbated by SHAP's inherent dependency on model outputs. To mitigate these biases, integrating robust statistical methods is crucial. Techniques such as Spearman's rho, Kendall's tau, Goodman-Kruskal's gamma, Somers' delta, and Hoeffding's dependence, combined with p-value analysis, offer unbiased assessments. Integrating these statistical methods alongside ML models ensures a more reliable evaluation of feature importance in environmental risk modeling. Consequently, future research should prioritize methodologies that combine ML with rigorous statistical validation to enhance accuracy and reduce biases.

1. Text

Letter to the Editor:

Pan et al. [16] conducted a study, "Dialogue between algorithms and soil: Machine learning unravels the mystery of phthalates pollution in soil," which presents critical points warranting further discussion. Their study aimed to predict the concentrations of phthalates (PAEs) in soil using various machine learning (ML) models, including Random Forest Regression (RFR), Gradient Boosting Regression Tree (GBRT), Extreme Gradient Boosting (XGBoost), Multilayer Perceptron (MLP), Support Vector Regression (SVR), and k-Nearest Neighbors (KNN). Their study found that the MLP model exhibited optimal performance (R^2 of 0.8637), followed by SVR (R^2 of 0.8132) and XGBoost (R^2 of 0.8096). The ranking of feature importance elements was assessed using SHapley Additive exPlanations (SHAP), providing a comprehensive interpretation of the factors influencing PAEs distribution. However, the distinct feature importances generated by different models like XGBoost, MLP, and SVR through SHAP suggest that their methodologies may be biased, which is not a trivial issue.

While this letter acknowledges a significant contribution to the field of PAEs assessment by Pan et al. [16], it raises critical concerns regarding the interpretation of feature importances derived from three machine learning models (XGBoost, MLP, SVR) and SHAP. They assessed predictive accuracy considering several metrics such as R^2, MSE, and MAE. However, it is crucial to recognize that predictive accuracy and feature importance are fundamentally distinct concepts, and high predictive accuracy does not inherently guarantee the reliability of feature importance. Pan et al.'s research aims to explore feature-PAE concentration relationships via feature importance, yet the high prediction accuracy demonstrated at the outset may imply reliability in the subsequent SHAP interpretation. While many similar papers follow this pattern, the main point of this letter is that high prediction accuracy does not guarantee the reliability of feature importance interpretation. This pitfall has already been pointed out in over 100 peer-reviewed papers, with Lipton's article serving as an essential introduction to the fundamental issues [8,9,11,17,21]. A detailed discussion and supporting references are provided in the supplementary material.

XGBoost, like other tree-based models such as RFR and GBDT,

https://doi.org/10.1016/j.jhazmat.2025.138366
Received 12 March 2025; Received in revised form 15 April 2025; Accepted 20 April 2025
Available online 22 April 2025

exhibits inherent biases in feature importance calculations due to its tree building process, which can overemphasize the importance of features used in earlier splits [1,2,19]. This could lead to a skewed perception of environmental factors' importance. MLP deep learning models also intrinsically possess non-negligible biases due to their complex architectures and the way they learn feature representations [5]. Additionally, SVR can exhibit biases due to its reliance on kernel methods, which transform the input space in ways that can obscure the true importance of original features [4].

Furthermore, Pan et al. claim that the divergence in feature importance rankings across XGBoost, MLP, and SVR enhances the value of the analysis, which warrants careful interpretation. In Section 3.3, 'Identification of the important feature, ' they state: "The variations among these models not only reflected the complexity of PAEs behavior in soil but also highlighted each algorithm's unique advantages in processing non-linear relationships and feature interactions. Although different key features were identified by each model, the significant influences of meteorological conditions, soil characteristics, and anthropogenic factors on PAEs distribution were collectively indicated, wherein the complexity of PAEs behavior in the environment was reflected." Their assertion that these inconsistencies are 'unique advantages' often raises further concerns. Indeed, such model discrepancies are not always epistemologically problematic; when multiple valid relationships or interactions exist within the data, different models capturing different aspects can contribute to a more comprehensive understanding. However, in this particular study, where the rankings of key features differ significantly and the reasons for these differences are not clearly demonstrated, these inconsistencies complicate the determination of truly important features. Especially when high predictive accuracy is observed alongside significant variability in feature importance, it may simply indicate that the models are making predictions based on different mechanisms, thus necessitating cautious interpretation.

Additionally, SHAP values, while intended to elucidate feature importance, inherit and may even exacerbate biases from the underlying machine learning model [10,3,6,7]. The function of 'explain = SHAP (model)' underscores this dependency. As SHAP relies on the model's output for its explanations, its inherent vulnerability to model biases demands careful interpretation of the results and assessment of the analysis's reliability. Several feature analysis methods exist besides SHAP, offering the advantage of lower computational cost. For example, Permutation Importance evaluates feature importance by randomly shuffling the values of each feature and measuring the impact on the model's performance. Local Interpretable Model-agnostic Explanations (LIME) generates local surrogate models to explain individual predictions, providing insights into the model's local behavior. Integrated Gradients is designed for deep learning models, attributing the model's predictions to input features by calculating gradients along the path from a baseline input to the actual input. Despite the existence of these simpler alternative methods, SHAP is widely utilized as one of the most consistent frameworks due to its comprehensive game-theoretic principles and versatility. However, it is important to recognize that all these methods, including SHAP, are susceptible to model bias, necessitating careful application to ensure reliable interpretations. While there are bias mitigation methods, they cannot completely eliminate biases in feature importances derived from ML models.

The crux of the matter is that validating feature importance is exceptionally challenging due to the absence of ground truth values, as different models employ distinct methodologies, inevitably leading to model-specific biases and varying rankings [13]. This issue is particularly noticeable in Pan et al.'s study, given their complex feature sets. High dimensionality and potential collinearity significantly impede the interpretation of machine learning models, especially in the context of PAEs assessment [18].

These factors not only complicate the isolation of individual feature effects but also diminish the perceived importance of predictive variables. This complexity extends beyond theoretical considerations, manifesting in practical applications where researchers may find it challenging to identify the true determinants of PAEs assessment, potentially leading to conclusions about underlying mechanisms that should be interpreted with caution. Furthermore, high dimensionality increases the risk of overfitting, causing models to capture noise rather than genuine signals and to emphasize spurious features in importance measures. This is particularly problematic in environmental research, which inherently involves noisy and complex datasets, requiring robust models that can distinguish between meaningful signals and random fluctuations. Additionally, the complexity of features renders importance measures highly sensitive to minor changes in data or model configurations, which may affect their stability and reliability. This instability can complicate the interpretation of model-derived insights and hinder the development of consistent and reproducible research findings.

Addressing these limitations demands attention to three key areas: the nature of data distribution, the statistical relationships between variables, and the statistical validation. Understanding data distribution is essential for choosing effective modeling strategies. Investigating complex relationships, particularly through non-parametric methods, is vital. Furthermore, validating findings statistically, using hypothesis testing and p-value analysis, guarantees that results are not coincidental. These three aspects are comprehensively addressed by robust statistical methods. Instead of relying solely on machine learning models and SHAP for feature selection, we advocate for the integration of unbiased, robust statistical methods, such as Spearman's rho and Kendall's tau, coupled with p-values [15,20]. These are particularly well-suited for assessing monotonic relationships. Other suitable non-parametric methods include Goodman-Kruskal's gamma, Somers' delta, and Hoeffding's dependence, effective for complex relationships like non-monotonic collinearity and interactions [12,14]. By prioritizing these statistical principles, researchers can enhance the reliability and validity of feature importance assessments in environmental risk modeling.

In conclusion, while machine learning techniques such as XGBoost, MLP, SVR and SHAP are powerful for feature selection, they possess inherent biases. In complex domains like environmental risk assessment, integrating robust statistical methods and rigorous validation is essential to complement machine learning's limitations. This integrated approach is crucial for achieving accurate and reliable insights. Future research should prioritize exploring innovative methodologies that combine the strengths of machine learning and statistical analysis.

CRediT authorship contribution statement

Oka Souichi: Writing – original draft, Investigation, Conceptualization. **Takefuji Yoshiyasu:** Writing – review & editing, Supervision, Project administration.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2025.138366.

Data availability

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

References

[1] 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.

[2] 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.

[3] Bilodeau, B., et al., 2024. Impossibility theorems for feature attribution. Proc Natl Acad Sci USA 121 (2), e2304406120. https://doi.org/10.1073/pnas.2304406120.

[4] Faragalli, A., et al., 2025. Do machine learning methods solve the main pitfall of linear regression in dental age estimation? Forensic Sci Int 367, 112353. https://doi.org/10.1016/j.forsciint.2024.112353.

[5] 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. https://doi.org/10.48550/arXiv.1801.01489.

[6] 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.

[7] Kumar, I., et al., 2021. Shapley residuals: quantifying the limits of the shapley value for explanations. Adv Neural Inf Process Syst 34, 26598–26608.

[8] Lenhof, K., et al., 2024. 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), bbae379. https://doi.org/10.1093/bib/bbae379.

[9] Lipton, Z.C., 2018. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. ACM Queue 16 (3), 31–57. https://doi.org/10.1145/3236386.3241340.

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

[11] Mandler, H., Weigand, B., 2024. A review and benchmark of feature importance methods for neural networks. ACM Comput Surv 56 (12), 318. https://doi.org/10.1145/3679012.

[12] Metsämuuronen, J., 2021. Directional nature of Goodman–Kruskal gamma and some consequences: identity of Goodman–Kruskal gamma and Somers delta, and their connection to Jonckheere–Terpstra test statistic. Behaviormetrika 48 (2), 283–307. https://doi.org/10.1007/s41237-021-00138-8.

[13] Musolf, A.M., et al., 2022. What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. Hum Genet 141 (9), 1515–1528. https://doi.org/10.1007/s00439-021-02402-z.

[14] Newson, R., 2006. Confidence intervals for rank statistics: Somers' D and extensions. Stata J 6, 309–334. https://doi.org/10.1177/1536867X0600600302.

[15] Okoye, K., Hosseini, S., 2024. Correlation tests in R: Pearson Cor, Kendall's Tau, and Spearman's Rho. In: Okoye, K., Hosseini, S. (Eds.), R programming: statistical data analysis in research. Springer Nature, pp. 247–277. https://doi.org/10.1007/978-981-97-3385-9_12.

[16] Pan, B., et al., 2025. Dialogue between algorithms and soil: machine learning unravels the mystery of phthalates pollution in soil. J Hazard Mater 482, 136604. https://doi.org/10.1016/j.jhazmat.2024.136604.

[17] Potharlanka, J.L., Bhat, M.N., 2024. Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms. Sci Rep 14 (1), 2923. https://doi.org/10.1038/s41598-024-53141-w.

[18] Touw, W.G., et al., 2013. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Brief Bioinforma 14 (3), 315–326. https://doi.org/10.1093/bib/bbs034.

[19] Ugirumurera, J., et al., 2024. Addressing bias in bagging and boosting regression models. Sci Rep 14 (1), 18452. https://doi.org/10.1038/s41598-024-68907-5.

[20] Yu, H., Hutson, A.D., 2024. A robust Spearman correlation coefficient permutation test. Commun Stat - Theory Methods 53 (6), 2141–2153. https://doi.org/10.1080/03610926.2022.2121144.

[21] Wood, D., et al., 2024. Model-agnostic variable importance for predictive uncertainty: an entropy-based approach. Data Min Knowl Discov 38, 4184–4216. https://doi.org/10.1007/s10618-024-01070-7.

Souichi Oka *

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

Yoshiyasu Takefuji

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

E-mail address: takefuji@keio.jp.

* Corresponding author.

E-mail address: souichi.oka@sciencepark.co.jp (S. Oka).