Science of the Total Environment 984 (2025) 179714

Contents lists available at ScienceDirect

Science of the Total Environment

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

Letter to the Editor

Letter to the Editor regarding "Prediction of PFAS bioaccumulation in different plant tissues with machine learning models based on molecular fingerprints" by Song et al. (2024), Sci. Total Environ. 950 175091

HIGHLIGHTS

- Machine learning predicts PFAS plant uptake.
- XGBoost/SHAP modeling shows predictive capability.
- Feature importance analysis critically evaluated.
- XGBoost & SHAP biases impact interpretations.
- Robust feature selection: non-parametric methods recommended

ARTICLE INFO

Editor: Wei Ouyang
Keywords:
PFASs
Plant uptake
Machine learning
Feature importance
Biases
Non-parametric statistics

ABSTRACT

Song et al. (2024), "Prediction of PFAS bioaccumulation in different plant tissues with machine learning models based on molecular fingerprints," employed machine learning methods, such as XGBoost and SHapley Additive exPlanations (SHAP), to predict PFAS bioaccumulation, reporting high predictive accuracy. However, this commentary critically examines their interpretation of feature importance, since high predictive accuracy does not guarantee reliable feature importance. Both XGBoost and SHAP are known to exhibit biases, such as overemphasizing features used in early splits and inheriting biases from the underlying model. Furthermore, the high dimensionality and potential collinearity of molecular fingerprints complicate SHAP interpretation, increasing overfitting risk and compromising SHAP value stability. To provide a general example, we conducted an independent simulation using a publicly available dataset of US industrial facilities and environmental compliance, demonstrating significant discrepancies between feature importance rankings from XGBoost and robust statistical tests. This commentary advocates for robust statistical methods coupled with *p*-values, including Spearman's rho, Kendall's tau, Goodman-Kruskal's gamma, Somers' delta, and Hoeffding's dependence, for feature selection. These non-parametric methods, which are independent of specific model assumptions and rely on data ranks, are better suited to capture complex relationships in high-dimensional data, providing a more reliable foundation for future PFAS bioaccumulation research.

1. Introduction

The recent publication by Song et al., "Prediction of PFAS bioaccumulation in different plant tissues with machine learning models based on molecular fingerprints," presents several critical issues that necessitate further discussion (Song et al., 2024). Their objective was to develop a machine learning model for predicting the bioaccumulation factors (BAFs) of *per-* and polyfluoroalkyl substances (PFASs) in various plant tissues, including roots, stems, leaves, and fruits. They demonstrated the key influential features affecting model predictions using the Extreme Gradient Boosting (XGB) model and SHapley Additive exPlanations (SHAP). This analysis revealed that the key influential features varied among different plant tissues. For instance, soil organic matter (OM) was the most significant factor for roots and stems. In contrast, ECFP-347 was the most influential for leaves and fruits. Their model demonstrated strong performance, with coefficients of determination (R^2) ranging from 0.82 to 0.93.

While Song et al. (2024) have made a significant contribution to the field of PFAS risk assessment, this paper raises critical concerns regarding the interpretation of feature importances derived from XGB and SHAP. Although they achieved high R^2 values, it is crucial to distinguish between predictive accuracy and the reliability of feature importance. As discussed extensively in over 100 peer-reviewed articles, high predictive accuracy does not guarantee the validity of feature importance rankings (Fisher et al., 2019). A detailed discussion and supporting references are provided in the supplementary material.

2. Limitations of XGB

XGB, like other tree-based models, exhibits inherent biases in feature importance calculations due to its tree-building process, which can overemphasize the importance of features used in earlier splits (Adler

https://doi.org/10.1016/j.scitotenv.2025.179714
Received 28 February 2025; Received in revised form 27 March 2025; Accepted 18 May 2025

Table 1
Feature importance rankings from XGB and statistical tests for penalized factories. The colored cells within the table identify key variables pertaining to a facility's environmental compliance status: FAC_COMPLIANCE_STATUS (the compliance status of the facility with environmental regulations), FAC_PROGRAMS_WITH_SNC (the number of environmental programs with significant non-compliance at the facility), and FAC_SNC_FLG (indicating whether the facility has significant non-compliance issues).

Rank	XGBoost		Spearman's Rho			Kendall's Tau		
	Variable	Importances	Variable	Coefficient	p-Value	Variable	Coefficient	p-Value
1	FAC_QTRS_WITH_NC	0.295	FAC_QTRS_WITH_NC	0.363	0.000	FAC_QTRS_WITH_NC	0.351	0.000
2	TRI_FLAG	0.125	TRI_FLAG	0.288	0.000	TRI_FLAG	0.288	0.000
3	AIR_FLAG	0.080	GHG_FLAG	0.279	0.000	GHG_FLAG	0.279	0.000
4	GHG_FLAG	0.073	FAC_COMPLIANCE_STATUS	0.264	0.000	FAC_PROGRAMS_WITH_SNC	0.264	0.000
5	Status	0.060	FAC_PROGRAMS_WITH_SNC	0.264	0.000	FAC_SNC_FLG	0.263	0.000
6	NPDES_FLAG	0.037	FAC_SNC_FLG	0.263	0.000	FAC_COMPLIANCE_STATUS	0.260	0.000
7	RCRA_FLAG	0.037	CWA	0.178	0.000	CWA	0.178	0.000
8	State	0.036	NPDES_FLAG	0.178	0.000	NPDES_FLAG	0.178	0.000
9	Federal Facility	0.036	SDWIS_FLAG	0.138	0.000	SDWIS_FLAG	0.138	0.000
10	Region	0.033	RCRA	0.118	0.000	RCRA	0.118	0.000
11	Industry	0.028	RCRA_FLAG	0.118	0.000	RCRA_FLAG	0.118	0.000
12	FAC_COMPLIANCE_STATUS	0.027	Status	−0.111	0.000	Status	−0.111	0.000
13	FAC_PROGRAMS_WITH_SNC	0.027	CAA	0.101	0.000	CAA	0.101	0.000
14	FAC_POP_DEN	0.024	AIR_FLAG	0.101	0.000	AIR_FLAG	0.101	0.000
15	FAC_PERCENT_MINORITY	0.023	Region	−0.062	0.000	Region	−0.054	0.000
16	EJSCREEN_FLAG_US	0.023	State	0.056	0.000	State	0.047	0.000
17	SDWIS_FLAG	0.019	Federal Facility	0.030	0.000	Federal Facility	0.030	0.000
18	RCRA	0.015	FAC_PERCENT_MINORITY	0.029	0.000	FAC_PERCENT_MINORITY	0.025	0.000
19	FAC_SNC_FLG	0.000	EJSCREEN_FLAG_US	0.016	0.000	EJSCREEN_FLAG_US	0.016	0.000
20	CAA	0.000	FAC_POP_DEN	0.007	0.031	FAC_POP_DEN	0.006	0.031
21	CWA	0.000	Industry	−0.002	0.528	Industry	−0.002	0.528

and Painsky, 2022; Alaimo Di Loro et al., 2023; Ugirumurera et al., 2024). In the specific context of Song et al.'s study, the high dimensionality of molecular fingerprints increases the risk of overfitting, leading the model to capture noise rather than true signal. Put simply, a common drawback of machine learning models, including XGB, is their propensity to overfit by prioritizing less relevant features in a greedy pursuit of marginal gains in predictive accuracy. Even with regularization techniques employed to mitigate overfitting, these inherent architectural biases persist.

3. Limitations of SHAP

Crucially, SHAP values, while seemingly insightful, inherit and can even exacerbate biases from the underlying model (Bilodeau et al., 2024; Huang and Marques-Silva, 2024; Kumar et al., 2021; Lones, 2024). This dependency is evident in the SHAP calculation itself, which directly utilizes the function 'explain = SHAP(model)'. Since SHAP sorely relies on the model's output for its explanations, it is inherently vulnerable to the model's biases. For instance, overfitting caused by XGB can manifest in SHAP values, thereby highlighting features that appear spuriously important due to the model's overfitting. Consequently, SHAP presents a significant drawback: it lacks the capacity to rectify the biases inherent in XGB and instead reinforces these biases indiscriminately. Therefore, interpreting SHAP values as definitive indicators of genuine feature importance is problematic.

4. Validation challenges

Fundamentally, the absence of ground truth values for feature importance makes validation extremely challenging. Different models use distinct methodologies for calculating feature importance, resulting in model-specific biases and varying rankings. As previously stated, machine learning models exhibit a propensity to introduce overfitting biases in their endeavor to improve predictive accuracy. Consequently, feature importance rankings that demonstrate high predictive accuracy do not necessarily reflect genuine associations. It is essential to

Table 2
Variable descriptions (alphabetical order).

Variable	Description
AIR_FLAG	Indicates whether the facility has air pollution permits or violations.
CAA	Refers to the Clean Air Act, a U.S. federal law that regulates air emissions.
CWA	Refers to the Clean Water Act, a U.S. federal law that regulates water pollution.
EJSCREEN_FLAG_US	Indicates potential environmental justice concerns based on EPA's EJSCREEN tool.
FAC_COMPLIANCE_STATUS	The compliance status of the facility with environmental regulations.
FAC_PERCENT_MINORITY	The percentage of minority population living near the facility.
FAC_POP_DEN	The population density near the facility.
FAC_PROGRAMS_WITH_SNC	The number of environmental programs with significant non-compliance at the facility.
FAC_QTRS_WITH_NC	The number of quarters with non-compliance at the facility.
FAC_SNC_FLG	Indicates whether the facility has significant non-compliance issues.
Federal Facility	Indicates whether the facility is owned or operated by the federal government.
GHG_FLAG	Indicates whether the facility reports greenhouse gas emissions.
Industry	The industry sector of the facility.
NPDES_FLAG	Indicates if facility has water discharge permit.
RCRA	Refers to the U.S. law regulating solid and hazardous waste.
RCRA_FLAG	Indicates whether the facility is regulated under RCRA.
Region	The EPA region where the facility is located.
SDWIS_FLAG	Indicates if facility is under drinking water regulations.
State	The U.S. state where the facility is located.
Status	The operational status of the facility.
TRI_FLAG	Indicates whether the facility reports to the Toxic Release Inventory.

recognize that numerous prior studies have misinterpreted this crucial aspect. This issue is particularly salient in Song et al.'s study, which employs complex molecular fingerprints as features. The high dimensionality and potential collinearity of these fingerprints significantly complicate the interpretation of SHAP values. Molecular fingerprints, encoding structural information, can have thousands of dimensions, making it difficult to isolate individual feature effects. Collinearity can cause SHAP values to distribute importance across correlated features, thus diluting the perceived importance of individual features. The complexity of molecular fingerprints also renders SHAP values susceptible to even small changes in the data or model, leading to significant shifts in their values and compromising their stability and reliability as indicators of feature importance. For example, minor dataset variations or slight model parameter modifications can substantially alter SHAP values, reducing their dependability for consistent feature importance assessment.

5. Proposed solutions

To address these limitations, we propose focusing on three critical aspects: data distribution, statistical relationships between variables, and statistical validation. Understanding data distribution is crucial for selecting appropriate modeling techniques. Investigating statistical interactions, particularly with non-parametric approaches, is essential for capturing complex relationships. Furthermore, statistical validation, including hypothesis testing and *p*-value analysis, is vital for ensuring observed relationships are not due to chance. These three aspects are comprehensively addressed by robust statistical methods. Instead of relying on XGB and SHAP for feature selection, we advocate for unbiased, robust statistical methods, such as Spearman's rho and Kendall's tau, coupled with *p*-values (Okoye and Hosseini, 2024). 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 (Metsämuuronen, 2021).

Table 3
Prediction accuracy of XGB and statistical tests for penalized factories.

	XGBoost	Spearman's Rho	Kendall's Tau
All 21 variables	0.822–0.845	0.822–0.845	0.822–0.845
Top 10 variables	0.825–0.845	0.805–0.832	0.805–0.832

6. Simulation results

Given the proprietary nature of the dataset referenced in Song et al. (2024), we conducted an independent validation utilizing the publicly accessible PFAS Industry Sectors Dataset, a compilation from the U.S. Environmental Protection Agency's open data, aggregated within the PFAS Central Data Hub (2025). This dataset represents the most authoritative and reliable publicly available resource for PFAS-related analyses. As shown in Table 1, we analyzed 21 features to determine the factors influencing regulatory penalties across 7916 U.S. industrial facilities. Variable descriptions are provided in Table 2. Notably, while the top two feature importances exhibited concordance between XGB and statistical testing, substantial deviations were observed from the fourth rank onwards. Specifically, XGB ranked "Compliance Status" below twelfth, while statistical methods ranked it fourth to sixth.

7. Discussion

The counterintuitive bias displayed by XGB is demonstrably linked to its prioritization of marginal gains in prediction accuracy, as evidenced by the 20-fold cross-validation results presented in Table 3. With all 21 features, using the exact same variables, both XGB and statistical tests naturally achieved comparable prediction accuracies (0.822–0.845). However, when restricted to the top 10 features, XGB maintained its accuracy (0.822–0.845), while statistical test accuracy declined slightly (0.805–0.832). This discrepancy reveals that XGB, in its pursuit of a mere 1 % increase in prediction accuracy, unjustifiably downplayed the significance of "Compliance Status," a feature of undeniable critical importance. The inherent biases within XGB are thus unequivocally established. Given that SHAP values are derived from XGB outputs, it is logically and empirically sound to conclude that SHAP values will similarly inherit these biases, compromising the integrity of subsequent interpretations.

8. Conclusion

In summary, while the study demonstrates promising predictive performance, its reliance on XGB and SHAP for feature selection raises significant concerns about the validity of the reported feature importances. We recommend a reassessment using robust statistical methods. This approach, by prioritizing statistical rigor and minimizing model-specific biases, will not only address the potential discrepancies in the current findings but also provide a more robust foundation for future research in this area.

CRediT authorship contribution statement

Souichi Oka: Writing – original draft, Investigation, Conceptualization. **Yoshiyasu Takefuji:** 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 have no conflicts of interest to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2025.179714.

Data availability

Python code and data set are publicly available at GitHub site: https://github.com/souichi-oka/pfas-analysis.

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., et al., 2024. Impossibility theorems for feature attribution. Proc. Natl. Acad. Sci. U. S. A. 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. https://doi.org/10.48550/arXiv.1801.01489.

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. Proces. Syst. 34, 26598–26608.

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

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.

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.

PFAS Central Data Hub, 2025. PFAS industry sectors dataset for the US. https://pfascentral.org/data-hub/. (Accessed 27 March 2025).

Song, C., Gu, Q., Zhang, D., Zhou, D., Cui, X., 2024. Prediction of PFAS bioaccumulation in different plant tissues with machine learning models based on molecular fingerprints. Sci. Total Environ. 950, 175091. https://doi.org/10.1016/j.scitotenv.2024.175091.

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.

Souichi Oka[a,*], Yoshiyasu Takefuji[b]

[a] *SciencePark Corporation, 3-24-9 Iriya-Nishi Zama-shi, Kanagawa 252-0029, Japan*

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

[*] Corresponding author.

E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka), takefuji@keio.jp (Y. Takefuji).