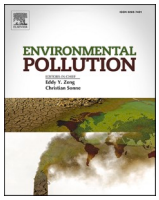
Environmental Pollution 397 (2026) 128013

Contents lists available at ScienceDirect

Environmental Pollution

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



Feature ranking instability in supervised model-SHAP pipelines compromises the reliability of clinical predictions

ARTICLE INFO

Keywords:
Gestational diabetes mellitus
Feature importance validity
LightGBM interpretability
SHAP analysis limitations
Supervised learning bias

The journal Environmental Pollution has published a substantial and rapidly increasing body of literature addressing machine learning interpretability, including 135 articles on feature importance (48 published in 2025 and 16 in 2026), 122 articles on feature selection (36 in 2025 and 13 in 2026), and 114 articles on SHapley Additive exPlanations (SHAP) (54 in 2025 and 16 in 2026). This remarkable publication trajectory signals an accelerating and widespread interest among environmental pollution researchers in moving beyond mere predictive modeling toward understanding the true associations and causal relationships between environmental variables, a scientifically meaningful but methodologically challenging endeavor.

Supervised machine learning models inherently possess two conceptually distinct dimensions of accuracy that are frequently conflated in environmental research. The first is target prediction accuracy, referring to the model's ability to correctly predict an outcome or response variable, which can be rigorously and objectively validated against known ground truth labels through established metrics such as RMSE, R^2, or AUC. The second is feature importance accuracy, referring to the model's attribution of relative influence or contribution to each input variable in driving predictions. A critical yet widely overlooked distinction is that while target prediction accuracy benefits from verifiable ground truth for validation, feature importance estimates derived from supervised models have no equivalent ground truth benchmark against which their correctness can be tested. This fundamental absence of a validation standard for feature importance creates a dangerous blind spot, where researchers may report and interpret feature rankings with unwarranted confidence, leading to systematically erroneous conclusions about which environmental variables truly matter and potentially misdirecting policy decisions, monitoring strategies, and scientific understanding.

SHapley Additive exPlanations (SHAP) has gained widespread adoption as a post-hoc explainability tool, often perceived as an objective and trustworthy method for quantifying variable contributions. However, this perception contains a fundamental misconception that must be clearly understood. The defining computational relationship in SHAP, expressed as explainer = SHAP(model), reveals a critical structural dependency: SHAP does not operate independently of the underlying model but is entirely and irrevocably bound to it. SHAP is mathematically constrained to explain the predictions of a given supervised model, meaning it can only redistribute and decompose whatever patterns and associations the model has already encoded. Because the model's feature importance is itself unvalidated against any ground truth, SHAP inherits these pre-existing inaccuracies and biases wholesale. Furthermore, rather than correcting or auditing these biases, SHAP may amplify and formalize them, presenting model-specific artifacts as if they were objective, ground-truth-validated measures of variable importance. In essence, applying SHAP to a model with unreliable feature attributions does not produce reliable explanations but instead produces confidently presented, mathematically coherent, yet potentially misleading narratives about environmental variable relationships.

The environmental science community must therefore internalize a fundamental principle that runs counter to common practice: high target prediction accuracy does not guarantee, and provides no direct evidence for, reliable or accurate feature importances. A model can achieve excellent predictive performance by exploiting complex, redundant, or spurious statistical patterns in training data while simultaneously producing distorted, unstable, or ecologically meaningless feature rankings. The absence of ground truth for feature importance means that neither the model itself nor SHAP-based explanations derived from it can be self-validated through conventional accuracy metrics. Researchers must therefore avoid conflating predictive performance with explanatory validity, recognize that SHAP explanations are model-dependent rather than ground-truth-dependent, and seek independent validation frameworks, such as controlled simulations with known variable relationships, sensitivity analyses, or cross-method convergence testing, before drawing scientific conclusions about variable importance in environmental systems. Without these safeguards, the growing literature on feature importance and SHAP in environmental pollution research risks constructing an elaborate but scientifically unreliable edifice of explanations built upon unvalidated foundations.

Establishing valid associations requires two mandatory criteria: consistency and dose-response relationships. However, existing studies have systematically overlooked these validations (Ioannidis, 2008; Prasad and Jena, 2013; Roberts et al., 2019; Lai et al., 2025; Prada et al.,

https://doi.org/10.1016/j.envpol.2026.128013
Received 13 February 2026; Received in revised form 23 March 2026; Accepted 24 March 2026
Available online 24 March 2026

2025; Stamatakis et al., 2025; Ye et al., 2024), compromising their conclusions. This paper addresses these shortcomings by theoretically examining the underlying pitfalls and empirically demonstrating, through a controlled showcase, the severe distortions in feature importance estimates produced by supervised models—both with and without SHAP-based interpretability.

Gan et al. investigated maternal prenatal mercury exposure from rice and its impact on newborn neurobehavioral development using various machine learning techniques (Gan et al., 2026). Decision trees, Random Forest, XGBoost, and LightGBM models achieved AUCs of 0.833-0.909, outperforming other approaches. LightGBM demonstrated the best results with an AUC of 0.909 (95% CI: 0.858-0.961) and accuracy of 0.849 (95% CI: 0.775-0.907), with all metrics exceeding 0.80. To interpret the LightGBM model's predictions and identify key factors affecting neurobehavioral outcomes, researchers utilized SHAP values, providing valuable insights into feature contributions.

A critical limitation of Gan et al.'s approach lies in their reliance on supervised models like LightGBM, which exhibit a fundamental disconnect between target prediction accuracy and feature importance accuracy. While prediction accuracy can be objectively validated against labeled outcomes, feature importance lacks corresponding ground truth validation mechanisms, potentially leading to misleading interpretations. Moreover, the implementation of SHAP values (expressed functionally as explain = SHAP(model = XGBoost)) inherently constraints explanatory outputs to faithfully reflect the underlying supervised model's structure, thereby potentially inheriting and amplifying any biases in feature importance assessments without independent validation.

To assess true associations between environmental variables, rigorous testing of consistency and dose-response relationships is mandatory for validating reliable and meaningful outcomes. A critical but widely neglected aspect of this process is that examining ordered sets of features is essential, as the ranked sequence of variable importance carries substantive scientific meaning that unordered sets fundamentally cannot capture. Existing studies have largely failed in this regard by examining non-ordered sets of features, effectively discarding the hierarchical structure of variable contributions and rendering their consistency assessments methodologically inadequate and scientifically unreliable. This paper introduces a leave-top-feature-out approach as a principled framework for implementing consistency and dose-response relationship testing while explicitly preserving and evaluating the ordered structure of feature rankings. The procedure operates as follows: the top five features are first identified from the complete dataset, referred to as set1, representing the baseline feature importance ranking. The highest-ranked feature is then systematically removed from the complete dataset to create a reduced dataset, from which the top four features are re-selected and designated as set2. The degree of agreement or disagreement in feature ranking orders between set1 and set2 is then evaluated to assess consistency. The underlying rationale of this approach is that removing the most dominant feature from the dataset creates a substantial perturbation in the feature importance landscape, as the remaining features must now redistribute their explanatory contributions in the absence of the previously dominant variable. If the feature importance rankings remain consistent and the expected dose-response relationships are preserved despite this perturbation, this provides meaningful evidence that the identified associations reflect genuine and stable relationships rather than model-specific artifacts or statistical coincidences. Conversely, substantial instability in rankings following the removal of the top feature signals unreliable feature importance attributions that should not be interpreted as true associations.

Table 1 presents cross-validation accuracy results examining the prediction contributions and feature ranking stability across multiple approaches: XGBoost (XGB), XGBoost with SHAP interpretation (XGB-SHAP), LightGBM (LGBM), LGBM with SHAP, feature agglomeration (FA), highly variable gene selection (HVGS), and Spearman's correlation. Due to the unavailability of datasets fromGan et al., our analysis of publicly accessible gestational diabetes datasets (Kaggle dataset) reveals critical methodological concerns. Specifically, LightGBM models—both with and without SHAP interpretability layers—exhibit problematic instability in feature ranking orders across cross-validation iterations. Conversely, unsupervised dimensionality reduction approaches (FA, HVGS) and nonparametric statistical methods such as Spearman's correlation demonstrate substantially greater consistency in feature rankings, thereby circumventing label-dependent distortions inherent to supervised learning algorithms.

Table 1
Cross-validation accuracy and feature rankings per algorithm.

Method	CV5 Accuracy	Top 5 Features	Top 4 Features
XGB	0.9719	Case Number, Gestation in previous Pregnancy, Dia BP, OGTT, BMI	OGTT, HDL, Dia BP, Prediabetes
XGB-SHAP	0.9711	Case Number, BMI, OGTT, HDL, Dia BP	OGTT, HDL, Sys BP, BMI
LGBM	0.9705	Case Number, OGTT, BMI, HDL, Dia BP	BMI, OGTT, Sys BP, HDL
LGBM-SHAP	0.9694	Case Number, OGTT, HDL, BMI, Gestation in previous Pregnancy	HDL, OGTT, Sys BP, BMI
FA	0.9716	HDL, Case Number, Prediabetes, PCOS, BMI	Case Number, Prediabetes, PCOS, BMI
HVGS	0.9699	Case Number, OGTT, Sys BP, HDL, BMI	OGTT, Sys BP, HDL, BMI
Spearman	0.9713	OGTT, BMI, HDL, Case Number, Prediabetes	BMI, HDL, Case Number, Prediabetes

These findings illuminate a fundamental methodological gap pervading current predictive modeling literature: the critical need for rigorous dual assessment of (1) reproducibility and consistency in feature importance rankings across model iterations, and (2) biological plausibility through dose-response relationship validation between predictors and clinical outcomes. These essential evaluation criteria remain largely absent from existing gestational diabetes prediction studies, potentially compromising both the interpretability and clinical utility of reported models.

When strong collinearity exists among features, removing the top-ranked feature may redistribute its explanatory weight onto correlated substitutes rather than revealing genuinely independent contributors. However, this redistribution itself serves as a meaningful consistency test: a reliable feature selection method should produce stable rankings regardless of whether collinear features are present or absent. Additionally, the method may exhibit reduced reliability when the number of features is small or the sample size is extremely limited, and should therefore be applied with caution in accordance with the specific data context.

For purposes of reproducibility and transparency, Python code, gdm.py is publicly available at GitHub (gdm.py. https).

AI use

Not applicable.

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.

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

The authors do not have permission to share data.

References

Gan, C., Feng, X., Yang, K.L., Sun, G., Luo, W., Yang, Q., Zhang, W., Shi, Y., Wang, L., Xiong, M., Abdelhafiz, M.A., 2026. Maternal prenatal mercury exposure from rice and its association with newborn neurobehavioral development. Environ. Pollut. 390, 127465. https://doi.org/10.1016/j.envpol.2025.127465.

Gdm.py. https://github.com/y-takefuji/diabetes.

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

Kaggle dataset: 'gestational diabetic dat set. xlsx'. https://www.kaggle.com/datasets/sumathisanthosh/gestational-diabetes-mellitus-gdm-data-set.

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.

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. https://doi.org/10.1016/j.jid.2018.12.023.

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. https://doi.org/10.1136/bjsports-2024-108484.

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 completed this research and wrote the program and this article.

According to ScholarGPS (prior 5 years) on productivity, 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.

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