Science of the Total Environment 1000 (2025) 180475

Contents lists available at ScienceDirect

Science of the Total Environment

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

Letter to the Editor

A composite theory-guided framework for robust feature attribution in $PM_{2.5}$ ionic composition modeling

ARTICLE INFO

Keywords:
$PM_{2,5}$ ionic composition modeling
Feature agglomeration
Highly variable gene selection
Spearman's rank correlation
SHAP explainability

ABSTRACT

Kang et al. (2025) developed a theory guided framework that fuses satellite observations, land use regression and extreme gradient boosting to map $PM_{2.5}$ ionic species across Taiwan with SHAP for feature attribution. Despite high predictive accuracy, model specific importance scores from tree based methods may misrepresent true associations due to hyperparameter sensitivity, multicollinearity and data imbalance. We advocate a composite strategy combining unsupervised feature agglomeration to cluster correlated predictors, highly variable gene selection to identify dominant covariates and Spearman rank correlation with *p*-value testing to quantify monotonic relationships without distributional assumptions. This pipeline yields stable interpretable importance rankings under bootstrapping and cross validation. SHAP attributions remain useful for exploring interactions but require independent validation with input perturbations, alternative models or simulated ground truth data to ensure reliability.

Kang Lo et al. (2025) developed a theory-guided framework that fuses satellite observations, land-use regression (LUR), and machine learning to map the spatiotemporal distribution of $PM_{2.5}$ ionic species across Taiwan. First, LUR draws on domain knowledge to screen and construct candidate predictors; next, extreme gradient boosting (XGBoost) builds a highly nonlinear, high-resolution prediction model; and finally, Shapley additive explanations (SHAP) decompose the fitted ensemble's output to quantify each feature's marginal impact. By combining hypothesis-driven feature selection with powerful modeling and transparent attribution, their approach delivers both accurate estimates and interpretable insights into the drivers of $PM_{2.5}$ composition.

Despite the appeal of combining LUR, XGBoost and SHAP, key theoretical and empirical challenges arise because the importance scores produced by XGBoost are purely model-specific indicators of how each variable contributes to reducing prediction error within that particular ensemble, not measures of true associations or causal effects. In contrast, predictive accuracy can be directly verified by comparing model outputs to held-out PM2.5 measurements and computing well-understood metrics such as root mean squared error or R^2. Feature importance scores, however, are generated internally, often according to gain, cover or impurity-reduction criteria, and there is no independent benchmark against which to validate them. As a result, even a model that forecasts exceptionally well can assign exaggerated importance to variables that interact favorably with the chosen algorithm settings, especially under multicollinearity, unbalanced data or particular choices of tree depth and learning rate. In other words, high target prediction accuracy does not guarantee reliable feature importances (Parr et al., 2024; Watson and Wright, 2021; Molnar et al., 2022; Lipton, 2018; Fisher et al., 2019; Lenhof et al., 2024; Mandler and Weigand, 2024; Potharlanka and Bhat, 2024; Wood et al., 2024). A review of more than three hundred peer-reviewed studies has documented these systematic biases in tree-based models, revealing that high-variance or correlated features frequently receive inflated importance while genuinely relevant but less variable predictors are underestimated. Because feature importance in supervised learning reflects each variable's contribution to the model's predictions rather than its real-world association with the outcome, treating those scores as evidence of ecological or causal drivers risks drawing misleading conclusions.

SHAP explanations using the call explain = SHAP(model = XGBoost) assign each predictor a numerical contribution by averaging its marginal impact across all possible subsets of features. Although this approach provides a detailed breakdown of how the fitted XGBoost ensemble arrives at each prediction, it does not eliminate the biases that already exist in the model's internal structure. SHAP treats the learner as a black box and derives attribution values directly from its response surface. If the base XGBoost model has overemphasized a correlated or high variance feature or has underweighted a weaker but truly important variable, those distortions will be inherited by the SHAP values and can even be magnified by the exhaustive subset averaging (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). In other words, SHAP does not act as a debiasing layer but merely decomposes whatever patterns and artifacts the model has captured. To gain confidence that high SHAP scores reflect genuine ecological or causal drivers rather than algorithmic quirks, researchers must complement SHAP with independent validation— for example, perturbing inputs to test attribution stability, employing simulated datasets with known feature effects or using orthogonal causal inference methods. Without such checks, interpreting SHAP values from a biased learner remains unsafe.

Because no single supervised learner can recover unbiased true associations from complex observational data, we recommend a composite theory guided strategy alongside or instead of Shapley additive explanations for XGBoost. The first step applies unsupervised feature

https://doi.org/10.1016/j.scitotenv.2025.180475
Received 1 August 2025; Received in revised form 22 August 2025; Accepted 8 September 2025
Available online 11 September 2025

engineering to tame multicollinearity and reduce dimensionality by using feature agglomeration (FA) to group highly correlated predictors into clusters for more stable, lower dimensional representations and highly variable gene selection (HVGS) adapted to environmental covariates to identify the variables with the greatest spatiotemporal variance. Next, predictor and outcome relationships are quantified with non-target supervised statistical methods Spearman's rank correlation combined with rigorous *p*-value testing; this nonlinear nonparametric statistic captures any monotonic association without imposing linearity or distributional assumptions and yields feature rankings that remain consistent under bootstrapping and cross validation.

Unsupervised methods like FA and HVGS operate without labels and leverage variance structure to simplify data before any outcome is considered. By clustering correlated variables and retaining the most variable, informative features, they reduce multicollinearity and noise, yielding stable, low-dimensional representations that are less susceptible to label-driven bias or model misspecification. In contrast, supervised feature importance reflects contributions to prediction rather than true associations and can be distorted by confounding, feedback, or correlated predictors. Pairing the unsupervised step with Spearman's rank correlation and rigorous *p*-value testing quantifies monotonic relationships without strong distributional assumptions, producing feature rankings that are more stable under resampling and closer to underlying associations than rankings from a single supervised learner or post-hoc tools like SHAP for XGBoost.

Lacking access to the datasets used by Kang et al., we evaluated feature selection on the MNIST benchmark (70,000 samples; 784 features). Empirically, feature agglomeration (FA) and highly variable gene selection (HVGS) produced notably more stable feature rankings, whereas XGBoost exhibited greater instability in its ranked importances. Across methods, Random Forest achieved a cross-validated accuracy of 0.8861 ± 0.0025 but was highly unstable, XGBoost reached 0.8172 ± 0.0034 and was likewise highly unstable, FA achieved 0.8368 ± 0.0021 with stable rankings, HVGS reached 0.8441 ± 0.0023 with stable rankings, and Spearman attained 0.5196 ± 0.0030, also with stable rankings. For the stability assessment, we repeatedly selected the top 30 features from the full set across the five methods, performed cross-validation, then removed the single highest-ranked feature from the full set to form a reduced dataset, reselected the top 29 features, and compared the resulting ranking orders. For reproducibility and transparency, the Python script mniststability.py, which reports cross-validation accuracy and the top-10 feature rankings for XGBoost, FA, HVGS, and Spearman, is publicly available on GitHub (GitHub, 2025).

In contrast, SHAP attributions inherit distortions from the underlying XGBoost model due to factors such as sensitivity to hyperparameters, multicollinearity, data imbalance and variance, which can cause unstable importance rankings across repeated fits. By integrating FA, HVGS and Spearman correlation, researchers can produce reproducible, interpretable importance estimates that are less vulnerable to the algorithmic quirks of post hoc attribution methods, while reserving SHAP values as a complementary diagnostic for complex interactions only after validating their rankings against input perturbations, alternative modeling choices and simulated ground truth data.

CRediT authorship contribution statement

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

Authors' contributions

Yoshiyasu Takefuji completed this research and wrote this article.

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.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

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

Not applicable.

Funding

This research has no fund.

Code availability

Not applicable.

Declaration of competing interest

The author has no conflict of interest.

Data availability

Not applicable.

References

Bilodeau, B., Jaques, N., Koh, P.W., Kim, B., 2024. Impossibility theorems for feature attribution. Proc. Natl. Acad. Sci. Proc. Natl. Acad. Sci. USA 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, 2025. mniststability.py. https://github.com/y-takefuji/mnist/blob/main/mniststability.py.

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.

Kang Lo, T.-H., Lin, F.-C., Wang, F.-C., Shiu, Y.-S., Chen, C.-C., Lin, Y.-C., Huang, C.-S., Liao, H.-T., Wu, Y.-L., Wu, C.-F., 2025. Integrating satellite information, land use regression, and machine learning to estimate the spatiotemporal variation of ionic composition in PM2.5 across Taiwan. Sci. Total Environ. 994. https://doi.org/10.1016/j.scitotenv.2025.180050. Article 180050.

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.

Lenhof, K., Eckhart, L., Rolli, L.M., Lenhof, H.P., 2024. Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. Brief. Bioinform. 25 (5), bbae379. https://doi.org/10.1093/bib/bbae379.

Létoffé, O., Huang, X., Marques-Silva, J., 2025. Towards trustable SHAP scores. Proceedings of the AAAI Conference on Artificial Intelligence 39 (17), 18198–18208. https://doi.org/10.1609/aaai.v39i17.34002.

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

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

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.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., 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, vol. 13200. Springer, p. 4. https://doi.org/10.1007/978-3-031-04083-2_4.

Parr, T., Hamrick, J., Wilson, J.D., 2024. Nonparametric feature impact and importance. Inform. Sci. Inf. Sci. 653, 119563. https://doi.org/10.1016/j.ins.2023.119563.

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.

Watson, D.S., Wright, M.N., 2021. Testing conditional independence in supervised learning algorithms. Mach. Learn. 110 (8), 2107–2129. https://doi.org/10.1007/s10994-021-06030-6.

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

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.

Yoshiyasu Takefuji[a,*]

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

[*] Corresponding author.
E-mail address: takefuji@keio.jp.