Applied Soil Ecology xxx (xxxx) xxx

Contents lists available at ScienceDirect

Applied Soil Ecology

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

Letter to the Editor

Beyond SHAP: Why supervised models require external feature validation and the case for leave-top-feature-out (LTFO) in soil health assessment☆

ARTICLE INFO

Keywords:
Shapley additive explanations (SHAP)
Leave-top-feature-out (LTFO)
Feature importance validation
Bradford hill criteria
Soil health assessment

ABSTRACT

Li et al. (2026) proposed a soil health assessment framework using interpretable machine learning, wherein XGBoost achieved the highest accuracy and, combined with SHAP, identified seven key soil indicators. This paper argues that such supervised model–SHAP integration is methodologically flawed: while target prediction accuracy is verifiable against ground truth, feature importance accuracy is not, leaving SHAP to inherit and amplify unvalidated biases from the underlying model. Leave-top-feature-out serves as an external validation framework, assessing feature consistency and dose-response relationships. Using a public soil microbiome dataset, we compared seven algorithms—RF, RF-SHAP, XGBoost, XGBoost-SHAP, Feature Agglomeration, HVGS, and Spearman correlation. All achieved similar R^2 (0.995), yet supervised/SHAP methods showed unstable rankings after the highest feature removal, while unsupervised and Spearman methods remained perfectly stable. These results demonstrate that high target accuracy does not guarantee valid feature importance, and that LTFO provides essential external validation where SHAP alone cannot.

Li et al. (2026) explored novel soil health assessment framework and microbial community characterization for paddy field by interpretable machine learning. They employed multiple supervised models where the extreme gradient boosting (XGBoost) model achieved the highest prediction accuracy and integrated with Shapley additive explanation (SHAP) to identify seven key soil indicators.

This paper raises theoretical and methodological concerns regarding the use of supervised models with/without SHAP for interpretation. Li et al. must understand the fundamental principles of supervised models. Supervised models inherently possess two distinct types of accuracy: target prediction accuracy and feature importance accuracy. While target predication accuracy can be rigorously validated against ground truth labels, feature importance lacks any corresponding ground truth for accuracy validation, indicating the absence of ground truth in feature assessment, potential leading to erroneous interpretations without external feature validation frameworks.

Furthermore, the implementation of explainer = SHAP(model) suggests that SHAP is forced to solely rely on a given supervised model with unvalidated feature importance, faithfully inheriting and potentially amplifying biases in feature importance derived from the supervised model. While SHAP is a powerful tool, SHAP cannot correct unvalidated feature importance, implying that external feature validation frameworks are needed to ensure true associations. To accurately calculate true associations, grounded in the extended Bradford Hill criteria (Hill, 1965), two mandatory components must be examined: feature consistency and dose-response relationships (Hill, 1965; Ioannidis, 2003; Loughman et al., 2026; Takefuji, 2025; Takefuji, 2026a; Takefuji, 2026b; Takefuji, 2026c; Thapa et al., 2026). This paper recommends the use of leave-top-feature-out (LTFO) approaches: select top n features from full set (set1), remove the highest feature from full set to create a reduced dataset, re-select top n-1 features from the reduced dataset (set2), and compare ordered sets of feature importance rankings between two sets. Removing the highest feature creates a substantial impact on feature ranking orders to examine feature consistency and dose-response relationships. The LTFO has been successfully validated across multiple domains and disciplines using different datasets to demonstrate its robustness (Takefuji, 2025; Takefuji, 2026a; Takefuji, 2026b; Takefuji, 2026c).

We empirically examine the quality of selected features through three analytical dimensions: cross-validation for assessing target prediction accuracy, feature consistency, and dose-response relationships via the LTFO (Leave-Two-Features-Out) approach, applied across multiple methods, including supervised models (Random Forest: RF, RF-SHAP, XGBoost: XGB, XGB-SHAP), unsupervised models (Feature Agglomeration: FA, Highly Variable Gene Selection: HVGS), and a nonparametric, non-target-prediction method such as Spearman's correlation. These analyses are conducted using a publicly available soil microbe dataset (Soil_microbe_dataset.csv) comprising 100,000 samples and 13 features, with 'β-Glucosidase (μmol/g/h)' serving as the target variable (Kaggle, 2026). For a fair comparison, default settings including random number seeds are used across all algorithms, without scaling, transformation, or hyperparameter tuning. For reproducibility and transparency, the Python code (cvshap.py) is publicly available on GitHub (2026).

As shown in Table 1, across all seven feature selection algorithms evaluated, predictive accuracy for β-glucosidase activity was

☆ According to ScholarGPS, 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.

https://doi.org/10.1016/j.apsoil.2026.107454
Received 28 May 2026; Received in revised form 5 August 2026; Accepted 10 September 2026
0929-1393/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Table 1
cross-validation accuracy and feature rankings across algorithms.

Method	CV6 R^2	Top 6 feature rankings	Top 5 feature rankings
RF	0.995	Organic_C (%), CO2_Emission (μg/g/day), Bacteria_Abundance (%), Fungi_Abundance (%), Soil_pH, C_N_Ratio	CO2_Emission (μg/g/day), **Fungi_Abundance (%), Bacteria_Abundance (%),** Soil_pH, C_N_Ratio
RF-SHAP	0.995	Organic_C (%), CO2_Emission (μg/g/day), Bacteria_Abundance (%), Soil_pH, Fungi_Abundance (%), C_N_Ratio	CO2_Emission (μg/g/day), **Fungi_Abundance (%), Bacteria_Abundance (%), Soil_pH**, C_N_Ratio
XGB	0.995	CO2_Emission (μg/g/day), Organic_C (%), Fungi_Abundance (%), Land_Use_Type_Organic, Bacteria_Abundance (%), Total_N (%)	Organic_C (%), Fungi_Abundance (%), **Bacteria_Abundance (%), Land_Use_Type_Traditional**, Total_N (%)
XGB-SHAP	0.995	Organic_C (%), CO2_Emission (μg/g/day), Bacteria_Abundance (%), Total_N (%), Fungi_Abundance (%), Soil_pH	CO2_Emission (μg/g/day), **Fungi_Abundance (%), C_N_Ratio, Bacteria_Abundance (%),** Soil_pH
FA	0.995	NH4_Nitrate (μg/g), C_N_Ratio, Soil_Depth_cm, Urease (μmol/g/h), Organic_C (%), Soil_pH	C_N_Ratio, Soil_Depth_cm, Urease (μmol/g/h), Organic_C (%), Soil_pH
HVGS	0.995	NH4_Nitrate (μg/g), C_N_Ratio, Fungi_Abundance (%), Bacteria_Abundance (%), Soil_Depth_cm, CO2_Emission (μg/g/day)	C_N_Ratio, Fungi_Abundance (%), Bacteria_Abundance (%), Soil_Depth_cm, CO2_Emission (μg/g/day)
spearman	0.995	Organic_C (%), CO2_Emission (μg/g/day), C_N_Ratio, Urease (μmol/g/h), Total_N (%), NH4_Nitrate (μg/g)	CO2_Emission (μg/g/day), C_N_Ratio, Urease (μmol/g/h), Total_N (%), NH4_Nitrate (μg/g)

Bold entries indicate features with inconsistent and unstable rankings.

remarkably consistent, with each method achieving an identical cross-validated R^2 of 0.995 when using its respective top-6 ranked features. This uniformity indicates that regardless of the underlying feature selection strategy—whether supervised (RF, RF-SHAP, XGB, XGB-SHAP), unsupervised (FA, HVGS), or correlation-based (Spearman)—the selected feature subsets retained sufficient information to explain nearly all variance in the target enzyme activity, suggesting a high degree of redundancy among the top-ranked soil and microbial predictors.

Despite this convergence in predictive performance, marked differences emerged in the stability and consistency of feature rankings across algorithm categories. The supervised tree-based models, both with and without SHAP-based interpretation, exhibited inconsistent and unstable rankings after feature removal. For RF, the top-6 set (Organic_C, CO2_Emission, Bacteria_Abundance, Fungi_Abundance, Soil_pH, C_N_Ratio) only partially persisted into the top-5 after removing the highest-ranked feature, with the remaining five features preserving their relative order but demonstrating sensitivity to the removal step. XGB displayed the most pronounced instability, with *Land_Use_Type_Organic* dropping out of the top-5 entirely and being replaced by a categorically distinct dummy variable (*Land_Use_Type_Traditional*)—indicating that the model's ranking of categorical land-use encodings was highly sensitive to minor perturbations in the feature space rather than reflecting a robust underlying signal. Similarly, XGB-SHAP showed a reordering in which *C_N_Ratio* re-entered the top-5 ranking at a different relative position than in RF-SHAP, and *Soil_pH* shifted rank, further underscoring the lack of reproducibility in SHAP-derived importance rankings once the dominant feature (Organic_C) was excluded. These inconsistencies suggest that supervised importance metrics—whether derived from impurity-based splits or SHAP attributions—are sensitive to feature collinearity and re-partitioning of explained variance among correlated predictors (e.g., Organic_C, CO2_Emission, Bacteria/Fungi_Abundance), leading to rank instability despite stable overall model accuracy.

In contrast, the unsupervised feature selection methods (FA, HVGS) and the Spearman correlation-based approach demonstrated perfect consistency and stability in feature rankings. For all three methods, the top-5 feature set obtained after removing the single highest-ranked feature was an exact, order-preserving subset of the original top-6 ranking, with no reordering, substitution, or rank exchange among the remaining features. FA consistently prioritized NH4_Nitrate, C_N_Ratio, Soil_Depth_cm, Urease, Organic_C, and Soil_pH, retaining the identical relative order across both rankings. HVGS likewise preserved its full ranking order (NH4_Nitrate → C_N_Ratio → Fungi_Abundance → Bacteria_Abundance → Soil_Depth_cm → CO2_Emission) after feature removal, as did the Spearman correlation ranking (Organic_C → CO2_Emission → C_N_Ratio → Urease → Total_N → NH4_Nitrate). This perfect rank preservation reflects the fact that these methods derive importance from static, data-intrinsic properties—variance, cluster-distance structure, or monotonic correlation with the target—that are invariant to the removal of any single feature, unlike model-based importance metrics that re-estimate feature contributions relative to the remaining predictor set after each model refit.

Taken together, these findings indicate that while all seven algorithms converge on statistically equivalent predictive performance ($R^2 = 0.995$), only the unsupervised and correlation-based methods provide stable, reproducible feature rankings suitable for robust biological interpretation, whereas supervised model-based rankings (RF, RF-SHAP, XGB, XGB-SHAP)—despite their equally high accuracy—are more susceptible to rank instability driven by feature collinearity and re-partitioning of importance following feature removal.

Table 2 summarizes the key differences between SHAP and LTFO. SHAP explains feature importance *within* the model, whereas LTFO functions as an *external* feature validation framework. Unlike SHAP, which offers no independent means of validation, LTFO treats the resulting performance change as the validation itself. Moreover, while SHAP inherits and operates under the assumptions of the underlying model, LTFO actively interrogates and challenges those assumptions. Consequently, SHAP produces an attribution score that cannot be independently verified, whereas LTFO yields a measurable, empirically grounded performance delta. Epistemologically, this distinction reflects a deeper contrast: SHAP operates within a closed, self-referential loop, while LTFO operates within an open, empirical loop.

CRediT authorship contribution statement

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

Author contribution

Yoshiyasu Takefuji completed this research and wrote the program and this article.

Table 2
key differences between SHAP and LTFO.

Dimension	SHAP	LTFO
Scope	Explains within the model	Tests outside the model
Validation	No independent validation	Performance change is the validation
Assumptions	Inherits model assumptions	Challenges model assumptions
Output	Unverifiable attribution score	Measurable performance delta
Epistemology	Closed self-referential loop	Open empirical loop

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.

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.

Acknowledgments

None.

Code availability

Python code is publicly available at GitHub..

Data availability

The authors do not have permission to share data.

References

GitHub, 2026. cvshap.py. https://github.com/y-takefuji/soil_microbe.

Hill, A.B., 1965. The environment and disease: association or causation? Proc. R. Soc. Med. 58 (5), 295–300. https://doi.org/10.1177/003591576505800503.

Ioannidis, J.P.A., 2003. Genetic associations: false or true? Trends Mol. Med. 9 (4), 135–138. https://doi.org/10.1016/S1471-4914(03)00030-3.

Kaggle, 2026. Soil_microbe_dataset.csv. https://www.kaggle.com/datasets/isuranga/soil-microbe-dataset.

Li, Q., Jiang, N., Zhu, K., Kong, Y., Zhu, C., Li, J., Cao, X., Tian, W., Zhu, L., Yu, Y., Zhang, J., 2026. Novel soil health assessment framework and microbial community characterization for paddy field by interpretable machine learning. Appl. Soil Ecol. 223, 107022. https://doi.org/10.1016/j.apsoil.2026.107022.

Loughman, J., Bullimore, M.A., Flitcroft, D.I., 2026. Myopia and eye disease: from correlation to causation. Prog. Retin. Eye Res. 114, 101497. https://doi.org/10.1016/j.preteyeres.2026.101497.

Takefuji, Y., 2026a. Evaluating linear parametric Poisson regression vs nonparametric unsupervised learning in ulcerative colitis data. Gastroenterology. https://doi.org/10.1053/j.gastro.2026.04.011. Advance online publication.

Takefuji, Y., 2025. Beyond XGBoost and SHAP: unveiling true feature importance. J. Hazard. Mater. 488, 137382. https://doi.org/10.1016/j.jhazmat.2025.137382.

Takefuji, Y., 2026b. The reliability gap: why high predictive accuracy doesn't guarantee stable feature importance. Mar. Pollut. Bull. 226, 119398. https://doi.org/10.1016/j.marpolbul.2026.119398.

Takefuji, Y., 2026c. Ensuring reliable feature importance in food chemistry AI. Food Chem. 509, 148515. https://doi.org/10.1016/j.foodchem.2026.148515.

Thapa, S.K., Li, J., Mach, J., Hilmer, S.N., Kirkpatrick, C.M.J., 2026. Applying the Bradford Hill criteria to assess the independent causal roles of aging and medication in frailty progression: a systematic review. Drugs Aging 43 (1), 9–28. https://doi.org/10.1007/s40266-025-01273-7.

Yoshiyasu Takefuji[a,*]

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

[*] Corresponding author at: AI laboratory, SciencePark Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa, 252-0029, Japan.
E-mail address: takefuji@keio.jp.