Computers and Electronics in Agriculture 240 (2026) 111226

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Beyond the black box: Enhancing feature explainability in machine learning with SHAP and complementary approaches

Yoshiyasu Takefuji

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

ARTICLE INFO

Keywords
SHAP
Machine learning
Feature importance
Model interpretability
Policy implications

ABSTRACT

SHapley Additive Explanations (SHAP) are widely used to interpret machine learning models in agriculture and environmental decision-making, but SHAP inherits model misspecification, confounding, and distribution shift, risking unstable and policy-misleading importance rankings. This opinion advances a concrete, model-agnostic workflow that goes beyond SHAP: pair unsupervised structure checks (e.g., feature agglomeration, highly variable feature selection) with nonlinear, nonparametric association metrics and pre-registered sensitivity/stability analyses, then audit attributions with domain-informed negative controls and decision impact tests. The novelty lies in reframing SHAP from a standalone explainer to one component of a robustness protocol explicitly designed to reduce label-driven and proxy bias. This integrated approach yields more reliable variable importance, tighter uncertainty communication, and clearer links between attribution changes and real policy choices.

In 2025, Computers and Electronics in Agriculture reported that 48 of its published articles employed SHapley Additive Explanations (SHAP) to interpret machine-learning models—up sharply from 25 in 2024 and just 8 in 2023—highlighting the journal's accelerating interest in transparent, data-driven insights for agricultural challenges. Interpretable machine learning here means not only achieving high predictive accuracy but also understanding which input factors—such as land-use density, traffic volume or ambient temperature—drive the model's forecasts. SHAP adapts the Shapley value concept from cooperative game theory by treating each feature as a "player" in the prediction "game" and allocating credit in proportion to its average marginal contribution across all possible feature coalitions. Because Shapley values satisfy properties like additivity (ensuring that the sum of individual contributions equals the difference between a model's output and a baseline expectation) and symmetry (guaranteeing that interchangeable features receive equal attribution), SHAP provides a principled, mathematically grounded way to open the so-called "black box." By doing so, it helps researchers strike a balance between model performance and substantive insight into urban systems.

Building on this momentum, Li et al. (2025) conducted an investigation into the impact of stand structure on forest net primary productivity by employing a comprehensive integrated approach that combines multiple machine learning models with SHAP (SHapley Additive exPlanations) and Dynamic Structural Equation Modeling (DSEM). SHAP effectively identifies key variables and quantifies not only their significance but also the direction and magnitude of their impact. This study synergistically merges various machine learning feature selection techniques with SHAP to assess the relative importance of different stand structure factors on forest Net Primary Productivity (NPP), providing a nuanced understanding of their influence.

While supervised machine learning models like DSEM are adept at identifying patterns and achieving strong predictive performance on unseen data, the feature importance scores they produce—whether based on split gains, permutation drops, or SHAP values—lack intrinsic verifiability against an external "ground truth." In predictive tasks, we can validate accuracy by comparing forecasts with actual outcomes, but no equivalent standard exists to unequivocally establish the real-world significance of each input variable. As a result, high predictive accuracy does not necessarily ensure that the importance rankings are dependable or unbiased (Parr et al., 2024; Watson & Wright, 2021; Molnar et al., 2022; Lipton, 2018; Fisher, Rudin, & Dominici, 2019; Lenhof et al., 2024; Mandler & Weigand, 2024; Potharlanka et al., 2024; Wood et al., 2024). For example, features that correlate strongly with the target variable but lack causal influence can still receive disproportionately high importance scores. Additionally, the interplay among features can lead to ambiguous credit-sharing, complicating the assessment of their individual contributions. Various model parameters—such as tree depth, learning rate, and random seed—can also exert substantial influence on importance weights. Therefore, practitioners must recognize that feature importance reflects the model's internal logic rather

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.compag.2025.111226
Received 19 July 2025; Received in revised form 26 August 2025; Accepted 10 November 2025

than serving as an objective measure of each variable's scientific or policy significance. Furthermore, given the linear parametric nature of DSEM, the resulting feature importances and other metric scores may be significantly distorted when applied to nonlinear or nonparametric data contexts, leading to skewed interpretations.

Despite SHAP's rigorous game-theoretic underpinnings and its adherence to fairness axioms, its attributions remain tied exclusively to the chosen model and data sample. There is no independent ground truth for "true" feature importance; SHAP values quantify marginal contributions to the model's predictions, not bona fide causal or associative strengths in the system under study. Consequently, importance scores can be unstable: small changes in data splitting, sampling strategies, hyperparameter configurations or even the random seed may produce materially different rankings. Correlated features can divide or inflate credit arbitrarily, imbalanced data can skew allocations toward dominant classes, and overfitting can embed spurious patterns that SHAP dutifully explains. In practice, therefore, SHAP outputs should be treated as hypothesis generators rather than definitive answers. To build confidence in the insights derived, researchers must perform robustness checks such as sensitivity analyses, cross-model comparisons, domain-expert vetting and, where appropriate, formal causal inference methods before translating SHAP attributions into policy decisions.

The SHAP explanatory wrapper (e.g., explain=SHAP(model)) operates by querying the fitted DSEM model to estimate each feature's contribution to individual predictions. Because SHAP values are computed using the same conditional expectations and structural insights that DSEM has learned, any systematic distortions in feature splitting, sample weighting, or interaction handling will be carried through and, in some cases, magnified in the resulting importance scores (Bilodeau et al., 2024; Hooshyar & Yang, 2024; Huang & Marques-Silva, 2024; Kumar et al., 2021; Létoffé et al., 2025; Lones, 2024; Molnar et al., 2022; Wu, 2025). For instance, if a DSEM model inadvertently overfits a spurious interaction between two correlated predictors, SHAP will assign both features credit for that relationship, despite the possibility that this interaction lacks grounding in the underlying data-generating process. As a result, SHAP inherits the model's inductive biases—such as preferences for certain types of splits, heuristics for handling missing values, and various regularization settings—and presents them as "explanations." Therefore, it is essential to interpret SHAP attributions with an awareness of the DSEM model's known strengths and limitations.

To mitigate reliance on a single supervised model's internal logic and thereby move closer to uncovering true associations rather than artifacts of a particular predictive algorithm, this paper advocates a two-stage, model-agnostic pipeline. First, we apply unsupervised techniques such as feature agglomeration (which groups variables by similarity in high-dimensional space) and highly variable gene selection (which retains only those features exhibiting the greatest dispersion across samples). These methods reduce dimensionality and filter out noisy or redundant variables without referencing the outcome of interest. In the second stage, we compute pairwise Spearman rank correlations, together with their associated p-values, to capture nonlinear, monotonic relationships and to formally test their statistical significance. By combining a data-driven feature-reduction step with a robust, nonparametric association metric, we aim to generate hypotheses about genuine variable interdependencies that do not depend on the potentially biased mechanics of any single supervised learner.

There are substantial differences between the approaches proposed in this paper and the Li et al.'s pipeline (2025) that combines SHAP with a Dynamic Structural Equation Model (DSEM) to capture temporal variation. Li et al.'s pipeline remains fundamentally supervised—despite ensembling multiple machine learning models with SHAP—while DSEM is a linear, parametric framework. When the linearity and distributional assumptions of DSEM are violated by the nonlinear, nonparametric nature of real-world data, downstream metrics (including explained variance) can be distorted. Moreover, applying SHAP across multiple supervised models (for example, explain = SHAP(model) for model-1, model-2, …, model-n) can propagate and even amplify biases in feature importance if those models share misspecification or label leakage.

By contrast, the proposed method emphasizes unsupervised components—such as feature agglomeration and highly variable gene selection—that avoid label-driven bias and can outperform supervised methods when labels are unreliable or absent. These are complemented by non-target, supervised, nonlinear, and nonparametric diagnostics, such as Spearman's rank correlation with p-values, to provide robustness without imposing restrictive parametric assumptions.

The growing adoption of SHAP in agricultural studies reflects an urgent need for interpretable machine learning models capable of generating insights that inform sustainable practices. By effectively allocating credit to each feature based on its marginal contribution, SHAP transforms black-box predictors into transparent tools that enable researchers to draw meaningful connections between factors such as stand density and species composition and their influence on forest productivity forecasts. However, the integration of SHAP with Dynamic Structural Equation Modeling (DSEM), as demonstrated by Li et al. (2025), has limitations due to DSEM's linear parametric nature, which can distort outcomes when applied to nonlinear or nonparametric data. This distortion can lead SHAP to inherit and amplify biases in feature importances derived from DSEM. As a result, while the findings may uncover relationships among various stand structure factors and forest Net Primary Productivity (NPP), caution must be exercised in interpreting these insights. Overall, it is crucial to complement this methodology with robust validation techniques to enhance informed decision-making and develop a comprehensive understanding of how to optimize agricultural and forestry productivity sustainably over time.

SHAP's adherence to fairness axioms and its game-theoretic rigor do not insulate it from the biases and artifacts of the underlying supervised learner. Importance attributions are shaped by model choices such as tree depth, splitting criteria, and regularization settings, as well as data characteristics including feature correlation and imbalanced classes. Because there is no external ground truth for feature importance, SHAP values reflect the model's internal logic rather than objective causal effects. Small changes in training-test splits, hyperparameters, or random seeds can yield substantially different rankings, and spurious interactions discovered by an overfitted ensemble may be misrepresented by SHAP as genuine signals.

To address these concerns, we advocate employing a multi-faceted, model-agnostic strategy that complements SHAP with additional methods. First, unsupervised feature-reduction techniques such as agglomeration and highly variable gene selection help eliminate noise and redundancy independent of any target variable. Second, nonparametric association measures such as Spearman rank correlations paired with formal p-value testing facilitate the quantification of monotonic relationships without relying on supervised predictions. Lastly, robustness checks including sensitivity analyses, feature ranking stability tests to examine the consistency of remaining ranks after removing top features, and expert validation should accompany every SHAP-based interpretation. By integrating game-theoretic explanations with dimension reduction and rigorous statistical testing, researchers can reduce their dependence on the peculiarities of a single model and ultimately move closer to uncovering true variable interdependencies that support robust, data-driven policy.

SHAP estimates how features contribute to a model's predictions, but those attributions reflect the model's learned relationships rather than the true data-generating process. Decisions in forest management, crop monitoring, and climate adaptation require accurate, causal associations to guide action. Identifying which features most influence these decisions is therefore critical; if SHAP-based rankings are biased or misinterpreted, interventions may target the wrong levers, misallocate resources, and ultimately worsen outcomes. To address this, the paper complements SHAP with unsupervised structure discovery (e.g., feature agglomeration and highly variable gene selection) and non-target,

nonlinear, nonparametric diagnostics (e.g., Spearman's rank correlation with p-values), which together mitigate label-driven bias and improve the stability of feature rankings, whereas purely supervised pipelines are more prone to instability and can propagate misguidance into practice.

CRediT authorship contribution statement

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

Funding

This research has no fund.

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 7,222 in parallel algorithms. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

Declaration of competing interest

The author declares 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

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

References

Bilodeau, B., Jaques, N., Koh, P.W., Kim, B., 2024. Impossibility theorems for feature attribution. Proc. Natl. Acad. Sci. 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.

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.

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.

Li, T., Wu, Y., Ren, F., Tian, L., Li, M., 2025. Assessing the impact of stand structure on forest net primary productivity: a multiple machine learning-SHAP models and DSEM integrated approach. Comput. Electron. Agric. 236, 110427. https://doi.org/10.1016/j.compag.2025.110427.

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. https://doi.org/10.1007/978-3-031-04083-2_4.

Parr, T., Hamrick, J., Wilson, J.D., 2024. Nonparametric feature impact and importance. 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.