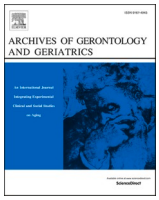
Archives of Gerontology and Geriatrics 135 (2025) 105873

Contents lists available at ScienceDirect

Archives of Gerontology and Geriatrics

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



Beyond SHAP: Reliable feature selection methods for clinical prediction models

ARTICLE INFO

Keywords:
Feature importance validation
Information theory
Monotonic relationships
Model-agnostic methods
Clinical prediction modeling

ABSTRACT

This study critically examines the limitations of model-dependent feature importance methods used in clinical prediction modeling, specifically addressing inconsistencies in Xu et al.'s (2025) depression prediction research. We demonstrate how algorithm selection fundamentally alters featured rankings despite similar prediction accuracies, revealing a methodological gap where accuracy validation exists but feature importance validation does not. We propose a comprehensive alternative framework combining statistical and information-theoretic approaches: (1) monotonic relationship detection using Spearman's correlation and Kendall's tau with p-value assessment, and (2) complex interaction analysis using Mutual Information and Effective Transfer Entropy. This dual methodology enables identification of both straightforward variable associations and complex nonlinear dependencies, providing more robust and reliable insights for clinical prediction models.

Dear editor,

Xu et al. (2025) investigated interpretable machine learning models for predicting depression in middle-aged and elderly Chinese arthritis patients through a nationwide prospective cohort study. Six machine learning algorithms, including XGBoost, logistic regression, KNN, decision tree, LightGBM, and random forest, were employed to develop depression risk prediction models for middle-aged and elderly arthritic individuals. The study incorporated demographic and clinical indicators alongside lifestyle variables and utilized the SHapley Additive exPlanations (SHAP) framework to enhance model interpretability. The final model achieved an AUC of 0.712, indicating relatively high predictive ability. The key predictors identified included age, life satisfaction, and comorbidities such as diabetes (Xu, 2025).

This paper raises critical concerns regarding the methodological approach to feature importance determination and the reliance on SHAP for interpretability in Xu et al.'s work. A fundamental limitation evident in their results is the model-specific nature of feature importances—different algorithms produced distinctly different feature rankings, highlighting a significant methodological challenge. This is clearly demonstrated in Xu et al.'s own findings, where their XGBoost model ranked lifestyle factors highly, while their logistic regression model emphasized demographic variables. Even more concerning, variables that appeared as top predictors in one model were nearly insignificant in others, despite all models achieving similar prediction accuracy. This inconsistency stems from the absence of ground truth values for feature importance validation, unlike prediction accuracy which can be directly validated against known outcomes. High target prediction accuracy does not guarantee reliable feature importances, as documented in over 100 peer-reviewed articles addressing this critical issue (Lipton, 2018; Fisher, 2019; Lenhof, 2024).

While supervised machine learning models have established mechanisms to validate prediction accuracy against ground truth values, no parallel validation mechanism exists for assessing the accuracy of feature importance rankings. Consequently, each model employs its own methodology for calculating feature importance, potentially leading to algorithm-specific biases rather than revealing true biological or clinical significance. This inconsistency fundamentally undermines the reliability of conclusions regarding which factors truly drive depression risk in the studied population, regardless of the sophistication of interpretability methods like SHAP that are applied post-hoc to these inherently inconsistent models.

Moreover, SHAP itself inherently amplifies existing biases in feature importances due to the function of explain=SHAP(model), meaning that any algorithmic biases in the underlying model are perpetuated and sometimes magnified in the resulting explanations. This cascading effect of model-specific biases through interpretability frameworks creates a false sense of scientific validity around feature importance rankings that may be entirely dependent on algorithmic choices rather than true causal relationships in the studied population (Bilodeau, 2024; Huang, 2024; Kumar 2021).

To accurately determine true associations between the target variable and features, this paper advocates for the use of nonlinear nonparametric methods grounded in information theory, such as Effective Transfer Entropy (ETE) (Li et al., 2024). Unlike existing machine learning models, these methods are better equipped to capture complex interactions and nonmonotonic patterns among multiple variables. Moreover, ETE provides directional information, enabling a deeper understanding of causal relationships, and offers a robust alternative for uncovering genuine connections that current models with SHAP explanations fail to address.

Researchers must grasp the fundamental principles of machine learning from a ground truth perspective. Applying linear methods to nonlinear data or using parametric models on nonparametric data can introduce significant biases, resulting in flawed outcomes and misleading conclusions. In cases where ground truth values are unavailable, it is essential for researchers to adopt multifaceted approaches to ensure the reliability and validity of their results and interpretations.

https://doi.org/10.1016/j.archger.2025.105873
Received 7 April 2025; Received in revised form 21 April 2025; Accepted 25 April 2025
Available online 26 April 2025

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

AI use

Not applicable.

Authors' contributions

Yoshiyasu Takefuji completed this research and wrote this article.

CRediT authorship contribution statement

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

Declaration of competing interest

The author has no conflict of interest.

Funding

This research has no fund.

Data availability

Not applicable

References

Bilodeau, B., Jaques, N., Koh, P. W., & Kim, B. (2024). Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences, 121*(2), Article 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. *Journal of Machine Learning Research, 20*, 177.

Huang, X., & Marques-Silva, J. (2024). On the failings of Shapley values for explainability. *International Journal of Approximate Reasoning, 171*, Article 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. *Advances in Neural Information Processing Systems, 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. *Briefings in Bioinformatics, 25*(5), bbae379. https://doi.org/10.1093/bib/bbae379

Li, W. X., Lin, Q. H., Zhang, C. Y., Han, Y., & Calhoun, V. D. (2024). A new transfer entropy method for measuring directed connectivity from complex-valued fMRI data. *Frontiers in Neuroscience, 18*, Article 1423014. https://doi.org/10.3389/fnins.2024.1423014

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

Xu, X., Huang, H.-Y., Wang, S.-Y., Tan, S.-Y., Chen, H.-H., Zhou, M.-M., & Qian, M.-J. (2025). Interpretable machine learning model for predicting depression in middle-aged and elderly Chinese arthritis patients: A nationwide prospective cohort study. *Archives of Gerontology and Geriatrics*. , Article 105810. https://doi.org/10.1016/j.archger.2025.105810

Yoshiyasu Takefuji[1]
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
E-mail address: takefuji@keio.jp.

[1] **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 highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.