Journal of Infection 91 (2025) 106575

Contents lists available at ScienceDirect

Journal of Infection

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

Letter to the Editor

Limitations of XGBoost-SHAP integration for interpretable machine learning in antimicrobial resistance prediction

Yuan et al. investigated machine learning applications for predicting antimicrobial resistance in Enterobacterales bloodstream infections.[1] Their research implemented XGBoost machine learning models to forecast resistance to seven antibiotics in bloodstream infection cases. Their study featured SHapley Additive exPlanations (SHAP) plots illustrating feature importance and their impacts on model output, particularly for amoxicillin resistance prediction at the time of blood culture sampling. Additional SHAP plots demonstrated how the time elapsed since the last resistant isolate influenced predictions of resistance to the same antibiotic.[1]

This paper's methodology reveals potential limitations in interpretation stemming from the combination of XGBoost with SHAP analysis. Yuan et al. appear to overlook certain fundamental machine learning principles critical to proper interpretation. While supervised machine learning models like XGBoost can be validated against ground truth values (labels) for prediction accuracy, the feature importances derived from these models lack corresponding ground truth for accuracy validation. It's essential to understand that feature importance in models such as XGBoost reflects contributions to prediction outcomes rather than true causal associations between variables. High prediction accuracy does not necessarily guarantee reliable feature importance rankings, as no ground truth exists to verify these attributions.[2–5]

The implementation approach (explain=SHAP(model)) indicates that SHAP analysis wholly depends on the underlying XGBoost model, inheriting and potentially amplifying any biases present in the model's feature importance calculations. This dependency can lead to misleading interpretations of variable relationships.[6–10] Further analysis demonstrates that feature ranking is inconsistent and unstable when top features are systematically removed, highlighting the model-specific nature of XGBoost's feature importance metrics rather than reflecting true biological or clinical relationships.

While SHAP is a powerful interpretability method, its application with XGBoost presents significant reliability concerns. The outputs from SHAP with XGBoost can be misleading because they fundamentally reflect model-specific feature utilization patterns rather than true causal mechanisms or biological relationships. When features are systematically removed from the model, the remaining feature importance rankings often shift dramatically, demonstrating instability in these attributions. This phenomenon occurs because XGBoost redistributes importance among available features rather than preserving consistent relationships between variables and outcomes. Furthermore, XGBoost's tendency to favor features with higher cardinality (more unique values) can artificially inflate the importance of certain variables regardless of their true predictive value. In clinical contexts such as antimicrobial resistance prediction, these algorithmic artifacts can lead to incorrect conclusions about which patient factors truly drive resistance patterns. In the absence of methods for accurately calculating true associations between variables, a more robust approach would involve a multifaceted framework using unsupervised machine learning models such as feature agglomeration (FA) and highly variable gene selection (HVGS) to complement SHAP, followed by nonlinear nonparametric statistical methods such as Spearman's correlation with p-values to identify monotonic relationships among variables. While FA, HVGS, and Spearman's correlation offer greater stability in feature importance rankings, SHAP analyses based solely on XGBoost inevitably suffer from instability in feature rankings that limit their clinical interpretability and applicability.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

This research has no funding.

Author contributions

Yoshiyasu Takefuji completed this research and wrote this article.

Code availability

Not applicable.

Data availability

Not applicable.

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.

https://doi.org/10.1016/j.jinf.2025.106575

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

Not applicable.

References

1. Yuan K, Luk A, Wei J, Walker AS, Zhu T, Eyre DW. *Machine learning and clinician predictions of antibiotic resistance in Enterobacterales bloodstream infections. J Infect* 2025;**90**(2):106388. https://doi.org/10.1016/j.jinf.2024.106388
2. Fisher A, Rudin C, Dominici F. *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* 2019;**20**:177.
3. Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. *Bias in artificial intelligence algorithms and recommendations for mitigation. PLoS Digit Health* 2023;**2**(6):e0000278. https://doi.org/10.1371/journal.pdig.0000278
4. Ugirumurera J, Bensen EA, Severino J, Sanyal J. *Addressing bias in bagging and boosting regression models. Sci Rep* 2024;**14**(1):18452. https://doi.org/10.1038/s41598-024-68907-5
5. Alaimo Di Loro P, Scacciatelli D, Tagliaferri G. *2-step Gradient Boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy. Stat Methods Appl* 2023;**32**:237–70. https://doi.org/10.1007/s10260-022-00643-4
6. Wu L. *A review of the transition from Shapley values and SHAP values to RGE. Statistics* 2025:1–23. https://doi.org/10.1080/02331888.2025.2487853
7. Bilodeau B, Jaques N, Koh PW, Kim B. *Impossibility theorems for feature attribution. Proc Natl Acad Sci USA* 2024;**121**(2):e2304406120. https://doi.org/10.1073/pnas.2304406120
8. Huang X, Marques-Silva J. *On the failings of Shapley values for explainability. Int J Approx Reason* 2024;**171**:109112. https://doi.org/10.1016/j.ijar.2023.109112
9. Hooshyar D, Yang Y. *Problems with SHAP and LIME in interpretable AI for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. IEEE Access* 2024;**12**:137472–90. https://doi.org/10.1109/ACCESS.2024.3463948
10. Lones MA. *Avoiding common machine learning pitfalls. Patterns* 2024;**5**(10):101046. https://doi.org/10.1016/j.patter.2024.101046

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 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.