Critical Analysis of Random Forest Feature Selection: Implications for Gut Microbiome Cancer Research

Dear Editors:

The choice of data analysis tools is pivotal for deriving accurate outcomes and robust conclusions. Without a thorough understanding of machine-learning principles, researchers risk falling into common pitfalls, leading to erroneous conclusions. Li et al[1] elucidated the role of gut archaea in colorectal cancer across diverse populations. In their study, feature selection was performed using the random forest-based Boruta package, v7.0.0 (Analytics Vidhya), with default settings to identify the optimal set of archaeal biomarkers. Subsequently, correlations among the "confirmed features" identified by Boruta were calculated, and only those with correlation coefficients below 0.7 were retained for constructing the predictive model.[1]

While recognizing the innovative models exploring the role of gut archaea in colorectal cancer, this paper raises significant concerns regarding the use of random forest for feature selection, primarily because of the model-specific nature that can lead to erroneous conclusions. Researchers, including Li et al, must grasp fundamental theoretical principles of machine learning. Although supervised machine learning allows for validation of target prediction accuracy through known ground-truth values, the calculation of feature importances lacks similar validation, resulting in potential inaccuracies.

Given that feature importance metrics lack established ground-truth values, different models often employ varying methodologies, which inherently introduces bias in their assessments. Li et al should consider 2 critical points: First, high predictive accuracy does not necessarily imply that the corresponding feature importances are reliable,[2,3] and, second, feature importance measures are intrinsically susceptible to bias.[2–5] More than 100 peer-reviewed studies have documented significant biases in feature importances derived from machine-learning models, including random forest.[2–5] Although the paper acknowledges high target prediction accuracy, it contends that the feature importances generated by random forest are fundamentally unreliable.

In the absence of definitive ground-truth values, it is essential to consider 3 key statistical components: the underlying data distribution, the nature of the relationships among variables, and the validation of statistical significance via *P* values. This paper recommends robust, nonlinear, and nonparametric methods that address these elements, including Spearman's correlation,[6] Kendall's tau,[7] Goodman-Kruskal gamma,[8] Somers' D,[8] and Hoeffding's D,[6] each supplemented by *P*-value assessments. Notably, although feature importances from random forest are confined to a 0 to 1 range, reflecting only the strength of association, measures such as Spearman's, Kendall's, Goodman-Kruskal gamma, and Somers' D span from −1 to 1, providing not only the magnitude but also the directional information of the association.

YOSHIYASU TAKEFUJI
Faculty of Data Science
Musashino University
Tokyo, Japan

References

1. Li T, et al. Gastroenterology 2025;168:525–538.e2.
2. Lipton ZC. Queue 2018;16:31–57.
3. Fisher A, et al. J Mach Learn Res 2019;20:177.
4. Nazer LH, et al. PLOS Digit Health 2023;2:e0000278.
5. Strobl C, et al. BMC Bioinformatics 2007;8:25.
6. Fujita A, et al. J Bioinform Comput Biol 2009;7:663–684.
7. Chen S, et al. R Soc Open Sci 2022;9:211346.
8. Metsämuuronen J. Behaviormetrika 2021;48:283–307.

Conflicts of interest
The author discloses no conflicts.

https://doi.org/10.1053/j.gastro.2025.03.043