Ann Rheum Dis 000 (2025) e1–e2

Contents lists available at ScienceDirect

Annals of the Rheumatic Diseases

journal homepage: https://www.sciencedirect.com/journal/annals-of-the-rheumatic-diseases

Correspondence

Evaluating the limitations of random forest and SHAP in predicting treatment responses in systemic lupus erythematosus

Bachali et al [1] explored the responsiveness of systemic lupus erythematosus (SLE) patients to iberdomide based on molecular endotypes. Their study employed a random forest (RF) model to predict treatment responses to the highest dose of iberdomide (0.45 mg once daily), utilising Shapley Additive Explanations (SHAP) analysis to identify significant factors influencing these predictions.

This paper raises significant concerns about the application of RF models in conjunction with SHAP, emphasising the model-specific nature of these methodologies, which can lead to erroneous conclusions. It is essential for Bachali et al [1] to understand the fundamental theoretical principles of machine learning. While supervised machine learning models such as RF require labelled training data (known as 'ground truth' or 'gold standard' values) to evaluate prediction accuracy of patient response to iberdomide, the feature importances derived from these models lack such reference standards for validation. Ground truth refers to information that is regarded as real or true, typically obtained through direct measurement, such as clinically validated SLE disease activity scores or molecular markers of treatment response. For example, in this study, the actual observed clinical responses of SLE patients to iberdomide treatment would constitute the ground truth data for evaluating the model's predictive performance.

This distinction is crucial; high prediction accuracy might indicate strong model performance but does not necessarily validate the reliability of the feature importances. Thus, a model can produce accurate predictions while still misrepresenting the significance of individual features.

The absence of universally accepted ground truth values means that different models utilise varied methodologies for calculating feature importances, which can introduce significant biases in the results. Numerous studies—over 100 peer-reviewed articles [2–6]—have documented that the feature importances derived from models are often severely biased, which can lead to erroneous conclusions about the data's underlying relationships, as detailed in the supplementary document accompanying this study. Furthermore, the use of the function `explain = SHAP(model)` implies that SHAP does not merely calculate feature importance but may also inherit—and potentially exacerbate—any biases present in the original model. This can result in an amplified distortion of feature importance, misleading researchers who rely on these measures for interpreting their results [7–9].

In light of these concerns, this paper advocates for the adoption of nonlinear and nonparametric robust statistical methods such as Spearman's correlation, Kendall's tau, Goodman-Kruskal gamma, Somers' D, and Hoeffding's D, each accompanied by *P* values to assess statistical significance. These methods offer a more reliable means of evaluating relationships between variables because they do not rely on the same assumptions as parametric methods and are less sensitive to outliers and skewed data distributions. Employing these robust statistical techniques could lead to more accurate interpretations of feature importance and treatment response, ultimately strengthening the conclusions drawn from studies of SLE patients and their responsiveness to iberdomide.

Our analytical framework deliberately employs multiple complementary approaches to enhance the validity and clinical utility of our findings. Rather than relying solely on correlation coefficients, we implement a multifaceted strategy that strengthens the robustness of our conclusions.

This comprehensive approach combines standardised correlation analyses with unstandardised effect measures, providing multiple lines of evidence that reinforce each other. For analysing SLE patient responses to iberdomide, we triangulate evidence using: (1) standardised correlation coefficients to assess relationship strengths, (2) unstandardised absolute differences in disease activity scores for direct clinical interpretation, (3) changes in molecular marker levels that maintain their biological meaning, and (4) cross-validation techniques to ensure reproducibility of findings.

By integrating these diverse analytical approaches, we not only enhance the validity of our findings but also provide clinically meaningful measures that directly translate to patient care. This methodological triangulation helps mitigate the limitations of any single analytical approach while offering complementary perspectives on the relationships between molecular endotypes and treatment response. Such a multifaceted strategy aligns with best practices in clinical research by providing both statistical rigour and practical clinical utility.

Competing interests

The author has no conflict of interest.

Handling editor Josef S. Smolen.

https://doi.org/10.1016/j.ard.2025.04.005
Received 17 March 2025; Revised 1 April 2025; Accepted 2 April 2025

Contributors

YT completed this research and wrote this article.

Funding

This research had no funding.

Patient consent for publication

Not applicable.

Ethics approval

Not applicable.

Provenance and peer review

Peer-reviewed, revised and accepted.

Data availability statement

Not applicable.

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

Not applicable.

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

Orcid

Yoshiyasu Takefuji: http://orcid.org/0000-0002-1826-742X

REFERENCES

[1] Bachali P, Daamen A, Korish S, Hu Y, Schafer P, Grammer A, et al. Responsiveness of systemic lupus erythematosus subjects to iberdomide based on molecular endotypes. Ann Rheum Dis. 2025 ISSN 0003-4967. doi: 10.1016/j.ard.2025.01.044.

[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] Steiner PM, Kim Y. The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases. J Causal Inference 2016;4(2):20160009. doi: 10.1515/jci-2016-0009.

[4] Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007;8:25. doi: 10.1186/1471-2105-8-25.

[5] Salles T, Rocha L, Gonçalves M. A bias-variance analysis of state-of-the-art random forest text classifiers. Adv Data Anal Classif 2021;15:379–405. doi: 10.1007/s11634-020-00409-4.

[6] Huti M, Lee T, Sawyer E, King AP. An investigation into race bias in random forest models based on breast DCE-MRI derived radiomics features. Clin Image Based Proced Fairness AI Med Imaging Ethical Philos Issues Med Imaging (2023) 2023;14242:225–34. doi: 10.1007/978-3-031-45249-9_22.

[7] Bilodeau B, Jaques N, Koh PW, Kim B. Impossibility theorems for feature attribution. Proc Natl Acad Sci U S A 2024;121(2):e2304406120. doi: 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. doi: 10.1016/j.ijar.2023.109112.

[9] Lones MA. Avoiding common machine learning pitfalls. Patterns (N Y) 2024;5 (10):101046. doi: 10.1016/j.patter.2024.101046.

Yoshiyasu Takefuji *

Faculty of Data Science, Musashino University, Tokyo, Japan

*Correspondence to Prof Yoshiyasu Takefuji, Faculty of Data Science, Musashino University, Tokyo, Japan.

E-mail address: takefuji@keio.jp