International Immunopharmacology 163 (2025) 115238

Contents lists available at ScienceDirect

International Immunopharmacology

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

Addressing bias in biomarker discovery for inflammatory bowel diseases: A multi-faceted analytical approach

ARTICLE INFO

Editor Name: Dr. J.-P. Gorvel
Keywords:
Inflammatory bowel disease
Biomarker discovery
Machine learning
Deep learning
Feature importance

ABSTRACT

Xiang-Guang et al. investigate the identification of novel biomarkers linked to M1 macrophage infiltration in inflammatory bowel diseases (IBD). Utilizing advanced bioinformatics and machine learning techniques, the researchers developed predictive models and employed the SHAP algorithm to assess feature importance, revealing that the top ten features corresponded exclusively to host genes. However, significant concerns regarding the model-specific nature of SHAP assessments raise doubts about the reliability of feature importance. To address these issues, we advocate for a multifaceted approach combining feature agglomeration (FA), highly variable gene selection (HVGS), and Spearman's correlation for a more accurate analysis. This integrated methodology aims to enhance our understanding of biological factors in IBD and improve diagnostic and therapeutic strategies.

Xiang-Guang et al. conducted an investigation into identifying novel biomarkers associated with M1 macrophage infiltration for the diagnosis of inflammatory bowel diseases (IBD) [1]. They developed predictive models utilizing advanced bioinformatics and deep learning techniques to pinpoint potential biomarkers that could significantly enhance management outcomes for patients with IBD. By employing the SHAP (Shapley Additive Explanations) algorithm to evaluate feature importance within their models, they found that the top ten identified features were exclusively host genes, revealing an intriguing focus on the host's biological makeup rather than direct microbial influences [1]. This finding underscores the potential for purely genetic biomarkers in understanding and diagnosing IBD, but also highlights the need for further research to explore the underlying mechanisms of host-microbe interactions in this context.

However, this paper raises significant theoretical and empirical concerns regarding the employment of deep learning alongside SHAP for assessing feature importance, primarily due to the model-specific nature of these assessments. Such an approach can lead to erroneous interpretations and misguided conclusions. In supervised machine learning models, including deep learning, target prediction accuracy can be validated against known ground truth values. In contrast, the feature importance extracted from these models lacks a comparable ground truth for accuracy validation. Thus, while a model may exhibit high predictive accuracy, the reliability of its identified feature importance remains questionable. This distinction between target prediction accuracy and feature importance reliability is crucial, as it is possible for different models to yield disparate feature importance results without any true underlying association between the variables. Over 300 peer-reviewed articles documented this non-negligible biases in feature importances derived from models [2–8].

The function explain = SHAP(model) indicates that SHAP relies heavily on the underlying model, inheriting and potentially amplifying existing biases in feature importance calculations [9–16]. As a result, the feature importances produced may reflect not only the genuine contributions of each feature to predictions but also the idiosyncrasies and biases of the particular model used. This highlights a critical limitation: while SHAP can provide insights into how features contribute to predictions, it does not necessarily indicate the true causal relationships or associations among the variables, warranting caution in the interpretation of its results. Thus, researchers must be vigilant in differentiating between predictive utility and actual biological relevance when utilizing SHAP for feature importance analysis in deep learning models.

In light of these concerns, this paper advocates for the use of multifaceted approaches that incorporate unsupervised machine learning techniques, such as feature agglomeration (FA) and highly variable gene selection (HVGS). Feature agglomeration is a method used to group similar features together, which helps reduce dimensionality and enhance model performance by eliminating redundancy. This can lead to more robust models by focusing on composite features that capture essential biological signals without being influenced by noise. Meanwhile, highly variable gene selection focuses on identifying genes that exhibit substantial variability across samples, which can serve as significant indicators of biological or pathological states. This approach is crucial in contexts such as IBD, where specific gene expressions may correlate with disease severity or response to treatment.

Furthermore, following the identification of these important features, utilizing nonlinear nonparametric statistical methods, such as Spearman's correlation, can enhance the analysis. Spearman's correlation assesses the strength and direction of the association between ranked variables, making it particularly suitable for biological data where relationships may not be linear. This method does not assume a normal distribution of the data, allowing for a more flexible analysis that can uncover meaningful connections between features in a way that normal parametric tests might miss. By implementing these robust analytical strategies, researchers can derive more reliable insights into the biological factors at play in IBD, ultimately paving the way for more

https://doi.org/10.1016/j.intimp.2025.115238
Received 13 July 2025; Accepted 13 July 2025

effective diagnostics and treatment options.

AI use

Not applicable.

Authors' contribution

Yoshiyasu Takefuji completed this research and wrote this article.

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

CRediT authorship contribution statement

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

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

Code availability

Not applicable.

Funding

This research has no fund.

Declaration of competing interest

The author declare 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

[1] Xiang-Guang Li, Huantao Li, Ding Luo, et al., Identification of novel biomarkers linked to M1 macrophage infiltration in the diagnosis of inflammatory bowel diseases, Int. Immunopharmacol. 162 (2025) 115138, https://doi.org/10.1016/j.intimp.2025.115138.
[2] Z.C. Lipton, The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57, https://doi.org/10.1145/3236386.3241340.
[3] A. Fisher, C. Rudin, F. Dominici, 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 (2019) 177.
[4] K. Lenhof, L. Eckhart, L.M. Rolli, H.P. Lenhof, 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 (2024), https://doi.org/10.1093/bib/bbae379.
[5] H. Mandler, B. Weigand, A review and benchmark of feature importance methods for neural networks, ACM Comput. Surv. 56 (12) (2024) 318, https://doi.org/10.1145/3679012.
[6] J.L. Potharlanka, M NB., Feature importance feedback with deep Q process in ensemble-based metaheuristic feature selection algorithms, Sci. Rep. 14 (2024) 2923, https://doi.org/10.1038/s41598-024-53141-w.
[7] D. Wood, T. Papamarkou, M. Benatan, et al., Model-agnostic variable importance for predictive uncertainty: an entropy-based approach, Data Min. Knowl. Disc. 38 (2024) 4184–4216, https://doi.org/10.1007/s10618-024-01070-7.
[8] L.H. Nazer, R. Zatarah, S. Waldrip, J.X.C. Ke, M. Moukheiber, A.K. Khanna, et al., Bias in artificial intelligence algorithms and recommendations for mitigation, PLOS Digit Health. 2 (6) (2023) e0000278, https://doi.org/10.1371/journal.pdig.0000278.
[9] L. Wu, 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.
[10] B. Bilodeau, N. Jaques, P.W. Koh, B. Kim, Impossibility theorems for feature attribution, Proc. Natl. Acad. Sci. USA 121 (2) (2024) e2304406120, https://doi.org/10.1073/pnas.2304406120.
[11] X. Huang, J. Marques-Silva, On the failings of Shapley values for explainability, Int. J. Approx. Reason. 171 (2024) 109112, https://doi.org/10.1016/j.ijar.2023.109112.
[12] D. Hooshyar, Y. Yang, 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 (2024) 137472–137490, https://doi.org/10.1109/ACCESS.2024.3463948.
[13] M.A. Lones, Avoiding common machine learning pitfalls, Patterns 5 (10) (2024) 101046, https://doi.org/10.1016/j.patter.2024.101046.
[14] C. Molnar, et al., General pitfalls of model-agnostic interpretation methods for machine learning models, in: A. Holzinger, R. Goebel, R. Fong, T. Moon, K. R. Müller, W. Samek (Eds.), xxAI - beyond Explainable AI. xxAI 2020 vol. 13200, Springer, 2022, https://doi.org/10.1007/978-3-031-04083-2_4. Lecture Notes in Computer Science.
[15] I. Kumar, C. Scheidegger, S. Venkatasubramanian, S. Friedler, Shapley residuals: quantifying the limits of the shapley value for explanations, Adv. Neural Inf. Proces. Syst. 34 (2021) 26598–26608.
[16] O. Létoffé, X. Huang, J. Marques-Silva, Towards trustable SHAP scores, Proceedings of the AAAI Conference on Artificial Intelligence 39 (17) (2025) 18198–18208, https://doi.org/10.1609/aaai.v39i17.34002.

Yoshiyasu Takefuji [*,1]
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

[*] Corresponding author.
E-mail address: takefuji@keio.jp.

[1] ORCID: 0000-0002-1826-742X