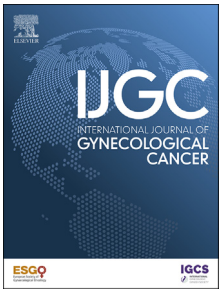
LETTER

Correspondence on "Large-scale analysis to identify risk factors for ovarian cancer" by Madakkatel et al



Received 23 June 2025, Accepted 26 June 2025; Available online xxx

Madakkatel and colleagues[1] (2025) identified ovarian cancer predictors from 2920 features using a CatBoost Gradient Boosting Decision Tree (GBDT) and SHapley Additive exPlanations (SHAP) values. However, their methodology raises concerns about model bias and the reliability of feature importance. It is a common misconception that even highly accurate machine learning models like GBDT, despite strong predictive performance, reliably yield trustworthy feature rankings. This principle is widely recognized, and validated by over 300 peer-reviewed articles (details in Supplementary Material). Boosting algorithms often produce biased importance scores by overemphasizing features used in early splits.[2] SHAP values, used for explainability, inherit and may amplify these biases.[3] This "GBDT-SHAP pipeline" distorts interpretation, increasing the risk that features reflect model artifacts, not true causal drivers. Moreover, validating feature importance is inherently difficult due to the absence of ground truth, undermining the stability and reproducibility of findings. To address these concerns, we advocate for multi-faceted analytical strategies that combine machine learning with robust statistical validation methods. For instance, approaches like highly variable gene selection and feature agglomeration can better capture complex, latent patterns in the data.[4,5] Integrating such methods alongside machine learning is essential for generating reliable insights in health risk assessment, ensuring findings are not only predictive but also interpretable and reproducible.

Funding/Support This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions Writing – original draft, Conceptualization: SO. Writing – review & editing, Supervision, Project administration: YT

Declaration of Competing Interests None declared.

Data Availability No new data were generated or analyzed in support of this research.

Supplemental Material Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijgc.2025.102000.

Souichi Oka[a,*] **Yoshiyasu Takefuji**[b]

[a]*Research and Development Planning Department, Science Park Corporation, Kanagawa, Japan*
[b]*Musashino University, Faculty of Data Science, Tokyo, Japan*

[*]***Correspondence to*** *Dr Souichi Oka, Science Park Corporation, 3-24-9 Iriya-Nishi Zama-shi, Kanagawa, Japan.*
E-mail: souichi.oka@sciencepark.co.jp (S. Oka)

REFERENCES

1. Madakkatel I, Lumsden AL, Mulugeta A, Mäenpää J, Oehler MK, Hyppönen E. Large-scale analysis to identify risk factors for ovarian cancer. *Int J Gynecol Cancer*. 2025. Published online January 6. https://doi.org/10.1136/ijgc-2024-005424.
2. Bilodeau B, Jaques N, Koh PW, Kim B. Impossibility theorems for feature attribution. *Proc Natl Acad Sci U S A*. 2024;121(2):e2304406120. https://doi.org/10.1073/pnas.2304406120.
3. 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.
4. Arora JK, Opasawatchai A, Teichmann SA, Matangkasombut P, Charoensawan V. Computational workflow for investigating highly variable genes in single-cell RNA-seq across multiple time points and cell types. *Star Protoc*. 2023;4(3):102387. https://doi.org/10.1016/j.xpro.2023.102387.
5. Zhang J, Wu X, Hoi SCH, Zhu J. Feature agglomeration networks for single stage face detection. *Neurocomputing*. 2020;380:180–189. https://doi.org/10.1016/j.neucom.2019.10.087.

DOI of original article: https://doi.org/10.1136/ijgc-2024-005424.

https://doi.org/10.1016/j.ijgc.2025.102000