



Letter to the Editor

Letter Re: Development of an artificial intelligence-generated, explainable treatment recommendation system for urothelial carcinoma and renal cell carcinoma to support multidisciplinary cancer conferences

ARTICLE INFO

Keywords:

Clinical oncology
Machine Learning
Feature importance
Model bias
Multi-faceted statistical validation

To the Editor,

Duwe et al. presents a promising AI-based system for generating explainable treatment recommendations in urothelial and renal cell carcinomas, aimed at supporting decision-making in multidisciplinary cancer conferences [1]. While their approach demonstrates strong predictive performance, it raises important concerns regarding the reliability of model interpretability that warrant further discussion. Their system utilized various machine learning (ML) and deep learning techniques to train a classifier to mimic treatment recommendations (TR), achieving excellent F1-scores for both urothelial carcinoma (UC) and renal cell carcinoma (RCC) across different treatment categories (e.g., UC: 'Surgery' 0.81, 'Anti-cancer drug' 0.83, 'Gemcitabine/Cisplatin' 0.88; RCC: 'Anti-cancer drug' 0.92, 'Nivolumab' 0.78, 'Pembrolizumab/Axitinib' 0.89). While the study offers valuable insights and provided explainability through visualized clinical feature importance scores, it also presents two key methodological concerns that require further analysis.

First, the feature importance rankings varied substantially across models (XGBoost, CatBoost, Random Forest, SoftOrdering CNN), suggesting model-specific biases in how clinical features are interpreted. This inconsistency raises concerns about the stability and generalizability of the system's explanations. Second, although predictive accuracy was high, it is essential to recognize that accuracy alone does not validate the reliability of feature attribution. Prediction performance and feature importance are two aspects that are conceptually distinct. As supported by over 300 peer-reviewed articles, strong predictive metrics do not necessarily imply trustworthy or consistent feature importance rankings [2,3]. A more detailed discussion and supporting references are provided in the supplementary material.

XGBoost, Catboost and Random Forest, like other tree-based models, exhibit inherent biases in feature importance due to their tree-building process, which can overemphasize features used in earlier splits [4]. This could lead to a skewed perception of clinical factors' importance.

Deep learning models like TabPFN, TabNet, SoftOrdering CNN, and FCN also show significant biases. This is largely because of their complex architecture and a tendency to overfit when optimized for high predictive accuracy [5]. In clinical datasets, which are often noisy and high-dimensional, this overfitting can lead models to capture spurious patterns rather than meaningful clinical signals, resulting in biased and potentially misleading importance scores.

The findings in this study raise serious doubts about the central claim made by Duwe et al. that their AI system offers both high accuracy and interpretability in treatment recommendations. While the reported F1-scores indicate strong predictive performance, the inconsistent feature importance across models reveals a lack of reliability in identifying key clinical factors. This instability undermines the credibility of the system's explanations. If different models highlight conflicting aspects of the data due to inherent biases, then the rationale behind treatment decisions may reflect model-specific artifacts rather than clinically meaningful insights. Without rigorous validation of feature attribution across architectures, the system risks producing explanations that are not only inconsistent but potentially misleading.

Additionally, SHAP values, while intended to elucidate feature importance, inherit and may even exacerbate biases from the underlying machine learning model [6]. The function of 'explain = SHAP(model)' underscores this dependency. Since SHAP relies on the model's output for its explanations, it is inherently vulnerable to the model's biases. This can lead to flawed interpretations and undermine the reliability of the analysis. Furthermore, although machine learning models often prioritize predictive accuracy, achieving high accuracy does not ensure the reliability of derived feature importances.

A major challenge in validating feature importance is the lack of ground truth. Different models rely on distinct methodologies, which introduce model-specific biases and lead to inconsistent rankings. This issue is particularly evident in Duwe et al.'s study, which employed complex and high-dimensional feature sets—77 patient input

DOI of original article: <https://doi.org/10.1016/j.ejca.2025.115367>.

<https://doi.org/10.1016/j.ejca.2025.115733>

Received 11 July 2025; Accepted 13 August 2025

Available online 20 August 2025

0959-8049/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.



parameters for UC and 76 for RCC. Such complexity, combined with potential collinearity, impairs interpretability and increases the risk of overfitting. In clinical research, where data are inherently noisy and intricate, models may capture spurious patterns rather than meaningful signals, resulting in unreliable importance scores. Furthermore, the sensitivity of feature importance to small changes in data or model configurations undermines its stability, posing serious challenges to reproducibility and clinical credibility.

To overcome methodological limitations and improve the reliability of health risk assessments, a more robust and multi-dimensional analytical framework is essential. This framework should reflect the complexity of clinical data and incorporate methods capable of capturing non-linear relationships. Unsupervised techniques such as Feature Agglomeration (FA) and, where applicable, Highly Variable Gene Selection (HVGS) [7,8] offer valuable alternatives. In addition, non-parametric statistical methods like Spearman's rho and Kendall's tau [9,10] can detect monotonic associations without assuming linearity, enhancing both precision and interpretability. These approaches are particularly useful in translational biomarker research, where clear and trustworthy insights must inform clinical decisions. Their interpretability also facilitates communication across diverse healthcare stakeholders, helping translate statistical findings into actionable outcomes. Ultimately, integrating these methods is key to generating insights that are accurate, reproducible, and clinically meaningful.

In conclusion, while machine learning methods such as ensemble models, deep learning, and SHAP are powerful for treatment recommendation, they carry biases that limit interpretability. In clinical oncology, where decisions must be both accurate and transparent, integrating rigorous statistical methods is essential. A multi-faceted approach that combines predictive strength with interpretive clarity marks a foundational shift in how AI supports clinical decision-making.

CRedit authorship contribution statement

Yoshiyasu Takefuji: Project administration, Supervision, Writing – review & editing. **Souichi Oka:** Conceptualization, Writing – original draft. **Takuma Yamazaki:** Investigation.

Funding

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

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.

Appendix A. Supporting information



Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.ejca.2025.115733](https://doi.org/10.1016/j.ejca.2025.115733).


Data Availability

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

References

- [1] Duwe G, Mercier D, Kauth V, et al. Development of an artificial intelligence-generated, explainable treatment recommendation system for urothelial carcinoma and renal cell carcinoma to support multidisciplinary cancer conferences. *Eur J Cancer* 2025;220:115367. <https://doi.org/10.1016/j.ejca.2025.115367>.
- [2] Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *ACM Queue* 2018;16(3):31–57. <https://doi.org/10.1145/3236386.3241340>.
- [3] Musolf AM, et al. What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. *Hum Genet* 2022;141(9):1515–28. <https://doi.org/10.1007/s00439-021-02402-z>.
- [4] Ugirumura J, Bensen EA, Severino J, Sanyal J. Addressing bias in bagging and boosting regression models. *Sci Rep* 2024;14:18452. <https://doi.org/10.1038/s41598-024-68907-5>.
- [5] 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. <https://doi.org/10.48550/arXiv.1801.01489>.
- [6] 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>.
- [7] Zhang J, Wu X, Hoi SCH, Zhu J. Feature agglomeration networks for single stage face detection. *Neurocomputing* 2020;380:180–9. <https://doi.org/10.1016/j.neucom.2019.10.087>.
- [8] Xie Y, Jing Z, Pan H, et al. Redefining the high variable genes by optimized LOESS regression with positive ratio. *BMC Bioinforma* 2025;26:104. <https://doi.org/10.1186/s12859-025-06112-5>.
- [9] Yu H, Hutson AD. A robust spearman correlation coefficient permutation test. *Commun Stat Theory Methods* 2024;53:2141–53. <https://doi.org/10.1080/03610926.2022.2121144>.
- [10] Okoye K, Hosseini S. Correlation tests in R: Pearson cor, Kendall's tau, and Spearman's rho. In: Okoye K, Hosseini S, editors. *R Programming: statistical data analysis in research*. New York: Springer Nature; 2024. p. 247–77. https://doi.org/10.1007/978-981-97-3385-9_12.

Souichi Oka^{*} , Takuma Yamazaki 
Science Park Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa 252-0029, Japan

Yoshiyasu Takefuji 
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
E-mail address: takefuji@keio.jp.

^{*} Corresponding author.
E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka),
tyamazaki@sciencepark.co.jp (T. Yamazaki).