Letter to the Editor

JOURNAL OF HEPATOLOGY

Reevaluating feature selection in machine learning-based radiomics for hepatocellular carcinoma: Bridging the gap between predictive accuracy and biological relevance

To the Editor:

Vithayathil *et al.* conducted a groundbreaking investigation demonstrating that machine learning-based radiomic models significantly outperform traditional clinical biomarkers in predicting immunotherapy outcomes for hepatocellular carcinoma.[1] Their methodologically sophisticated approach evaluated seven distinct machine learning algorithms (logistic regression, naïve Bayes, neural network, random forest, support vector machine, XGBoost, and ridge regression) in combination with thirteen diverse feature selection techniques, including regression-based methods (LASSO, elastic net), dimensional reduction approaches (PCA), feature importance algorithms (Boruta, RFE), and statistical correlation methods (Mutual Information (MI), Pearson, Spearman, Kendall, ANOVA F-test, variance threshold). Model optimization was rigorously performed using grid-search with 5-fold cross-validation to identify optimal hyper-parameters, establishing a comprehensive framework for radiomic-based prediction that substantially advances precision medicine approaches in liver cancer immunotherapy.[1]

While this paper acknowledges the powerful prediction models developed by Vithayathil *et al.*, it raises fundamental epistemological concerns regarding the use of machine learning for feature selection. The model-specific nature of feature selection algorithms can lead to potentially erroneous interpretations and conclusions about biological significance. A critical distinction exists in validation methodology: supervised machine learning models can be validated against established ground truth values for prediction accuracy, whereas feature importance rankings and selections lack comparable ground truth validation mechanisms. This methodological gap creates a significant risk: high prediction accuracy may create a false impression of reliable feature selection, when in reality, no objective standard exists to confirm these selections represent true biological relationships.

Vithayathil *et al.* must recognize the fundamental distinction between predictive contribution and biological association – feature importance metrics primarily quantify contributions to the prediction mechanism rather than establishing true causal or associative relationships between features and outcomes. This epistemological gap creates a significant interpretive hazard in radiomics research. Feature importance rankings and selections derived from machine learning models are inherently susceptible to model-specific biases and algorithmic assumptions that cannot be validated against objective biological standards. Over 100 peer-reviewed articles documented non-negligible errors in feature importances and selections.[2–7]

Furthermore, dimension reduction techniques like PCA operate exclusively on feature distributions and internal relationships among features, without consideration of target-feature associations, potentially obscuring rather than revealing clinically meaningful relationships in the radiomic data, regardless of their statistical efficiency.[8,9]

Pearson's correlation, ANOVA F-test, and variance threshold are linear parametric methods that may fundamentally distort analyses when applied to complex non-linear non-parametric biological data that rarely conform to linearity assumptions. These methods can systematically underrepresent or entirely miss non-linear relationships prevalent in biomedical data. In contrast, MI, Spearman's correlation, and Kendall's tau offer non-linear non-parametric approaches better suited to biological complexity. This paper specifically advocates for a complementary analytical framework employing Spearman and Kendall correlations[10] to detect monotonic pairwise relationships between targets and features, supplemented by MI analysis to identify nonmonotonic complex relationships among multiple variables. This methodological triangulation provides more robust validation than any single feature selection approach, particularly in clinical contexts where biological interpretation is paramount.

Yoshiyasu Takefuji*

Musashino University, Data Science Department, Tokyo, Japan

*Corresponding author. Address: 3-3-3 Ariake, Koto-Ku, Tokyo 135-8181, Japan.

E-mail address: takefuji@keio.jp

Received 15 May 2025; Accepted 19 May 2025; Available online xxx

https://doi.org/10.1016/j.jhep.2025.05.025

Financial support

The authors did not receive any financial support to produce this manuscript.

Conflict of interest

The author has no conflict of interest.

Please refer to the accompanying ICMJE disclosure forms for further details.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhep.2025.05.025.

References

[1] Vithayathil M, Koku D, Campani C, et al. Machine learning based radiomic models outperform clinical biomarkers in predicting outcomes after immunotherapy for hepatocellular carcinoma. J Hepatol 2025. https://doi.org/10.1016/j.jhep.2025.04.017. [Epub ahead of print].

[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] Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health 2023;2(6):e0000278. https://doi.org/10.1371/journal.pdig.0000278.

[4] Ugirumurera J, Bensen EA, Severino J, et al. Addressing bias in bagging and boosting regression models. Sci Rep 2024;14(1):18452. https://doi.org/10.1038/s41598-024-68907-5.

[5] Adler AI, Painsky A. Feature importance in gradient boosting trees with cross-validation feature selection. Entropy (Basel) 2022;24(5):687. https://doi.org/10.3390/e24050687.

[6] Özkale MR. Iterative algorithms of biased estimation methods in binary logistic regression. Stat Pap 2016;57:991–1016. https://doi.org/10.1007/s00362-016-0780-9.

[7] Wallace ML, Mentch L, Wheeler BJ, et al. Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. BMC Med Res Methodol 2023;23(1):144. https://doi.org/10.1186/s12874-023-01965-x.

[8] Dyer EL, Kording K. Why the simplest explanation isn't always the best. Proc Natl Acad Sci U S A 2023;120(52):e2319169120. https://doi.org/10.1073/pnas.2319169120.

[9] Yao Y, Ochoa A. Limitations of principal components in quantitative genetic association models for human studies. eLife 2023;12:e79238. https://doi.org/10.7554/eLife.79238.

[10] Okoye K, Hosseini S. Correlation tests in R: Pearson cor, Kendall's tau, and Spearman's Rho. In: R Programming. Singapore: Springer; 2024. https://doi.org/10.1007/978-981-97-3385-9_12.