European Journal of Radiology 191 (2025) 112308

Contents lists available at ScienceDirect

European Journal of Radiology

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

Correspondence

Complementing interpretable machine learning with synergistic analytical strategies for thyroid cancer recurrence prediction

ARTICLE INFO

Keywords:
Thyroid Cancer Recurrence Prediction
Machine Learning
XGBoost
SHapley Additive exPlanations
Feature importance
Statistical validation

ABSTRACT

This correspondence critically examines the methodology of Schindele et al. (2025) on thyroid cancer recurrence prediction. While their interpretable XGBoost model achieved a high predictive accuracy of 95.8% and a 0.947 AUROC, it is crucial to recognize that this predictive power does not justify the reliability of its derived feature importance rankings. As widely acknowledged in the literature, high predictive accuracy does not guarantee unbiased or reliable feature attribution. We underscore that gradient boosting decision tree (GBDT) models, including XGBoost, are prone to inherent biases in feature importance estimation, often due to overfitting. Furthermore, SHapley Additive exPlanations (SHAP), a widely adopted explainable AI (XAI) technique, can inherit and even amplify these biases, given its model-dependent nature. This raises concerns about the interpretive validity of the identified risk factors. To mitigate these methodological limitations, we advocate for integrative analytical frameworks that combine machine learning with robust statistical and non-parametric approaches, such as Highly Variable Feature Selection (HVFS) and Independent Component Analysis (ICA). These multi-faceted strategies are indispensable for obtaining robust and interpretable insights into feature importance, warranting their prioritization in future research efforts.

1. Letter to the Editor

Schindele et al. (2025) present an interpretable XGBoost model for predicting thyroid cancer recurrence. However, their reliance on a single boosting algorithm and the subsequent interpretation of feature importance warrants further discussion [1]. They employed a data-driven machine learning (ML) approach, identifying predictors of differentiated thyroid cancer (DTC) recurrence from 114 clinical and biomarker features in a large patient cohort of 2,920 individuals. Specifically, they utilized an extreme gradient boosting (XGBoost) model as their primary ML algorithm, which achieved a high predictive accuracy of 95.8 % and an AUROC of 0.947 on the testing dataset. Subsequently, for factors that contributed to model prediction, SHapley Additive exPlanation (SHAP) values were used to interpret feature importance. Their analysis identified several categories of important features for thyroid cancer recurrence prediction, including tumor size, maximal thyroglobulin values, and maximal thyroglobulin antibody levels. While their approach represents a powerful and commonly adopted strategy for prediction, it raises concerns regarding the inherent biases of the chosen ML model and their subsequent impact on feature importance interpretation.

Although Schindele et al. (2025) have made a significant contribution through their large-scale, data-driven ML analysis to identify thyroid cancer recurrence risk factors, their reliance on boosting algorithms like XGBoost raises critical methodological concerns. While models such as XGBoost are widely adopted for their high predictive accuracy, it is crucial to recognize that this performance does not inherently validate the reliability of the derived feature importance. This disconnect is well-documented in the literature, with over 300 peer-reviewed studies supporting the assertion that predictive accuracy alone is insufficient for trustworthy feature attribution (see Supplementary Material).

XGBoost, an implementation of gradient boosting decision trees, chosen by the authors, like other tree-based models, exhibits inherent biases in feature importance calculations due to its tree building process. While XGBoost offers strong predictive power, it often generates feature importance scores that overemphasize features used in earlier splits [2–4]. Additionally, these scores are influenced by factors such as the model's splitting logic, its handling of feature interactions, and multicollinearity [5–7]. This model-dependent nature suggests that the ranked features primarily reflect what is most advantageous for optimizing the XGBoost model's predictive performance, rather than serving as genuine indicators of the underlying causal drivers of thyroid cancer recurrence risk. This can lead to a skewed interpretation of factor importance, further highlighted by the frequent observation that different ML models produce conflicting rankings of predictive features.

Additionally, SHAP values, a widely employed eXplainable AI (XAI) method to interpret ML-derived feature importances, inherit and exacerbate biases from the underlying ML model [8–12]. 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, leading to flawed interpretations and undermining the reliability of the analysis. Furthermore, Schindele et al.'s reliance on an XGBoost-SHAP pipeline, combining two inherently biased methods, represents a critical and frequently encountered pitfall. The claim that this pipeline successfully identified predictive features, even with nonlinearity and interactions considered, demands rigorous scrutiny. For this reason, the compounded biases from both the XGBoost model and the SHAP explanation method can severely exacerbate interpretability issues.

Fundamentally, validating feature importance is exceptionally

https://doi.org/10.1016/j.ejrad.2025.112308
Received 4 July 2025; Accepted 8 July 2025
Available online 11 July 2025

challenging due to the absence of ground truth values, inevitably introducing model-specific biases and yielding inconsistent rankings. This issue is particularly evident in Schindele et al.'s study, where their complex feature sets, high dimensionality, and potential collinearity profoundly impede ML model interpretation, especially for thyroid cancer recurrence risk assessment. These factors severely complicate isolating individual feature effects, simultaneously elevating overfitting risk and causing models to capture noise instead of genuine signals. Furthermore, the complexity of features renders importance measures highly sensitive to minor data or model changes, compromising stability and reliability, thereby hindering consistent and reproducible research findings.

To address these methodological pitfalls and ensure more accurate interpretations in health risk assessment, a more robust, multi-faceted analytical framework is indispensable. Such an approach should account for the multifaceted nature of complex health data and incorporate methodologies better suited to capturing non-linearity. More methodologically robust approaches include unsupervised learning techniques like Feature Agglomeration (FA) or, where applicable, Highly Variable Gene Selection (HVGS) [13,14]. Additionally, non-linear non-parametric statistical methods such as Spearman's rho or Kendall's tau would be highly beneficial [15,16]. These methods specifically detect monotonic relationships without imposing linearity assumptions, thereby capturing potentially non-linear associations with greater precision and reliability. Beyond their statistical appropriateness, such non-parametric approaches offer enhanced interpretability—a critical consideration in translational biomarker research where findings must guide clinical decision-making. This interpretability advantage proves particularly valuable when communicating complex relationships to diverse stakeholders across the healthcare continuum, facilitating more effective translation from statistical findings to actionable clinical insights. Ultimately, this multi-faceted approach is indispensable for generating accurate, reproducible, and clinically meaningful insights.

In conclusion, despite their powerful capabilities in feature selection, machine learning techniques like XGBoost and SHAP inherently possess biases and limitations, especially in complex domains such as health risk assessment. These challenges underscore the necessity of a multi-faceted approach, integrating robust statistical methodologies and rigorous validation. By complementing ML with robust statistical validation, researchers can enhance interpretability and ensure more reliable outcomes. Future research should, therefore, focus on pioneering hybrid methodologies that effectively integrate the strengths of both machine learning and traditional statistical analysis. Ultimately, such an integrative approach is paramount for achieving accurate, reproducible, and clinically meaningful insights.

Funding sources

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

CRediT authorship contribution statement

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

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.

Acknowledgements

We extend our sincere gratitude to Takuma Yamazaki and Nobuko Inoue of Science Park, Inc. for their invaluable assistance with the extensive literature review.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2025.112308.

Data availability

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

References

[1] A. Schindele, A. Krebold, U. Heiß, K. Nimptsch, E. Pfaehler, C. Berr, R. A. Bundschuh, T. Wendler, O. Kertels, J. Tran-Gia, C.H. Pfob, C. Lapa, Interpretable machine learning for thyroid cancer recurrence prediction: leveraging XGBoost and SHAP analysis, Eur. J. Radiol. 186 (2025) 112049, https://doi.org/10.1016/j.ejrad.2025.112049.

[2] J. Ugirumurera, E.A. Bensen, J. Severino, J. Sanyal, Addressing bias in bagging and boosting regression models, Sci. Rep. 14 (2024) 18452, https://doi.org/10.1038/s41598-024-68907-5.

[3] P. Alaimo Di Loro, D. Scacciatelli, G. Tagliaferri, 2-step gradient boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy, Stat. Methods Appl. 32 (2023) 237–270, https://doi.org/10.1007/s10260-022-00643-4.

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

[5] M. Huti, T. Lee, E. Sawyer, A.P. King, An investigation into race bias in random forest models based on breast DCE-MRI derived radiomics features, in: S. Wesarg, E. Puyol Antón, J.S.H. Baxter (Eds.), Clin. Image-Based Proced., Fairness AI Med. Imaging: Ethical Philos. Issues Med, Imaging, Springer, Cham, 2023, pp. 225–234, https://doi.org/10.1007/978-3-031-45249-9_22.

[6] T. Salles, L. Rocha, M. Gonçalves, A bias-variance analysis of state-of-the-art random forest text classifiers, Advances in Data Analysis and Classification 15 (2021) 379–405, https://doi.org/10.1007/s11634-020-00409-4.

[7] W.G. Touw, et al., Data mining in the life sciences with random forest: a walk in the park or lost in the jungle, Brief. Bioinform. 14 (2013) 315–326, https://doi.org/10.1093/bib/bbs034.

[8] B. Bilodeau, N. Jaques, P.W. Koh, B. Kim, Impossibility theorems for feature attribution, Proc. Natl. Acad. Sci. u.s.a. 121 (2024) e2304406120, https://doi.org/10.1073/pnas.2304406120.

[9] 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.

[10] I. Kumar, C. Scheidegger, S. Venkatasubramanian, S. Friedler, Shapley residuals: Quantifying the limits of the Shapley value for explanations, in: M.A. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, J. Wortman Vaughan (Eds.), Adv. Neural Inf. Process. Syst. 34 (2021) 26598–26608. https://doi.org/10.48550/arXiv.2106.10860.

[11] M.A. Lones, Avoiding common machine learning pitfalls, Patterns 5 (2024) 101046, https://doi.org/10.1016/j.patter.2024.101046.

[12] A.M. Musolf, E.R. Holzinger, J.D. Malley, J.E. Bailey-Wilson, What makes a good prediction? feature importance and beginning to open the black box of machine learning in genetics, Hum. Genet. 141 (2022) 1515–1528, https://doi.org/10.1007/s00439-021-02402-z.

[13] J. Zhang, X. Wu, S.C.H. Hoi, J. Zhu, Feature agglomeration networks for single stage face detection, Neurocomputing 380 (2020) 180–189, https://doi.org/10.1016/j.neucom.2019.10.087.

[14] Y. Xie, Z. Jing, H. Pan, et al., Redefining the high variable genes by optimized LOESS regression with positive ratio, BMC Bioinf. 26 (2025) 104, https://doi.org/10.1186/s12859-025-06112-5.

[15] H. Yu, A.D. Hutson, A robust Spearman correlation coefficient permutation test, commun. stat. theory, Methods 53 (2024) 2141–2153, https://doi.org/10.1080/03610926.2022.2121144.

[16] K. Okoye S. Hosseini Correlation tests in R: Pearson cor, kendall's tau . Spearman's Rho, In: K. Okoye, S. Hosseini, R Programming: Statistical Data Analysis in Research 2024 Springer Nature 247 277 10.1007/978-981-97-3385-9_12.

Souichi Oka[a,*,1], Yoshiyasu Takefuji[b,2]

[a] *Science Park Corporation, 3-24-9 Iriya-Nishi Zama-shi, Kanagawa 252-0029, Japan*

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

* Corresponding author at: Souichi Oka, Science Park Corporation, 3-24-9 Iriya-Nishi Zama-shi, Kanagawa 252-0029, Japan.

E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka), takefuji@keio.jp (Y. Takefuji).

[1] ORCID: 0009-0000-4840-5232.

[2] ORCID: 0000-0002-1826-742X.