Letter to the Editor

Revisiting AI Model Interpretability in Lung Cancer Screening: Challenges in Balancing Predictive Performance and Reliability

Souichi Oka,[a] Yoshiyasu Takefuji[b]

Clinical Lung Cancer, Vol. 000, No.xxx, 1–2

To the Editor,

Henriksen et al., in "Maximizing lung cancer screening in high-risk population leveraging ML-developed risk prediction algorithms: Danish retrospective validation of LungFlag," present an impressive machine learning (ML)-based risk prediction algorithm, LungFlag, designed to maximize lung cancer screening in high-risk populations and support early detection.[1] While their approach demonstrates acceptable predictive performance, it raises important concerns regarding the reliability of model interpretability that warrant further discussion. A review of the literature reveals that the LungFlag model, developed by Medial EarlySign, is based on the XGBoost algorithm and utilizes ML techniques to predict lung cancer risk.[2] Their model has achieved performance metrics in both LC fast-track clinic patients (Population A) and outpatients with chronic obstructive pulmonary disease (COPD) (Population B), with LungFlag achieving an AUC of 0.63 in Population A. They identified smoking, age, and COPD as key predictors, and further highlighted that LungFlag identified high-risk individuals who were generally younger compared to those identified by PLCOm2012. While their study offers valuable insights and provides explainability through visualized clinical feature importance scores via SHapley Additive exPlanations (SHAP) values, it also presents critical methodological concerns that require further analysis.

It is important to note that, although a positive accuracy (AUC of 0.63) was reported, predictive performance and the reliability of feature importance are conceptually distinct. As supported by over 300 peer-reviewed articles, even high predictive metrics do not necessarily imply trustworthy or consistent feature importance rankings.[3-5] A more detailed discussion and supporting references are provided in the supplementary material. Interpreting feature importance in complex ML models such as XGBoost presents notable methodological challenges, particularly in clinical applications.[6-9] While these models offer considerable predictive capabilities, their internal mechanisms, such as tree-building logic and handling of feature interactions, can introduce structural biases that may distort the interpretation of feature relevance. Specifically, XGBoost tends to overemphasize features used in early splits of decision trees, potentially leading to skewed assessments of clinical factor importance.

Relatedly, the issue of multicollinearity among predictors undermines the reliability of importance estimates. When features are highly correlated, tree-based models may arbitrarily favor one over others, obscuring the collective contribution these features make to the outcome. Such models often overlook joint variance and mutual interactions, resulting in interpretations that can be misleading, especially in high-dimensional clinical datasets. Exacerbating these concerns, ML models optimized for predictive accuracy are prone to overfitting in noisy and heterogeneous clinical environments. Overfitting risks capturing spurious patterns rather than meaningful clinical signals, further distorting feature importance scores and compromising the model's interpretability and reliability in real-world applications.

Additionally, Henriksen et al. feed the output of the XGBoost model into SHAP for post hoc explanation. While this is a widely adopted practice, it raises significant interpretability concerns. SHAP explanations are fundamentally model-dependent, as they are derived directly from the model's predictions.[10,11] This dependency means that any biases embedded in the model—such as those arising from tree-splitting heuristics or overfitting—are inherited and potentially amplified by SHAP. In the presence of collinearity, SHAP tends to distribute contributions across correlated features, thereby diluting their apparent importance. Furthermore, SHAP may generate feature combinations that do not exist in empirical data distribution, leading to misleading or clinically implausible interpretations.

A major challenge in validating feature importance lies in the absence of ground truth. Different models employ distinct mechanisms for estimating importance, resulting in inconsistent and

[a]Science Park Corporation, Kanagawa 252-0029, Japan
[b]Faculty of Data Science, Musashino University, Tokyo 135-8181, Japan

Submitted: Aug 4, 2025; Accepted: Sep 15, 2025; Epub: xxx

Address for correspondence: Souichi Oka, PhD, SciencePark Corporation, 3-24-9 Iriya-Nishi Zama-shi, Kanagawa 252-0029, Japan.
E-mail contact: souichi.oka@sciencepark.co.jp

1525-7304/$ - see front matter
https://doi.org/10.1016/j.cllc.2025.09.005

Please cite this article as: Souichi Oka, Yoshiyasu Takefuji, Revisiting AI Model Interpretability in Lung Cancer Screening: Challenges in Balancing Predictive Performance and Reliability, Clinical Lung Cancer, https://doi.org/10.1016/j.cllc.2025.09.005

often contradictory rankings. This issue is exacerbated in high-dimensional settings, where complex interactions and collinearity obscure the true contribution of individual features. In clinical datasets, which are inherently noisy and heterogeneous, models may capture spurious correlations rather than meaningful signals, producing unreliable and unstable importance scores. Moreover, the sensitivity of feature importance to minor changes in data or model configuration undermines reproducibility and poses serious risks to clinical credibility and decision-making.

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 nonlinear relationships. Unsupervised techniques such as Feature Agglomeration (FA) and, where applicable, Highly Variable Gene Selection (HVGS)[12,13] offer valuable alternatives. In addition, nonparametric statistical methods like Spearman's rho and Kendall's tau[14,15] 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 ML models such as XGBoost, as implemented in LungFlag, demonstrate strong predictive capabilities for risk assessment, their inherent biases and the limitations of post hoc explanation methods such as SHAP raise serious concerns about interpretability. SHAP explanations are derived directly from the model's predictions, which means they reflect the internal logic of the model rather than objective feature relevance. This dependency can compromise the reliability of insights, especially in clinical contexts where transparency is essential. In oncology, where decisions must be both accurate and explainable, relying solely on predictive performance is insufficient. Interpretability must be robust, reproducible, and clinically grounded. To achieve this, a comprehensive approach is needed that combines ML with rigorous statistical validation. Such integration is essential to ensure that AI-driven insights are not only accurate but also clinically meaningful, trustworthy, and actionable in high-stakes medical decision-making.

Data Statement

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

Disclosure

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.

CRediT authorship contribution statement

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

Acknowledgments

We extend our sincere gratitude to Takuma Yamazaki and Nobuko Inoue of Science Park, Inc. for their invaluable assistance with the extensive literature review. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cllc.2025.09.005.

References

1. M.B. Henriksen, O. Hilberg, C. Juul, et al. Maximizing lung cancer screening in high-risk population leveraging ML-developed risk prediction algorithms: Danish retrospective validation of LungFlag, clin. Lung cancer. (Year not specified in original for this format, assuming current or near future based on DOI) https://doi.org/10.1016/j.cllc.2025.05.017. Accessed 29 July, 2025.
2. Gould MK, Huang B, Chang S, Kuan AT, Chen H. Machine learning for early lung cancer identification using routine clinical and laboratory data. *Am J Respir Crit Care Med*. 2021;204:438–446. doi:10.1164/rccm.202007-2791OC.
3. Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *ACM Queue*. 2018;16:31–57. doi:10.1145/3236386.3241340.
4. 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. doi:10.48550/arXiv.1801.01489.
5. Musolf AM, Holzinger ER, Malley JD, Bailey-Wilson JE. What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. *Hum Genet*. 2022;141:1515–1528. doi:10.1007/s00439-021-02402-z.
6. Ugirumurera J, Bensen EA, Severino J, Sanyal J. Addressing bias in bagging and boosting regression models. *Sci Rep*. 2024;14:18452. doi:10.1038/s41598-024-68907-5.
7. Alaimo Di Loro P, Scacciatelli D, Tagliaferri G. 2-step gradient boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy. *Stat Methods Appl*. 2023;32:237–270. doi:10.1007/s10260-022-00643-4.
8. Huti M, Lee T, Sawyer E, King AP, et al. An investigation into race bias in random forest models based on breast DCE-MRI derived radiomics features. In: Wesarg S, Puyol Antón E, Baxter JSH, et al., eds. *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging*. Cham: Springer; 2023:225–234. doi:10.1007/978-3-031-45249-9_22.
9. Adler AI, Painsky A. Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy*. 2022;24:687. doi:10.3390/e24050687.
10. Huang X, Marques-Silva J. On the failings of Shapley values for explainability. *Int J Approx Reason*. 2024;171:109112. doi:10.1016/j.ijar.2023.109112.
11. Kumar I, Scheidegger C, Venkatasubramanian S, Friedler S. Shapley residuals: quantifying the limits of the Shapley value for explanations. (Eds.). *Adv Neural Inf Process Syst*. 2021;34:26598–26608. doi:10.48550/arXiv.2106.10860.
12. Zhang J, Wu X, Hoi SHC, Zhu J. Feature agglomeration networks for single stage face detection. *Neurocomputing*. 2020;380:180–189. doi:10.1016/j.neucom.2019.10.087.
13. Xie Y, Jing Z, Pan H, et al. Redefining the high variable genes by optimized LOESS regression with positive ratio. *BMC Bioinform*. 2025;26:104. doi:10.1186/s12859-025-06112-5.
14. Yu H, Hutson AD. A robust Spearman correlation coefficient permutation test. *Commun Stat Theory Methods.*. 2024;53:2141–2153. doi:10.1080/03610926.2022.2121144.
15. Okoye K, Hosseini S. Correlation tests in R: Pearson Cor, Kendall's tau, and Spearman's rho. In: Okoye K, Hosseini S, eds. *R Programming: Statistical Data Analysis in Research*. NY: Springer Nature; 2024:247–277. doi:10.1007/978-981-97-3385-9_12.

Please cite this article as: Souichi Oka, Yoshiyasu Takefuji, Revisiting AI Model Interpretability in Lung Cancer Screening: Challenges in Balancing Predictive Performance and Reliability, Clinical Lung Cancer, https://doi.org/10.1016/j.cllc.2025.09.005