Ultrasound in Medicine & Biology 52 (2026) 252–253

Contents lists available at ScienceDirect

Ultrasound in Medicine & Biology

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

Letter to the Editor

Towards Reliable Feature Importance in Hashimoto's Thyroiditis Prediction: Reconstructing Machine Learning Frameworks

Chen et al. (2025) proposed a machine learning (ML) framework using XGBoost and SHapley Additive exPlanation (SHAP) for predicting Hashimoto's Thyroiditis (HT) stages [1]. While their study makes a valuable contribution, further constructive discussion on the interpretation of feature importance could help drive further advancement. For identifying stage-specific HT predictors, they analyzed 137 features (radiomic, clinical, and laboratory variables) derived from a patient cohort and their ultrasound images. XGBoost was selected over other ML algorithms like logistic regression, random forest, support vector machine, k-nearest neighbor, and artificial neural network due to its superior performance, achieving 95.8% accuracy and an AUROC of 0.947 on the test dataset. SHAP values were then used to evaluate feature importance, identifying key predictors such as first-order features from transverse ultrasound images, texture feature gray-level run length matrix from longitudinal views, and free thyroxine levels. Despite the widespread adoption of such frameworks, it is essential to recognize that high predictive performance does not guarantee reliable feature rankings. Inherent biases raise substantial concerns about the reliability of feature importance.

Numerous studies have highlighted the disconnect between predictive performance and meaningful attribution, reinforcing the need for rigorous, model-independent frameworks to support reproducible discovery and clinically relevant interpretation. Feature importance rankings often reflect artifacts of the prediction process rather than genuine causal relationships. Over-reliance on predictive accuracy to justify feature relevance is a well-documented issue, supported by over 300 peer-reviewed studies [2–5]. Details are discussed in the Supplementary Material.

XGBoost, like other tree-based ML models, has inherent biases in feature importance calculations, often overemphasizing features used in earlier splits [6–10]. These scores are also influenced by the model's splitting logic, feature interactions, and multicollinearity. SHAP values, a popular eXplainable AI (XAI) method, inherit and can worsen these biases because SHAP's explanations are directly dependent on the underlying model's output [11,12]. Therefore, relying on an XGBoost-SHAP pipeline combines two inherently biased methods, a common pitfall that can severely exacerbate interpretability issues and undermine the reliability of the analysis. The claim that this pipeline successfully identified predictive features, even with nonlinearity and interactions, warrants rigorous scrutiny given these compounded biases. Validating feature importance is inherently challenging due to the absence of ground truth, leading to model-specific biases and inconsistent rankings. In Chen et al.'s study, complex features, high dimensionality, and collinearity hinder ML interpretation for HT stage risk, increasing overfitting and reducing reliability due to sensitivity to small data or model changes.

To ensure accurate interpretations in health risk assessment, a robust analytical framework is essential. This approach should incorporate methodologies better suited for capturing complex associations within health data, such as unsupervised learning techniques including Feature Agglomeration (FA) and Highly Variable Gene Selection (HVGS) [13,14]. Additionally, non-parametric statistical methods like Spearman's rho or Kendall's tau would be highly beneficial [15,16]. These methods can detect various types of relationships, offering enhanced interpretability crucial for translating findings into actionable clinical insights. Ultimately, this multi-faceted approach is indispensable for generating accurate, reproducible, and clinically relevant insights that can truly advance health risk assessment.

CRediT authorship statement

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

Conflict of 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.

Acknowledgments

We extend our sincere gratitude to Nobuko Inoue and Takuma Yamazaki 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.

Data availability

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

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ultrasmedbio.2025.09.008.

DOI of original article: http://dx.doi.org/10.1016/j.ultrasmedbio.2025.09.007.

https://doi.org/10.1016/j.ultrasmedbio.2025.09.008
Received 20 July 2025; Accepted 9 September 2025

References

[1] Chen JH, Kang K, Wang XY, Chi JN, Gao XM, Li YX, et al. Development of radiomics-based risk prediction models for stages of Hashimoto's thyroiditis using ultrasound, clinical, and laboratory factors. Ultrasound Med Biol 2025. doi: 10.1016/j.ultrasmedbio.2025.05.025.

[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. doi: 10.48550/arXiv.1801.01489.

[3] 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. doi: 10.1145/3236386.3241340.

[4] Lones MA. Avoiding common machine learning pitfalls. Patterns 2024;5:101046. doi: 10.1016/j.patter.2024.101046.

[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;141:1515–28. doi: 10.1007/s00439-021-02402-z.

[6] 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;32:237–70. doi: 10.1007/s10260-022-00643-4.

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

[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 editors. In: Wesarg S, Puyol Antón E, Baxter JSH, editors. Clin Image-Based Proced, Fairness AI Med Imaging: Ethical Philos Issues Med Imaging. Cham: Springer; 2023. p. 225–34. doi: 10.1007/978-3-031-45249-9_22.

[9] Salles T, Rocha L, Gonçalves M. A bias-variance analysis of state-of-the-art random forest text classifiers. Adv Data Anal Classif 2021;15:379–405. doi: 10.1007/s11634-020-00409-4.

[10] Touw WG, et al. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle. Brief Bioinform 2013;14:315–26. doi: 10.1093/bib/bbs034.

[11] 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.

[12] Kumar I, Scheidegger C, Venkatasubramanian S, Friedler S. Shapley residuals: quantifying the limits of the Shapley value for explanations In: Ranzato MA, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors Adv Neural Inf Process Syst 2021;34:26598–608. doi: 10.48550/arXiv.2106.10860.

[13] Zhang J, Wu X, Hoi SCH, Zhu J. Feature agglomeration networks for single stage face detection. Neurocomputing 2020;380:180–9. doi: 10.1016/j.neucom.2019.10.087.

[14] Xie Y, Jing Z, Pan H, Xu X, Fang Q. Redefining the high variable genes by optimized LOESS regression with positive ratio. BMC Bioinformatics 2025;26:104. doi: 10.1186/s12859-025-06112-5.

[15] Yu H, Hutson AD. A robust Spearman correlation coefficient permutation test. Commun Stat Theory Methods 2024;53:2141–53. doi: 10.1080/03610926.2022.2121144.

[16] 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. Singapore: Springer Nature; 2024. p. 247–77. doi: 10.1007/978-981-97-3385-9_12.

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

[a] *Science Park Corporation, Kanagawa, Japan*

[b] *Faculty of Data Science, Musashino University, Tokyo, Japan*

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

E-mail address: souichi.oka@sciencepark.co.jp (S. Oka).