Computers in Biology and Medicine 196 (2025) 110710

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Letter to the Editor: Complementary statistical approaches for interpreting machine learning feature importance in osteoporosis risk

ARTICLE INFO

Keywords:
Osteoporosis risk
Machine learning
Predictive accuracy
Feature importance
Statistical validation
Ground truth

ABSTRACT

This paper comments on the valuable contribution by Carvalho and Gavaia regarding machine learning for osteoporosis risk prediction, particularly their use of a stacking ensemble model and feature importance analysis. While acknowledging the model's high predictive accuracy, we raise a crucial concern: high accuracy does not inherently validate the reliability of feature importance interpretation. We discuss how the interpretation of feature importance from complex, model-dependent methods like those used can be influenced by model structure and data characteristics, potentially overemphasizing certain variables or reflecting model-specific relevance rather than true underlying causal drivers of osteoporosis risk. Validating feature importance is inherently difficult due to the absence of ground truth for causal relationships. To address these limitations and move beyond purely model-dependent predictive importance, we propose integrating complementary statistical methodologies, such as Spearman's rho, Kendall's tau, Mutual Information, and Total Correlation. These impartial and resilient methods can offer more robust insights into variable relationships. By combining predictive ML modeling with these statistical approaches, we aim to advance the understanding of complex health outcomes like osteoporosis in biomedical and healthcare applications, providing a more dependable assessment of feature importance and model behavior.

1. Introduction

The recent paper by Carvalho and Gavaia in Computers in Biology and Medicine, "Enhancing osteoporosis risk prediction using machine learning: A holistic approach integrating biomarkers and clinical data," makes a valuable contribution to the prediction of osteoporosis risk [1]. Their study evaluates the potential of a stacking ensemble machine learning model integrating biomarkers and clinical data for predicting osteoporosis risk using data from NHANES cycles 2007–2014. However, the reliance on complex machine learning (ML) models and the interpretation of feature importance warrant further discussion.

Carvalho and Gavaia employed a stacking ensemble model combining four specialized ML models: Gradient Boosting, Random Forest, XGBoost, and LightGBM, with a logistic regression meta-classifier. They reported robust performance, achieving 93 % accuracy and an AUC of 0.94 through cross-validation. Beyond evaluating model performance, a key aspect of their work involved feature importance analysis based on the model's output, revealing influential predictors such as age, arm muscle circumference, and body weight. This raises critical concerns about potential bias in the ranked features.

Although Carvalho and Gavaia have offered a novel method for predicting osteoporosis risk, this paper highlights a concern regarding how feature importance derived from ML models is interpreted. Their study prominently features the high predictive accuracy of their stacking ensemble model and then proceeds to analyze its feature importance. However, it's vital to understand that achieving high predictive accuracy does not automatically confirm the reliability of the feature importance scores. While Carvalho and Gavaia's goal is to pinpoint key osteoporosis risk factors through feature importance, the initial impressive prediction accuracy of their model could inadvertently imply that the subsequent feature importance interpretation is reliable. As demonstrated by over 300 previous studies, strong predictive performance does not guarantee dependable feature importance interpretation [2–7]. More details and supporting literature can be found in the supplementary material. Drawing from Carvalho and Gavaia's research, we further examine this issue. We suggest incorporating complementary statistical methods to facilitate a more dependable understanding of feature importance, with the aim of improving methodologies in biomedical and healthcare fields.

2. Methodological limitations of ML

Interpreting complex machine learning models like Random Forest, Gradient Boosting, XGBoost, and LightGBM, particularly for understanding feature importance, presents methodological complexities. These models, while offering strong predictive power, can generate feature importance scores that are influenced by factors inherent in their structure and operation, such as splitting logic in tree-based models or the handling of feature interactions and multicollinearity [8–10]. This can lead to skewed assessments, potentially favoring certain variables or features with particular data structures, which can overemphasize the importance of features used in earlier splits [11–13].

The feature importance analysis employed in this study by Carvalho and Gavaia relies on the output of their stacking ensemble model. Consequently, the way the ensemble model integrates the predictions and feature handling characteristics of its base learners (Gradient Boosting, Random Forest, XGBoost, and LightGBM) can influence the perceived importance of features. While the reported feature importance

https://doi.org/10.1016/j.compbiomed.2025.110710
Received 9 May 2025; Accepted 1 July 2025
0010-4825/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

provides valuable insights into the model's decision-making process, it is inherently tied to the specific model architecture and the way these algorithms interpret relationships within the data. This model-dependent nature means the ranked features reflect what is important for that specific model's prediction, rather than definitively representing the true underlying causal drivers of osteoporosis risk.

Fundamentally, validating feature importance is challenging because the true causal relationships lack ground truth values. This highlights the difficulty of identifying the actual underlying causal factors or drivers using only feature analysis processes that are dependent on a specific model [14–18]. Acknowledging the limitations of relying solely on complex ML models for robust feature importance interpretation, this study suggests employing complementary statistical methodologies. These are intended to provide more objective insights into the connections between clinical variables and osteoporosis risk, thereby shifting the focus from model-specific predictive importance to understanding potential underlying mechanisms.

3. Proposed solutions

To address these limitations effectively, it is crucial to establish a comprehensive analytical framework that incorporates data characteristics, the statistical relationships between variables, and rigorous validation. Successful modeling and interpretation are contingent upon a deep understanding of the underlying biological and clinical processes involved in osteoporosis. Exploring complex associations between variables, particularly through non-parametric methods, is of paramount importance. Furthermore, verifying the statistical significance of findings via hypothesis testing and p-value analysis is essential to prevent drawing inaccurate conclusions.

Instead of relying exclusively on complex machine learning models and their built-in interpretability techniques for identifying key features and understanding model behavior, we propose a synergistic approach that incorporates impartial, resilient statistical methods, such as Spearman's rho and Kendall's tau, particularly adept at characterizing monotonic relationships [19,20]. For more complex dependencies, including non-monotonic interactions among variables, alternative non-parametric methods like Mutual Information and Total Correlation offer valuable insights [21–24]. Prioritizing these statistical principles, combined with ML and domain expertise, will substantially bolster the credibility and dependability of feature importance and model behavior assessments in domains like biomedical and healthcare engineering.

4. Conclusion

In conclusion, the study by Carvalho and Gavaia provides a valuable model and identifies features relevant for osteoporosis risk prediction using their chosen methodology. However, as with many machine learning applications in complex biological systems, interpreting the identified features as definitive drivers of osteoporosis risk variability warrants careful consideration of methodological limitations and the inherent challenges of validation. Addressing the limitations of interpreting features solely through model-dependent approaches requires a broader strategy. To deepen understanding of complex health outcomes like osteoporosis, we should integrate predictive ML modeling with complementary statistical methodologies.

CRediT authorship contribution statement

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

Ethics statement

An Ethics Statement is not applicable to this correspondence as it is a commentary and methodological discussion based on the analysis of a previously published study which utilized publicly available, de-identified data. This work did not involve any new collection of human or animal data, or interventions requiring ethical approval or informed consent.

Data availability

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

Funding sources

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. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.110710.

References

[1] F.R. Carvalho, P.J. Gavaia, Enhancing osteoporosis risk prediction using machine learning: a holistic approach integrating biomarkers and clinical data, Comput. Biol. Med. 192 (Part B) (2025) 110289, https://doi.org/10.1016/j.compbiomed.2025.110289.

[2] Z.C. Lipton, The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery, ACM Queue 16 (3) (2018) 31–57, https://doi.org/10.1145/3236386.3241340.

[3] A. Fisher, C. Rudin, F. Dominici, 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. 20 (2019) 177, https://doi.org/10.48550/arXiv.1801.01489.

[4] K. Lenhof, L. Eckhart, L.M. Rolli, H.P. Lenhof, Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer, Briefings Bioinf. 25 (5) (2024) bbae379, https://doi.org/10.1093/bib/bbae379.

[5] H. Mandler, B. Weigand, A review and benchmark of feature importance methods for neural networks, ACM Comput. Surv. 56 (2024), https://doi.org/10.1145/3679012. Article 318.

[6] J.L. Potharlanka, M.N. Bhat, Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms, Sci. Rep. 14 (1) (2024) 2923, https://doi.org/10.1038/s41598-024-53141-w.

[7] D. Wood, T. Papamarkou, M. Benatan, R. Allmendinger, Model-agnostic variable importance for predictive uncertainty: an entropy-based approach, Data Min. Knowl. Discov. 38 (2024) 4184–4216, https://doi.org/10.1007/s10618-024-01070-7.

[8] 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: Clinical Image Based Procedure Fairness AI Med Imaging Ethical Philos Issues Med Imaging, 2023, https://doi.org/10.1007/978-3-031-45249-9_22. Vol. 14242, pp. 225–234.

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

[10] W.G. Touw, J.R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, S.A.F. T. van Hijum, Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Briefings Bioinf. 14 (3) (2013) 315–326, https://doi.org/10.1093/bib/bbs034.

[11] 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.

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

[13] 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.

[14] B. Bilodeau, N. Jaques, P.W. Koh, B. Kim, Impossibility theorems for feature attribution, Proc. Natl. Acad. Sci. 121 (2024) e2304406120, https://doi.org/10.1073/pnas.2304406120.
[15] M.A. Lones, Avoiding common machine learning pitfalls, Patterns 5 (10) (2024) 101046, https://doi.org/10.1016/j.patter.2024.101046.
[16] A. Faragalli, L. Ferrante, N. Angelakopoulos, R. Cameriere, E. Skrami, Do machine learning methods solve the main pitfall of linear regression in dental age estimation? Forensic Sci. Int. 367 (2025) 112353 https://doi.org/10.1016/j.forsciint.2024.112353.
[17] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C.A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, B. Bischl, General pitfalls of model-agnostic interpretation methods for machine learning models, in: A. Holzinger, R. Goebel, R. Fong, T. Moon, K.R. Müller, W. Samek (Eds.), xxAI - Beyond Explainable AI, 13200, Springer, 2022, p. 4, https://doi.org/10.1007/978-3-031-04083-2_4.
[18] P.M. Steiner, Y. Kim, The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases, J. Causal Inference 4 (2016), https://doi.org/10.1515/jci-2016-0009. Article 20160009.
[19] K. Okoye, S. Hosseini, Correlation tests in R: pearson cor, Kendall's tau, and Spearman's rho, in: K. Okoye, S. Hosseini (Eds.), R Programming: Statistical Data Analysis in Research, Springer Nature, 2024, pp. 247–277, https://doi.org/10.1007/978-981-97-3385-9_12.
[20] H. Yu, A.D. Hutson, A robust Spearman correlation coefficient permutation test, Commun. Stat. Theor. Methods 53 (6) (2024) 2141–2153, https://doi.org/10.1080/03610926.2022.2121144.
[21] J.D. Gibson, Entropy and mutual information, in: Information Theoretic Principles for Agent Learning. Synthesis Lectures on Engineering, Science, and Technology, Springer, Cham, 2025, https://doi.org/10.1007/978-3-031-65388-9_2.
[22] T. Kerby, T. White, K.R. Moon, Learning local higher-order interactions with total correlation, in: Proc. 2024 IEEE 34th Int. Workshop Mach. Learn. Signal Process. (MLSP), 2024, pp. 1–6, https://doi.org/10.1109/MLSP58920.2024.10734758.
[23] Y. Shi, R. Golestanian, A. Vilfan, Mutual information as a measure of mixing efficiency in viscous fluids, Phys. Rev. Res. 6 (2024) L022050, https://doi.org/10.1103/PhysRevResearch.6.L022050.
[24] S. Tserkis, S.M. Assad, P.K. Lam, P. Narang, Quantifying total correlations in quantum systems through the pearson correlation coefficient, Phys. Lett. A 543 (2025) 130432, https://doi.org/10.1016/j.physleta.2025.130432.

Souichi Oka[a,*], Takuma Yamazaki[a], Yoshiyasu Takefuji[b]
[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.
E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka), tyamazaki@sciencepark.co.jp (T. Yamazaki), takefuji@keio.jp (Y. Takefuji).