Geriatric Nursing 73 (2026) 104318

Contents lists available at ScienceDirect

Geriatric Nursing

journal homepage: https://www.sciencedirect.com/journal/geriatric-nursing

Beyond predictive performance: Interpretability challenges and feature importance bias in XGBoost-based readmission models

ARTICLE INFO

Article history:
Received 10 February 2026
Received in revised form 23 May 2026
Accepted 12 August 2026
Available online xxx

Keywords:
Coronary heart disease
Readmission prediction
Machine learning
Feature importance bias
Model-agnostic validation

ABSTRACT

Song et al. report a machine-learning framework based on the eXtreme Gradient Boosting (XGBoost) algorithm for predicting 1-year unplanned readmissions among elderly patients with coronary heart disease (CHD). This commentary examines critical limitations in the interpretability and methodological robustness of such models. Extensive prior work has demonstrated that tree-based algorithms can exhibit structural biases in feature-importance estimates, particularly when predictors display substantial collinearity or heterogeneous measurement scales. SHAP explanations, by inheriting these model-embedded biases, may overstate the relevance of variables whose prominence arises from algorithmic artifacts rather than clinically coherent patterns. To enhance reliability, model-agnostic validation, non-parametric association analyses, and unsupervised strategies that mitigate multicollinearity should complement predictive modeling efforts. Strengthening interpretability frameworks is essential to ensure that machine-learning–derived insights meaningfully inform cardiovascular care and support reproducible clinical translation.

© 2026 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Letter to the Editor

Song et al. report a machine-learning framework based on the eXtreme Gradient Boosting (XGBoost) algorithm for predicting 1-year unplanned readmissions among elderly patients with coronary heart disease (CHD) .[1] Based on a retrospective single-center cohort of 2137 patients aged 60–99 years—including 482 readmitted within one year—they developed and validated a predictive framework integrating routinely available demographic, clinical, laboratory, and comorbidity-related variables. After excluding features with >80% missing data and applying ridge regression for feature selection, a final set of 38 predictors was retained to balance performance and clinical feasibility. In the independent test cohort, predictive performance reached an AUROC of 0.704, accompanied by high specificity (0.936). SHapley Additive exPlanations (SHAP) analysis identified length of stay, the age-adjusted Charlson comorbidity index, monocyte count, blood glucose, and red blood cell count as the most influential contributors to the risk of readmission. However, the utility and reliability of the derived feature-importance rankings warrant further discussion before any translation into clinical practice.

A growing body of literature has examined biases in feature importance across different modeling frameworks. Representative prior studies have highlighted multiple sources of such bias, including structural characteristics of tree-based models, [2–8] limitations of model-dependent explanation methods,[9–11] and broader challenges in reconciling predictive performance with reliable attribution.[12–15] These findings are consistent with recent trends in medical and healthcare-related machine learning, where growing attention has been paid to potential biases in feature importance and model interpretability.[16–23] In particular, feature importance measures may be influenced by model-specific characteristics, potentially limiting their ability to fully capture true feature-outcome relationships. Further details and additional references are provided in the Supplementary Material. This perspective reinforces the need to approach such findings cautiously when considering their application to real-world clinical settings and to recognize the inherent limitations highlighted by prior methodological evaluations. In the absence of public access to the dataset, we instead cite prior empirical studies showing that feature importance measures may not reliably capture true underlying relationships, even in well-performing models.[24] This does not critique the original study but highlights a general limitation of model-dependent attribution methods in complex clinical data.

Tree-based models, including XGBoost and random forests, can exhibit systematic biases in feature importance, particularly favoring variables that enable early splits within the model structure. This tendency becomes especially pronounced in high-dimensional clinical datasets characterized by complex correlations and multicollinearity, which can amplify the risk of overfitting and distort the apparent relevance of individual predictors. In Song et al.'s framework, the analytical approach incorporates a broad set of routinely collected demographic, clinical, laboratory, and comorbidity-related variables, many of which are inherently correlated, such as indices of disease burden, inflammatory markers, and metabolic measures. Under these conditions, top-ranked features may emerge not because of a stable or causal association with readmission risk, but because of their statistical position within the feature space. In the absence of an independent reference or external validation for feature stability, such rankings should be interpreted cautiously, as they may not reliably reflect robust and reproducible relationships.

https://doi.org/10.1016/j.gerinurse.2026.104318
0197-4572/$ – see front matter © 2026 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

When discussing SHAP in practice, it is useful to distinguish how its implementations operate. In particular, variants such as TreeSHAP and KernelSHAP are based on distinct approximation strategies and assumptions, which can lead to different behaviors depending on the model and data context. Recent studies have also explored methods to improve the robustness and reliability of model explanations, although this remains an active area of research. Furthermore, because SHAP is inherently contingent upon the model under interpretation, it inherits the biases embedded within that model. As a post-hoc explanation tool, it reflects the model's internal logic, which may not align with the underlying structure of the data. In the context of an XGBoost-based framework applied to clinical and laboratory variables for predicting 1-year unplanned readmission, SHAP values primarily describe how the trained model organizes and weights its inputs rather than how those inputs relate to established clinical knowledge or mechanisms. Moreover, SHAP explanations can be sensitive to differences in model architecture and hyperparameter choices, potentially producing inconsistent interpretations across otherwise similar models. Consequently, such outputs warrant rigorous scrutiny and complementary, model-agnostic validation to ensure that the reported attributions are stable and not artifacts of a particular modeling configuration or analytical pipeline. Alternative explainability approaches, including permutation feature importance, LIME, counterfactual explanations, and causal inference-based methods, may offer complementary perspectives on feature relevance, each with distinct assumptions and methodological trade-offs.

To address these limitations, a robust analytical framework is required that extends beyond model-dependent attribution and explicitly accounts for the structural complexity of clinical data. Techniques such as Highly Variable Gene Selection and Feature Agglomeration provide unsupervised approaches to dimensionality reduction, helping to mitigate multicollinearity by grouping or prioritizing correlated variables before downstream modeling.[25,26] In their study cohort analyzed, the numerous clinical, laboratory, and comorbidity-related variables exhibited substantial interdependence, and these preprocessing steps reduced redundancy while preserving relevant variation. Complementing these strategies with non-parametric association methods, including Spearman's rho or Kendall's tau, can facilitate the exploration of monotonic and non-linear relationships that may not be captured by model-specific assumptions. Together, these approaches offer a more stable foundation for identifying robust predictors and improving the interpretability and translatability of findings in cardiovascular research.

In summary, Song et al. advance data-driven risk stratification, yet intrinsic model biases and their propagation through SHAP analysis constrain the interpretability of downstream clinical inferences. In this context, biased feature importance may have direct clinical implications. Misinterpretation of feature rankings could lead to inappropriate weighting of clinical variables, potentially affecting risk stratification and treatment prioritization. For example, variables such as length of stay or laboratory markers identified as highly influential in Song et al.'s analysis may be overemphasized due to model-specific characteristics rather than stable clinical relevance, potentially distorting clinical interpretation. Furthermore, reliance on model-dependent attributions in real-world settings may introduce risks if such patterns do not reflect robust or clinically meaningful associations. To strengthen methodological coherence, analytical frameworks must prioritize dimensionality reduction and systematic stability assessments. Through these strategies, investigators can effectively decouple clinical signals from model artifacts and ensure the reliable translation of machine-learning insights into cardiovascular practice.

Funding source

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

Data availability

Data availability: No new data were created or analyzed in this study.

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.

CRediT authorship contribution statement

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

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.gerinurse.2026.104318.

References

1. Song X, Shi J, Zhu C, Xian F, Dong Z, Li J. XGBoost machine learning algorithm for predicting unplanned readmission in elderly patients with coronary heart disease. *Geriatr Nurs*. 2025;66:103609. https://doi.org/10.1016/j.gerinurse.2025.103609.
2. Ugirumurera J, Bensen EA, Severino J, Sanyal J. Addressing bias in bagging and boosting regression models. *Sci Rep*. 2024;14:18452. https://doi.org/10.1038/s41598-024-68907-5.
3. 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. https://doi.org/10.1007/s11634-020-00409-4.
4. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SAFT. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle. *Brief Bioinform*. 2013;14:315–326. https://doi.org/10.1093/bib/bbs034.
5. van Giffen B, Herhausen D, Fahse T. Overcoming the pitfalls and perils of algorithms: a classification of machine learning biases and mitigation methods. *J Bus Res*. 2022;144:93–106. https://doi.org/10.1016/j.jbusres.2022.01.076.
6. Nalenz M, Rodemann J, Augustin T. Learning de-biased regression trees and forests from complex samples. *Mach Learn*. 2024;113:3379–3398. https://doi.org/10.1007/s10994-023-06439-1.
7. Adler AI, Painsky A. Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy*. 2022;24:687. https://doi.org/10.3390/e24050687.
8. 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. https://doi.org/10.1007/s10260-022-00643-4.
9. Huang X, Marques-Silva J. On the failings of Shapley values for explainability. *Int J Approx Reason*. 2024;171:109112. https://doi.org/10.1016/j.ijar.2023.109112.
10. Bilodeau B, Jaques N, Koh PW, Kim B. Impossibility theorems for feature attribution. *Proc Natl Acad Sci U S A*. 2024;121:e2304406120. https://doi.org/10.1073/pnas.2304406120.
11. Hooshyar D, Yang Y. Problems with SHAP and LIME in interpretable AI for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access*. 2024;12:137472–137490. https://doi.org/10.1109/ACCESS.2024.3463948.
12. 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. https://doi.org/10.48550/arXiv.1801.01489.
13. 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. https://doi.org/10.1145/3236386.3241340.
14. Lenhof K, Eckhart L, Rolli LM, Lenhof HP. Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. *Brief Bioinform*. 2024;25:bbae379. https://doi.org/10.1093/bib/bbae379.
15. 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. https://doi.org/10.1007/s00439-021-02402-z.

16. Ansari ZA, Kumar KK, Fatima S, Siddiqui S, Mohsin SW. Explainable artificial intelligence for cross domain evaluation of predictive models in multi-disease diagnosis. *Discover Comput*. 2026;29:163. https://doi.org/10.1007/s10791-026-10061-9.
17. Ansari ZA, Ansari A, Nayak SC, Dehuri S, Cho SB, Favorskaya M. Empowering breast cancer diagnostics: sHAP-enhanced explainable AI. In: Nayak, Dehuri, Cho, Favorskaya, eds. *Green Artificial Intelligence and Industrial Applications (G-AIIA). Artificial Intelligence-Enhanced Software and Systems Engineering. Green Artificial Intelligence and Industrial Applications (G-AIIA). Artificial Intelligence-Enhanced Software and Systems Engineering*. 8, Cham: Springer; 2026. https://doi.org/10.1007/978-3-031-99882-9_22.
18. Ansari ZA, Ansari MSH, Khan A, Pant H, Fahad S, Prasad PVH. Explainable breast cancer diagnosis: integrating genetic algorithms with LIME-based machine learning. *Evol Intell*. 2026;19. https://doi.org/10.1007/s12065-025-01131-w.
19. Ansari ZA, Arif M, Rajaboina NB, Shaikh AA, Singh Y. Integrated ensemble strategy for breast cancer detection using dimensionality reduction technique. *Adv Distrib Comput Artif Intell J*. 2025;14:e31899. https://doi.org/10.14201/adcaij.31899.
20. Ansari ZA, Soni S, Fatima S, Siddiqui S, Prasad PVH. A multi-model deep learning framework for SEM-based defect detection in FAPbI3 perovskite thin films. *Sci Rep*. 2025;15:41909. https://doi.org/10.1038/s41598-025-25848-x.
21. Ansari ZA, Ansari MSH, Khan A, Jhanjhi NZ, Rathnamma G. Optimized and interpretable machine learning framework for early breast cancer detection. *Health Technol*. 2025;15:1135–1147. https://doi.org/10.1007/s12553-025-01011-4.
22. Harshith M, Ansari ZA, Fatima S, Siddiqui S, Swarna S, Reddy DRN, Mohsin SW. Federated microservices architecture with blockchain for privacy-preserving and scalable healthcare analytics. *Sci Rep*. 2026;16:9023. https://doi.org/10.1038/s41598-026-39837-1.
23. Ahmed R, Tripathi MMT, Ali ZA. A comprehensive review on the role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models. *Discover Artif Intell*. 2025;5:75. https://doi.org/10.1007/s44163-025-00307-8.
24. Oka S, Takefuji Y. Revisiting AI interpretability in precision oncology: why predictive accuracy does not ensure stable feature importance. *Cancers (Basel)*. 2026;18:593. https://doi.org/10.3390/cancers18040593.
25. Xie Y, Jing Z, Pan H, Xu X, Fang Q. Redefining the high variable genes by optimized LOESS regression with positive ratio. *BMC Bioinform*. 2025;26:104. https://doi.org/10.1186/s12859-025-06112-5.
26. Zhang J, Wu X, Hoi SCH, Zhu J. Feature agglomeration networks for single stage face detection. *Neurocomputing*. 2020;380:180–189. https://doi.org/10.1016/j.neucom.2019.10.087.

Souichi Oka, PhD*
Maito Suzuki, MSc
AI Research Laboratory, Science Park Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa, 252-0029, Japan
E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka), msuzuki@sciencepark.co.jp (M. Suzuki).

Yoshiyasu Takefuji, PhD
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan
E-mail address: takefuji@keio.jp (Y. Takefuji).

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