Computer Methods and Programs in Biomedicine 272 (2025) 109085

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Beyond predictive accuracy: Statistical validation of feature importance in biomedical machine learning

ARTICLE INFO

Keywords:
Respiratory exacerbations
Machine learning
Feature importance
Model interpretability
Statistical validation

ABSTRACT

In medical machine learning (ML), a fundamental methodological distinction exists between optimizing model performance for predictive tasks and pursuing causal inference for mechanistic interpretation. Achieving high predictive accuracy does not necessarily imply that a model can uncover the true physiological mechanisms underlying the data. This letter addresses a critical interpretational challenge in medical machine learning, building upon Yuyang Yan et al.'s valuable work on exacerbation classification in asthma and COPD. While their multi-feature fusion model, particularly comprising models such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), and Bidirectional Long Short-Term Memory (BiLSTM) demonstrates high predictive accuracy for respiratory exacerbations, we highlight that such performance alone does not guarantee reliable insights into feature importance. Complex tree-based models like RF, when interpreted via methods like SHapley Additive exPlanations (SHAP), can exhibit inherent biases, overemphasizing features used in early splits and reflecting what is important for their specific prediction rather than the true underlying physiological drivers. Validating feature importance remains challenging without ground truth, as different models often yield varying rankings. We argue that solely relying on model-dependent interpretations risks misrepresenting the actual mechanisms of complex medical phenomena. Therefore, we advocate for a robust analytical strategy that transcends mere predictive metrics. This involves a synergistic approach combining the predictive power of ML with impartial, complementary statistical methodologies—such as non-parametric correlation and mutual information—to ensure genuinely trustworthy scientific insights into the true drivers of respiratory exacerbations.

1. Letter to the editor

The application of machine learning (ML) in medicine typically serves two distinct methodological objectives. The first is performance optimization, which aims to maximize predictive metrics—such as accuracy or AUC—to develop effective diagnostic or prognostic models. The second is causal inference for mechanistic interpretation, which seeks to identify the true underlying drivers of a medical condition by analyzing the model's internal logic, such as feature importance or learned representations. A common methodological pitfall is to conflate these objectives, mistakenly assuming that a model with high predictive performance is also valid for mechanistic interpretation. This assumption can lead to misleading conclusions about disease etiology and hinder the development of truly explanatory models.

The recent paper by Yuyang Yan et al. (2025) makes a valuable contribution to the classification of exacerbations in chronic respiratory diseases [1]. Their study develops a multi-feature fusion model that integrates various acoustic, text, and spectral features. Employing a suite of models including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), and Bidirectional Long Short-Term Memory (BiLSTM), they achieved robust classification performance, with performance reaching up to an accuracy of 89.18 % and an F1 score of 82.50 %. A key aspect of their work was the subsequent feature importance analysis; to their credit, they applied SHapley Additive exPlanations (SHAP) to all four of their ML models, revealing influential predictors. While these models' predictive performance is noteworthy, it is crucial to recognize that high accuracy does not automatically validate its SHAP-based interpretations. This is a well-documented limitation in the ML field [2–7], as feature importance rankings can reflect model-specific biases rather than the true underlying drivers of a disease without robust statistical validation. Further details are available in the supplementary material.

Unpacking the feature importance derived from sophisticated machine learning architectures, particularly tree-based ensembles like RF, presents substantial methodological hurdles. Despite their formidable predictive capabilities, these algorithms yield feature importance scores that are inherently shaped by their internal mechanics, including their splitting heuristics and interaction management. A critical tendency of tree-based models is to disproportionately emphasize features that contribute only marginally to predictive accuracy. This bias is often exacerbated by early splits in the tree-building process, which can cause certain features to dominate the structure due to initial partitioning decisions rather than true relevance [8–13]. This has been empirically demonstrated using real-world datasets to result in misleading feature importance rankings, frequently underrepresenting variables with clear domain relevance [14].

The insights into feature importance presented by Yuyang Yan et al. are predicated on the outputs of their chosen best-performing ML model coupled with SHAP. Consequently, the intrinsic biases and operational characteristics of this particular ML model, alongside those of the

https://doi.org/10.1016/j.cmpb.2025.109085
Received 29 May 2025; Received in revised form 16 September 2025; Accepted 23 September 2025
Available online 24 September 2025

Explainable Artificial Intelligence (XAI) methodology itself, inevitably color the perceived significance of various features. Given that SHAP is intrinsically linked to the models it interprets, it can either mirror or amplify these inherent model biases. This inherent reliance on the model means that the hierarchy of features revealed primarily reflects what is critical for that specific model's predictions and explanations, rather than unequivocally representing the true underlying drivers of exacerbation [15–19].

Fundamentally, substantiating feature importance is fraught with challenges given the absence of ground truth. Distinct models, owing to their differing methodologies and intrinsic biases, frequently produce divergent rankings. Indeed, the very act of comparing feature importances from models known to yield inconsistent or unreliable rankings across different algorithms, even on the same dataset, further complicates the interpretation. While Yuyang Yan et al. primarily centered their discussion on their best-performing model's performance and SHAP-based interpretation, the broader issue of varying feature importances across different models persists. In the realm of medical investigations, intricate feature sets potentially exhibiting collinearity can further complicate interpretation and exacerbate these issues, making it arduous to confidently pinpoint the actual determinants solely through model-dependent analysis. These factors underscore that superior predictive accuracy does not, by itself, assure a trustworthy interpretation of feature importance from biased models and XAI techniques.

To illustrate the potential clinical risks of misinterpreting feature importance, we present the following illustrative scenario, consistent with the findings of the target study. Suppose SHAP analysis ranks Loudness above Mel-Frequency Cepstral Coefficients (MFCC) 6 in predicting COPD exacerbations from patient speech recordings. Loudness is intuitive and clinically interpretable, often linked to breathlessness, while MFCC6—reflecting subtle vocal tract changes—is abstract. Given this ranking, clinicians may reasonably trust SHAP and prioritize Loudness in their decision-making. However, Loudness typically declines only after significant respiratory distress, making it a late-stage indicator. In contrast, MFCC6 may capture early physiological changes, potentially offering a more proactive signal for intervention. This scenario illustrates how interpretability and SHAP-based rankings, while helpful, can lead to overlooking less intuitive features that may be more clinically valuable.

To navigate the identified limitations in discerning feature importance from complex ML and XAI approaches, a robust analytical framework is imperative. Such a framework must commence with a profound comprehension of the data's inherent characteristics and the fundamental physiological processes governing exacerbations. Beyond mere model-derived metrics, it is vital to rigorously investigate the statistical relationships between variables, encompassing complex, non-monotonic associations, by deploying appropriate non-parametric methods. Furthermore, rigorous statistical validation, incorporating techniques like hypothesis testing, is indispensable to confirm that findings are not mere artifacts of the modeling procedure but genuinely reflect underlying relationships.

Rather than exclusively leaning on ML models and their embedded interpretability tools for feature selection and behavioral understanding, we advocate for a synergistic paradigm. This entails seamlessly integrating the predictive power of ML with impartial and rigorous statistical methodologies, such as Spearman's rho and Kendall's tau, which are exceptionally adept at characterizing monotonic relationships [20–21]. Crucially, because these methods operate on data ranks rather than raw numerical values, they are inherently robust to outliers and non-normally distributed data—conditions frequently encountered in clinical datasets. This makes them well-suited to provide a stable, model-independent baseline for assessing feature associations. For delving into more intricate dependencies, including non-monotonic interactions among variables, alternative non-parametric avenues like Mutual Information and Total Correlation offer invaluable perspectives [22–25]. This is particularly vital in medicine, where biological relationships are often non-linear—such as U-shaped dose-response curves—and may be overlooked by simpler correlation metrics. Mutual Information can reveal these complex associations without making prior assumptions about their functional form. By prioritizing foundational statistical principles and employing methods capable of uncovering a diverse spectrum of variable relationships, researchers can significantly bolster the credibility and reliability of feature importance assessments and model interpretations within clinical modeling domains.

In practice, statistical validation should be the first step when interpreting the relationship between features and outcomes. Before turning to ML models or tools like SHAP, it is essential to assess each feature's association with the outcome using standard statistical methods—such as Spearman's rho and corresponding p-values. These provide an objective, model-independent basis for evaluating feature relevance. If ML models or SHAP are used for feature interpretation, their outputs must be critically examined. Specifically, the feature importance rankings they produce should be compared against the statistically derived rankings. Features that are emphasized by the model but lack statistical significance should be treated with caution, as they may reflect model-specific biases rather than genuine associations. In this sense, statistical validation is a necessary safeguard against overinterpreting model-driven explanations.

In conclusion, while Yuyang Yan et al. offer a highly accurate multi-feature fusion model for exacerbation classification, our letter underscores a crucial challenge in medical modeling. The inherent constraints of ML and XAI tools, notably the best-performing model and SHAP employed for interpretation, mean their derived features cannot be definitively assumed as the sole drivers of exacerbation changes [14]. High predictive performance in classification tasks does not inherently ensure reliable feature importance. Relying solely on such interpretations risks distorting the underlying mechanisms of complex medical conditions. Therefore, achieving truly credible scientific insights necessitates a robust strategy that thoughtfully combines ML's predictive strength with the objective rigor of complementary statistical methods.

Funding sources

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.

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.

CRediT authorship contribution statement

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

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.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2025.109085.

References

[1] Y. Yan, L. van Bemmel, F.M.E. Franssen, S.O. Simons, V. Urovi, Developing a multi-feature fusion model for exacerbation classification in asthma and COPD, Comput. Methods Programs Biomed. 268 (2025) 108796, https://doi.org/10.1016/j.cmpb.2025.108796.

[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, Brief. Bioinform. 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) 318, https://doi.org/10.1145/3679012. Article.

[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: Clin Image Based Proced Fairness, AI Med Imaging Ethical Philos Issues Med Imaging 14242 (2023) 225–234, https://doi.org/10.1007/978-3-031-45249-9_22.

[9] T. Salles, L. Rocha, M. Gonçalves, A bias-variance analysis of state-of-the-art random forest text classifiers, Adv. Data Anal. Classif. 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? Brief. Bioinform. 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] S. Oka, Y. Takefuji, Letter to the Editor regarding "prediction of PFAS bio-accumulation in different plant tissues with machine learning models based on molecular fingerprints", by Song et al. (2024), Sci. Total Environ. 950 (2025) 175091, Sci. Total Environ 984 (2025) 179714, https://doi.org/10.1016/j.scitotenv.2025.179714.

[15] B. Bilodeau, N. Jaques, P.W. Koh, B. Kim, Impossibility theorems for feature attribution, Proc. Natl. Acad. Sci. U. S. A. 121 (2) (2024) e2304406120, https://doi.org/10.1073/pnas.2304406120.

[16] 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.

[17] I. Kumar, C. Scheidegger, S. Venkatasubramanian, S. Friedler, Shapley residuals: quantifying the limits of the Shapley value for explanations, Adv. Neural Inf. Process. Syst. 34 (2021) 26598–26608.

[18] M.A. Lones, Avoiding common machine learning pitfalls, Patterns 5 (10) (2024) 101046, https://doi.org/10.1016/j.patter.2024.101046.

[19] 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, Eds., in: A. Holzinger, R. Goebel, R. Fong, T. Moon, K.R. Müller, W. Samek (Eds.), xxAI - Beyond Explainable AI, Springer, Cham, 2022, p. 4, https://doi.org/10.1007/978-3-031-04083-2_4.

[20] K. Okoye, S. Hosseini, Correlation tests in R: pearson Cor, Kendall's tau, and Spearman's rho, R Programming: Statistical Data Analysis in Research, in: K. Okoye, S. Hosseini (Eds.), Correlation tests in R: pearson Cor, Kendall's tau, and Spearman's rho, Springer Nature (2024) 247–277, https://doi.org/10.1007/978-981-97-3385-9_12.

[21] H. Yu, A.D. Hutson, A robust Spearman correlation coefficient permutation test, Commun. Stat. Theory M 53 (6) (2024) 2141–2153, https://doi.org/10.1080/03610926.2022.21211414.

[22] J.D. Gibson, Entropy and Mutual information, in: Information Theoretic Principles for Agent Learning, Springer, Cham, 2025, https://doi.org/10.1007/978-3-031-65388-9_2.

[23] T. Kerby, T. White, K.R. Moon, Learning local higher-order interactions with total correlation, Proc. 2024 IEEE 34th Int. Workshop Mach. Learn. Signal Process. (MLSP) (2024) 1–6. https://doi.org/10.1109/MLSP58920.2024.10734758.

[24] 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.

[25] 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,*], Nobuko Inoue[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), ninoue@sciencepark.co.jp (N. Inoue), takefuji@keio.jp (Y. Takefuji).