International Journal of Medical Informatics 208 (2026) 106238

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Limitations of SHAP-based interpretability in sepsis progression models and paths to more robust feature validation

To the Editor,

Zhou et al. developed machine learning models to predict progression from sepsis to septic shock using MIMIC-IV and eICU-CRD databases [1]. Their methodology encompassed constructing a cohort of adult ICU patients with sepsis, applying LASSO regression for feature selection, and training six supervised learning algorithms (logistic regression, naïve Bayes, random forests, support vector machines, XGBoost, and LightGBM). The MIMIC-IV dataset served as the training set with external validation performed on eICU data. SHAP analysis was employed to interpret feature contributions in their model.

However, their interpretability framework raises significant methodological concerns. Supervised machine learning models contain two distinct forms of accuracy: target prediction performance and feature importance reliability. While prediction accuracy can be validated against known outcomes, feature importance lacks a ground truth reference. Consequently, SHAP values characterize model behavior rather than genuine clinical associations. Research demonstrates these explanatory methods are sensitive to model architecture, predictor correlations, missing data patterns, and dataset perturbations—often producing unstable feature rankings across different algorithms. These vulnerabilities are particularly problematic in ICU datasets, where multicollinearity, non-random missingness, and treatment-dependent confounding represent common challenges [2–10].

To determine whether feature importance reflects true clinical relationships rather than computational artifacts, several validation criteria must be met [11–20]: consistency across cohorts and models, demonstration of dose–response relationships, biological plausibility, and appropriate temporal alignment between predictors and outcomes. Since supervised learning alone cannot fully establish these criteria—especially when outcome labels may be influenced by diagnostic delays, institutional practices, or intervention effects—feature importance should not rely solely on techniques like SHAP. Additionally, leave-top-1-out stress testing provides a direct means of probing robustness: after selecting features from the full set (set 1), the highest-ranked feature is removed to form a reduced dataset, and the top features are reselected (set 2). If importance rankings reflect true physiological relationships, this process should yield predictable and proportionate shifts. Erratic reordering instead suggests instability driven by correlated variables, model dependence, or label noise.

To enhance interpretability, we recommend supplementing the analysis with unsupervised feature stability approaches such as Feature Agglomeration and highly variable feature selection, complemented by non-targeted, nonlinear, nonparametric association tests like Spearman correlation with significance testing. These methods operate independently of outcome labels, helping avoid label-driven distortions and reducing model-specific artifacts. Integrating unsupervised and nonparametric techniques with supervised modeling would provide a more reliable understanding of variables truly driving progression from sepsis to septic shock.

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.

References

[1] H. Zhou, F. Li, X. Liu, Early prediction of septic shock in ICU patients using machine learning: Development, external validation, and explainability with SHAP, Int. J. Med. Inf. 206 (2026) 106169, https://doi.org/10.1016/j.ijmedinf.2025.106169.

[2] 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.

[3] 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, Statist. Meth. Appl. 32 (2023) 237–270, https://doi.org/10.1007/s10260-022-00643-4.

[4] 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.

[5] J.P. Ioannidis, Genetic associations: False or true? Trends Mol. Med. 9 (4) (2003) 135–138, https://doi.org/10.1016/S1471-4914(03)00030-3.

[6] J.P. Ioannidis, Why most discovered true associations are inflated, Epidemiology 19 (5) (2008) 640–648, https://doi.org/10.1097/EDE.0b013e31818131e7.

[7] Q. Lai, R. Dannenfelser, J.P. Roussarie, V. Yao, Disentangling associations between complex traits and cell types with seismic, Nat. Commun. 16 (1) (2025) 8744, https://doi.org/10.1038/s41467-025-63753-z.

[8] L.H. Nazer, R. Zatarah, S. Waldrip, J.X.C. Ke, M. Moukheiber, A.K. Khanna, et al., Bias in artificial intelligence algorithms and recommendations for mitigation, PLOS Digital Health 2 (6) (2023) e0000278, https://doi.org/10.1371/journal.pdig.0000278.

[9] D. Prada, B. Ritz, A.Z. Bauer, A.A. Baccarelli, Evaluation of the evidence on acetaminophen use and neurodevelopmental disorders using the Navigation Guide methodology, Environ. Health 24 (1) (2025) 56, https://doi.org/10.1186/s12940-025-01208-0.

[10] V. Prasad, A.B. Jena, Prespecified falsification end points: can they validate true observational associations? JAMA 309 (3) (2013) 241–242, https://doi.org/10.1001/jama.2012.96867.

[11] M.R. Roberts, S. Ashrafzadeh, M.M. Asgari, Research techniques made simple: Interpreting measures of association in clinical research, J. Invest. Dermatol. 139 (3) (2019) 502–511.e1, https://doi.org/10.1016/j.jid.2018.12.023.

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

[13] A.G. Sheik, A. Kumar, C.S. Srungavarapu, M. Azari, S.R. Ambati, F. Bux, A.K. Patan, Insights into the application of explainable artificial intelligence for biological wastewater treatment plants: Updates and perspectives, Eng. Appl. Artif. Intel. 144 (2025) 110132, https://doi.org/10.1016/j.engappai.2025.110132.

[14] A. Singh, K. Hatzikotoulas, N.W. Rayner, K. Suzuki, H.J. Taylor, L. Southam, et al., Correcting for genomic inflation leads to loss of power in large-scale genome-wide

https://doi.org/10.1016/j.ijmedinf.2025.106238
Received 11 December 2025; Received in revised form 18 December 2025; Accepted 22 December 2025
Available online 24 December 2025

association study meta-analysis, Genet. Epidemiol. 49 (6) (2025) e70016, https://doi.org/10.1002/gepi.70016.

[15] E. Stamatakis, M. Ahmadi, R.K. Biswas, B. Del Pozo Cruz, C. Thøgersen-Ntoumani, M.H. Murphy, A. Sabag, S. Lear, C. Chow, J.M.R. Gill, M. Hamer, Device-measured vigorous intermittent lifestyle physical activity (VILPA) and major adverse cardiovascular events: evidence of sex differences, Br. J. Sports Med. 59 (5) (2025) 316–324, https://doi.org/10.1136/bjsports-2024-108484.

[16] P.M. Steiner, Y. Kim, The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases, J. Causal Inference 4 (2) (2016) 20160009, https://doi.org/10.1515/jci-2016-0009.

[17] Y. Takefuji, Model-specific feature importances: Distinguishing true associations from target-feature relationships, J. Affect. Disord. 369 (2025) 390–391, https://doi.org/10.1016/j.jad.2024.10.019.

[18] J. Ugirumurera, E.A. Bensen, J. Severino, J. Sanyal, Addressing bias in bagging and boosting regression models, Sci. Rep. 14 (1) (2024) 18452, https://doi.org/10.1038/s41598-024-68907-5.

[19] B.R. Underwood, I. Lourida, J. Gong, S. Tamburin, E.Y.H. Tang, E. Sidhom, Deep Dementia Phenotyping (DEMON) Network. Data-driven discovery of associations between prescribed drugs and dementia risk: A systematic review Alzheimer's & Dementia 11 (1) (2025) 10.1002/trc2.70037 e70037.

[20] M.L. Wallace, L. Mentch, B.J. Wheeler, A.L. Tapia, M. Richards, S. Zhou, et al., Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction, BMC Med. Res. Method. 23 (1) (2023) 144, https://doi.org/10.1186/s12874-023-01965-x.

Yuto Arai[1,*], Yoshiyasu Takefuji[2]
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku,
Tokyo 135-8181, Japan

[*] Corresponding author.
E-mail addresses: 2550010@stu.musashino-u.ac.jp (Y. Arai),
takefuji@keio.jp (Y. Takefuji).

[1] ORCID: 0009-0006-3589-9994.

[2] ORCID: 0000-0002-1826-742X.