ELSEVIER

Journal of Clinical Epidemiology 178 (2025) 111619

Journal of Clinical Epidemiology

LETTER TO THE EDITOR

Addressing feature importance biases in machine learning models for early diagnosis of type 1 Gaucher disease

Tenenbaum et al proposed a machine learning model aimed at the early diagnosis of type 1 Gaucher disease [1]. Their methodology utilized multiple algorithms, including Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and logistic regression, to enhance diagnostic accuracy. Notably, they integrated LightGBM with SHAP (SHapley Additive exPlanations) to conduct a thorough analysis of feature importance [1]. However, it is important to note that although cross-validation is effective for assessing predictive accuracy, it does not serve as a validation method for feature importance analysis. This distinction is critical for understanding the robustness of the model's insights regarding the relative importance of various features.

Their algorithms face two critical challenges. First, feature importance measures derived from machine learning models such as RF, LightGBM, and logistic regression are inherently biased due to the model-specific nature of these algorithms [2–6]. This means that different models can yield varying feature importance scores, suggesting that the calculated feature importances may not accurately represent true associations between the target variable and the features. As a result, these measurements can mislead researchers about the underlying relationships within the data.

Second, SHAP inherits biases from the machine learning models it is derived from, which can further distort interpretations and conclusions drawn from the analysis [7]. Since SHAP relies on the output of these models with explain = SHAP (model), it is susceptible to the same biases that exist in the underlying feature importance measures. Consequently, this reliance can lead to erroneous conclusions and diminish the reliability of the findings.

To improve the robustness of their model, it is crucial to adopt methods that mitigate these biases. This may involve utilizing ensemble approaches that combine the strengths of multiple models or adopting alternative explanation techniques that are less model-dependent. By addressing these issues, researchers can enhance the validity of their conclusions and create more reliable diagnostic tools for type 1 Gaucher disease.

This paper advocates for using true associations employing robust statistical methods [8–10] such as Spearman's correlation with *P* values and/or Chi-squared tests with *P* values, bias-free approaches instead of biased feature importances from machine learning models. This paper reveals why RF and LightGBM induce feature importance biases.

RF, LightGBM, and logistic regression are popular machine learning models that offer valuable predictions. However, they can also introduce biases due to their algorithms and operational mechanics in features. RF assesses feature importance by measuring contributions to impurity reduction but can favor features with many levels, leading to misleading importance scores. Similarly, LightGBM's histogram-based algorithm tends to prioritize early splits, potentially sidelining equally important features, especially when handling categorical data. Logistic regression, assuming a linear relationship, can misrepresent nonlinear associations and is sensitive to feature scaling, causing further bias. High multicollinearity can also obscure true feature importance. Thus, although these models enhance predictive capabilities, understanding their biases is vital for interpreting results, improving the reliability of feature importance assessments, and ultimately guiding informed decision-making.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Conceptualization, Investigation, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

There are no competing interests for any author.

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

Data availability

No data was used for the research described in the article.

Funding: This research has no fund.

References

[1] Tenenbaum A, Revel-Vilk S, Gazit S, Roimi M, Gill A, Gilboa D, et al. A machine learning model for early diagnosis of type 1 Gaucher disease using real-life data. J Clin Epidemiol 2024;175:111517. https://doi.org/10.1016/j.jclinepi.2024.111517.

[2] Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinf 2007;8:25. https://doi.org/10.1186/1471-2105-8-25.

[3] Adnan MN. On reducing the bias of random forest. In: Chen W, Yao L, Cai T, Pan S, Shen T, Li X, editors. Advanced Data Mining and Applications. ADMA 2022. Lecture Notes in Computer Science, 13726. Cham: Springer; 2022. https://doi.org/10.1007/978-3-031-22137-8_14.

[4] Song J. Bias corrections for Random Forest in regression using residual rotation. J Korean Stat Soc 2015;44:321−6. https://doi.org/10.1016/j.jkss.2015.01.003.

[5] Fahad NM, Azam S, Montaha S, Hossain Mukta S. Enhancing cervical cancer diagnosis with graph convolution network: AI-powered segmentation, feature analysis, and classification for early detection. Multimed Tools Appl 2024;83:75343−67. https://doi.org/10.1007/s11042-024-18608-y.

[6] Chen J, Ooi LQRO, Tan TWK, Zhang S, Li J, Asplund CL, et al. Relationship between prediction accuracy and feature importance reliability: an empirical and theoretical study. Neuroimage 2023;274:120115. https://doi.org/10.1016/j.neuroimage.2023.120115.

[7] 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.

[8] Strzelecka A, Zawadzka D. The use of Chi-squared Automatic Interaction Detector (CHAID) analysis to identify characteristics of agricultural households at risk of financial self-exclusions. Procedia Comput Sci 2023;225:4443−52. https://doi.org/10.1016/j.procs.2023.10.442.

[9] Jiang J, Zhang X, Yuan Z. Feature selection for classification with Spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets. Expert Syst Appl 2024;249(Pt B):123633. https://doi.org/10.1016/j.eswa.2024.123633.

[10] Tyagi A, Salhotra R, Agrawal A, Vashist I, Malhotra RK. Use of Pearson and Spearman correlation testing in Indian anesthesia journals: an audit. J Anaesthesiol Clin Pharmacol 2023;39(4):550−6. https://doi.org/10.4103/joacp.joacp_13_22.

https://doi.org/10.1016/j.jclinepi.2024.111619