Journal Pre-proof

Reevaluating Feature Selection in Machine Learning Models for Identifying Disease-Modifying Agents in Obstructive Sleep Apnea

Yoshiyasu Takefuji Ph. D

PII: S0300-2896(25)00038-9

DOI: https://doi.org/doi:10.1016/j.arbres.2025.01.014

Reference: ARBRES 3735

To appear in: *Archivos de Bronconeumologia*

Received Date: 20 January 2025

Please cite this article as: Takefuji Y, Reevaluating Feature Selection in Machine Learning Models for Identifying Disease-Modifying Agents in Obstructive Sleep Apnea, *Archivos de Bronconeumología* (2025), doi: https://doi.org/10.1016/j.arbres.2025.01.014

Discussion letter

Title: Reevaluating Feature Selection in Machine Learning Models for Identifying Disease-Modifying Agents in Obstructive Sleep Apnea

Authors:
Yoshiyasu Takefuji[1]*

Affiliation:
[1]Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan.

*Corresponding Author: Yoshiyasu Takefuji, Ph. D., Professor, Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan.
ORCID: 0000-0002-1826-742X

TEXT:
Belmonte et al. conducted a comprehensive study on the synergistic integration of the miRNome, machine learning, and bioinformatics to identify potential disease-modifying agents for obstructive sleep apnea [1]. Their methodology involved variable selection utilizing random forests and sparse partial least squares. To enhance the robustness and consistency of their selection process, they implemented a rigorous approach by repeating each method 50 times [1].

While Belmonte et al. showcased innovative machine learning models for identifying potential disease-modifying agents for obstructive sleep apnea, this paper raises critical concerns regarding the reliance on feature selection from random forests due to the model-specific nature of feature importance. It is vital for Belmonte et al. to recognize that feature importance metrics derived from machine learning models can be inherently biased, as different models employ distinct methodologies for calculating these values due to the absence of ground truth values for accuracy validation [2,3].

Random forests, as a supervised machine learning approach, rely on ground truth values for validating prediction accuracy. However, the validation does not extend to the reliability of feature importance. High prediction accuracy does not necessarily indicate trustworthy feature importances; thus, caution is warranted when interpreting these metrics. Over 100 peer-reviewed articles discussed non-negligible biases in feature importances from models [2,3].

To accurately evaluate genuine associations between the target variable and the features, three critical elements must be considered: the distribution of the data, the statistical relationships among the variables, and the validation of statistical significance through p-values. Consequently, the choice between parametric and nonparametric methods, as well as linear and nonlinear approaches, plays a pivotal role in ensuring a robust and precise analysis.

In this context, the paper recommends the use of nonlinear and nonparametric robust statistical methods [4], such as Spearman's correlation and Kendall's tau, supplemented by p-values for significance testing. Additionally, before applying these statistical methods, it is essential to assess multicollinearity by executing the Variance Inflation Factor (VIF) analysis, which can help to eliminate features with collinearity and interactions, thereby reducing feature inflation [5].

While Belmonte et al. made significant strides in integrating the miRNome, machine learning, and bioinformatics for identifying disease-modifying agents in obstructive sleep apnea, it is crucial to address concerns regarding the reliability of feature selection from random forests. Acknowledging the inherent biases in feature importance metrics derived from machine learning models—especially in the absence of ground truth values for validation—is vital for producing credible results. To ensure accurate associations between the target variable and features, careful consideration of data distribution, variable relationships, and statistical significance is essential. By employing robust statistical methods, such as Spearman's correlation and Kendall's tau, alongside VIF analysis for multicollinearity, they can enhance the rigor and validity of their findings in this evolving field.

Declarations
Funding: This research has no fund.
Conflicts of interest/Competing interests: The author has no conflict of interest.
Ethics approval: Not applicable
Consent to participate: Not applicable
Consent for publication: Not applicable
Availability of data and material: Not applicable
Code availability: Not applicable
Authors' contributions: Yoshiyasu Takefuji completed this research and wrote this article
Artificial intelligence involvement: Not applicable

References
[1[Belmonte T, Benitez ID, García-Hidalgo MC, Molinero M, Pinilla L, Mínguez O, Vaca R, Aguilà M, Moncusí-Moix A, Torres G, Mediano O, Masa JF, Masdeu MJ, Montero-San-Martín B, Ibarz M, Martinez-Camblor P, Gómez-Carballa A, Salas A, Martinón-Torres F, Barbé F, Sánchez-de-la-Torre M, de Gonzalo-Calvo D. Synergic Integration of the miRNome, Machine Learning and Bioinformatics for the Identification of Potential Disease-Modifying Agents in Obstructive Sleep Apnea. Arch Bronconeumol. 2024; ISSN 0300-2896. doi:10.1016/j.arbres.2024.11.011.
[2] 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.
[3] Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007;8:25. doi:10.1186/1471-2105-8-25

[4] Okoye, K., Hosseini, S. (2024). Correlation Tests in R: Pearson Cor, Kendall's Tau, and Spearman's Rho. In: R Programming. Springer, Singapore. https://doi.org/10.1007/978-981-97-3385-9_12
[5] Salmerón-Gómez, R., García-García, C.B. & García-Pérez, J. A Redefined Variance Inflation Factor: Overcoming the Limitations of the Variance Inflation Factor. Comput Econ 65, 337–363 (2025). https://doi.org/10.1007/s10614-024-10575-8