EUROPEAN SOCIETY OF RADIOLOGY
European Radiology

LETTER TO THE EDITOR

Letter to the Editor: Navigating bias in machine learning—reevaluating feature importances through robust statistical analysis

Yoshiyasu Takefuji[1*]

Dear Editor,

In their paper, Jiang et al introduced the interpretable deep learning radiomics model (IDLR) for diagnosing various stages of Alzheimer's disease and predicting the progression from mild cognitive impairment to Alzheimer's disease [1]. By leveraging deep learning features, the IDLR identifies robust imaging biomarkers, thereby enhancing treatment strategies. Additionally, SHAP (SHapley Additive exPlanations) feature importance analysis shows that the model reliably ranks feature contributions across multiple clinical cohorts [1].

However, reliance on SHAP within machine learning frameworks can introduce model-specific biases that distort feature importance assessments [2, 3]. The feature importances yielded by deep learning algorithms may be skewed, potentially leading to erroneous conclusions about the actual relationships in the data [2–6]. While deep learning excels at making precise predictions based on known ground truth values, high accuracy in target predictions does not necessarily imply accurate assessments of feature importance. This discrepancy stems from the lack of corresponding ground truth values to validate the accuracy of the feature importance measures. In contrast to machine learning models, which assign feature importances on a scale from 0 to 1, Spearman's correlation, accompanied by *p*-values, ranges from −1 to 1. This range provides directional information and reflects the nonlinear, nonparametric nature of the relationships being analyzed. This difference in scaling highlights the need for cautious interpretation of feature importance derived from machine learning models.

This paper discusses inherent biases in machine learning-generated feature importances and stresses the importance of using robust statistical methods to focus on authentic associations between target variables and features. Techniques such as Spearman's correlation [7, 8] provide non-parametric insights that operate independently of machine learning models, ensuring more dependable results. Researchers must distinguish between the predictive capabilities of machine learning algorithms and the often-biased feature importances, which can misrepresent true relationships due to the models' specificities.

Several factors contribute to biased feature importance assessments in machine learning models, particularly deep learning. The complexity and non-linear relationships inherent in deep learning architectures complicate the identification of individual feature contributions. Calculated feature importances may, therefore, misrepresent true influences, rather than elucidate them.

Interactions between features further complicate these assessments. Many models, including ensemble methods and deep learning approaches, can implicitly capture feature interactions, making it challenging to isolate a single feature's significance. What may appear less important in isolation can gain relevance when considered alongside other features, leading to potentially misleading interpretations.

In conclusion, while machine learning, particularly deep learning, provides powerful tools for prediction, caution is necessary in interpreting feature importances. Addressing biases from model architecture, measurement methods, data characteristics, and feature relationships is crucial for improving model robustness and interpretability. The adoption of bias-free statistical methods is advocated to ensure feature importances accurately reflect true relationships within the data [7, 8].

This letter refers to the article available at https://doi.org/10.1007/s00330-024-11158-9. A reply to this letter is available at https://doi.org/10.1007/s00330-025-11798-5.

*Correspondence:
Yoshiyasu Takefuji
takefuji@keio.jp
[1]Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

Springer

Funding
This research received no funding.

Code availability
Not applicable

Compliance with ethical standards

Conflict of interest
The author declares no conflict of interest.

Informed consent
Not applicable

Ethics approval
Not applicable

Consent for publication
Not applicable

Study subjects or cohorts overlap
Not applicable

Methodology

- Letter to the editor

Received: 1 November 2024 Revised: 19 December 2024 Accepted: 24 January 2025
Published online: 04 July 2025

References

1. Jiang J, Li C, Lu J et al (2024) Using interpretable deep learning radiomics model to diagnose and predict progression of early AD disease spectrum: a preliminary [18F]FDG PET study. Eur Radiol https://doi.org/10.1007/s00330-024-11158-9
2. Bilodeau B, Jaques N, Koh PW, Kim B (2024) Impossibility theorems for feature attribution. Proc Natl Acad Sci USA 121: e2304406120. https://doi.org/10.1073/pnas.2304406120
3. Chen V, Yang M, Cui W, Kim JS, Talwalkar A, Ma J (2024) Applying interpretable machine learning in computational biology-pitfalls, recommendations and opportunities for new developments. Nat Methods 21:1454–1461. https://doi.org/10.1038/s41592-024-02359-7
4. Chen J, Ooi LQR, Tan TWK et al (2023) Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study. Neuroimage 274: 120115. https://doi.org/10.1016/j.neuroimage.2023.120115
5. Krawczuk J, Łukaszuk T (2016) The feature selection bias problem in relation to high-dimensional gene data. Artif Intell Med 66:63–71. https://doi.org/10.1016/j.artmed.2015.11.001
6. Fisher A, Rudin C, Dominici F (2019) 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 20:177
7. Jiang J, Zhang X, Yuan Z (2024) Feature selection for classification with Spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets. Expert Syst Appl 249: 123633. https://doi.org/10.1016/j.eswa.2024.123633
8. Yu H, Hutson AD (2024) A robust Spearman correlation coefficient permutation test. Commun Stat Theory Methods 53:2141–2153. https://doi.org/10.1080/03610926.2022.2121144

Publisher's Note