Sleep Medicine Reviews 83 (2025) 102152

Contents lists available at ScienceDirect

Sleep Medicine Reviews

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

Beyond explainable AI: Enhancing trust and robustness in machine learning for sleep apnea diagnosis[☆]

ARTICLE INFO

Handling editor: Monica Andersen

Keywords:
Sleep apnea diagnosis
Explainable AI limitations
Feature importance reliability
Unsupervised machine learning
Nonparametric statistics

ABSTRACT

Kilic et al. reviewed machine learning (ML) and deep learning (DL) for sleep apnea detection, emphasizing explainable AI (XAI) while noting challenges like Apnea-Hypopnea Index (AHI) discrepancies. This paper extends their critique, arguing that XAI tools like SHAP inherit model biases, and high prediction accuracy does not guarantee reliable feature importances, which inherently lack ground truth validation. To overcome these limitations and build clinical trust, we advocate for a comprehensive approach combining unsupervised ML (e.g., feature agglomeration, highly variable gene selection) with nonlinear nonparametric statistical methods (e.g., Spearman's correlation). This strategy robustly evaluates variable relationships and p-values, particularly for monotonic associations, mitigating misapplications stemming from assumption violations and inadequate interpretation of model ground truth, thus fostering real-world applicability.

Kilic et al. conducted a systematic review and meta-analysis on the diagnostic accuracy of machine learning (ML) and deep learning (DL) algorithms for detecting sleep apnea via electrocardiograms [1]. They underscored the vital importance of evaluating sensitivity and specificity, noting that high sensitivity minimizes missed apnea events while high specificity reduces false positives that could lead to unnecessary follow-up testing. By pooling individual study results, they were able to estimate overall diagnostic performance and examine sources of heterogeneity, such as differences in study design, patient populations, and signal preprocessing pipelines. To address the often-criticized black-box nature of DL models, the authors advocated for the integration of explainable AI methods such as SHAP, LIME, and GradCAM. These approaches assign contribution scores to input features or highlight salient regions in input signals, with the aim of making model decisions more transparent to clinicians. However, Kilic et al. also identified significant challenges—most notably, variability in Apnea-Hypopnea Index (AHI) thresholds used to define sleep apnea severity, the reliance on ICD coding rather than standardized polysomnography scoring, and the under-representation of subclinical or borderline cases. Such inconsistencies can limit the comparability of model performance across studies and erode physician trust, ultimately hindering real-world deployment [1].

Furthermore, while Kilic et al. highlighted the limitations of explainable AI techniques like SHAP in providing clarity about true variable relationships, primarily because SHAP values rely on estimated conditional expectations rather than known distributions, this paper advocates for a more comprehensive strategy. We propose combining unsupervised machine learning methods such as feature agglomeration and selection of highly variable genes or signal features with nonlinear, nonparametric statistical tools like Spearman's rank order correlation with p-values. Feature agglomeration first groups features that exhibit strong pairwise correlations, thereby reducing dimensionality and highlighting coherent clusters of physiological or genetic signals. In high-dimensional data settings, selecting features with the greatest variance focuses analysis on the most informative dimensions and mitigates noise from low-variance measurements. Following feature reduction, Spearman's correlation tests for monotonic relationships without assuming linearity or normality. Because it uses rank information, Spearman's rho remains robust to outliers and skewed distributions, and exact or permutation-based p values enable rigorous control of false discovery rates. By integrating these unsupervised and statistical methods alongside explainable AI, one can distinguish model-driven attributions from genuine data-driven associations.

Supervised ML models, including tree-based ensembles and DL networks, typically benefit from ground-truth labels—such as the presence or absence of sleep apnea confirmed by polysomnography—that enable calculation of sensitivity, specificity, accuracy, and AUC. In contrast, the feature importance scores extracted post hoc from these models do not have an objective ground truth against which to validate their correctness. This distinction gives rise to two separate notions of accuracy in supervised learning: the accuracy of outcome predictions and the reliability of feature importance measures. A model may achieve excellent prediction accuracy yet still assign misleading importance to features, whether due to confounding variables, overfitting on spurious correlations, or biases inherited from the training data [2–10]. Although explainable AI tools such as SHAP improve model transparency by attributing outputs to inputs, they are inherently model-specific and thus will propagate any biases present in the trained model [11–18]. In

[☆] **According to ScholarGPS,** Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 54th out of 395,884 scholars in neural networks (AI), 23rd out of 47,799 in parallel computing, and 14th out of 7222 in parallel algorithms. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

https://doi.org/10.1016/j.smrv.2025.102152
Received 21 July 2025; Accepted 3 August 2025
Available online 12 August 2025

practice, this means that reliance on a single explainability method can lead to overconfidence in features that are artifacts of the algorithm or dataset rather than true physiological drivers of sleep apnea.

This paper delineates three categories of methodological misapplication that can result in flawed conclusions and diminished clinical utility. First, violation of statistical or explainability tool assumptions, for example, using a test that presumes independent observations when data are longitudinal or neglecting to verify distributional requirements, can invalidate p values, confidence intervals, and attribution scores. Second, misinterpretation of model-derived ground truths occurs when practitioners treat feature importance scores or model outputs as if they were directly measured biological effects, without accounting for label noise, measurement error, or sampling bias. This is the category most in need of scrutiny, as it directly undermines the trustworthiness of scientific inferences and clinical decisions. Third, critical preprocessing errors—such as applying scaling or normalization inconsistently across training, validation, and test sets or performing feature selection on the entire dataset rather than within cross-validation folds—can introduce data leakage and artificially inflate both performance metrics and feature importance estimates. By addressing these misapplications and adopting a complementary toolkit of explainable AI, unsupervised clustering, and robust nonparametric inference, researchers can develop sleep apnea detection models that deliver not only high prediction accuracy but also dependable, interpretable insights for clinical practice.

Consent to participate

Not applicable.

Ethics approval

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.

AI use

Not applicable.

Funding

This research has no fund.

Conflicts of interest/competing interest

The author has no conflict of interest.

References

[1] Kilic ME, Arayici ME, Turan OE, Yilancioglu YR, Ozcan EE, Yilmaz MB. Diagnostic accuracy of machine learning algorithms in electrocardiogram-based sleep apnea detection: a systematic review and meta-analysis. Sleep Med Rev 2025;81:102097. https://doi.org/10.1016/j.smrv.2025.102097.

[2] Parr T, Hamrick J, Wilson JD. Nonparametric feature impact and importance. Inf Sci 2024;653:119563. https://doi.org/10.1016/j.ins.2023.119563.

[3] Watson DS, Wright MN. Testing conditional independence in supervised learning algorithms. Mach Learn 2021;110(8):2107–29. https://doi.org/10.1007/s10994-021-06030-6.

[4] Molnar C, König G, Herbinger J, et al. General pitfalls of model-agnostic interpretation methods for machine learning models. Springer International Publishing; 2022. https://doi.org/10.1007/978-3-031-04083-2_4.

[5] Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. ACM Queue 2018;16(3):31–57. https://doi.org/10.1145/3236386.3241340.

[6] 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.

[7] Lenhof K, Eckhart L, Rolli LM, Lenhof HP. Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. Briefings Bioinf 2024;25(5):bbae379. https://doi.org/10.1093/bib/bbae379.

[8] Mandler H, Weigand B. A review and benchmark of feature importance methods for neural networks. ACM Comput Surv 2024;56(12):318. https://doi.org/10.1145/3679012.

[9] Potharlanka JL, Bhat MN. Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms. Sci Rep 2024;14(1):2923. https://doi.org/10.1038/s41598-024-53141-w.

[10] Wood D, Papamarkou T, Benatan M, et al. Model-agnostic variable importance for predictive uncertainty: an entropy-based approach. Data Min Knowl Discov 2024;38:4184–216. https://doi.org/10.1007/s10618-024-01070-7.

[11] Wu L. A review of the transition from Shapley values and SHAP values to RGE. Statistics 2025:1–23. https://doi.org/10.1080/02331888.2025.2487853.

[12] Bilodeau B, Jaques N, Koh PW, Kim B. Impossibility theorems for feature attribution. Proc Natl Acad Sci USA 2024;121(2):e2304406120. https://doi.org/10.1073/pnas.2304406120.

[13] Huang X, Marques-Silva J. On the failings of Shapley values for explainability. Int J Approx Reason 2024;171:109112. https://doi.org/10.1016/j.ijar.2023.109112.

[14] Hooshyar D, Yang Y. Problems with SHAP and LIME in interpretable AI for Education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. IEEE Access 2024;12:137472–90. https://doi.org/10.1109/ACCESS.2024.3463948.

[15] Lones MA. Avoiding common machine learning pitfalls. Patterns 2024;5(10):101046. https://doi.org/10.1016/j.patter.2024.101046.

[16] Molnar C, et al. General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W, editors. xxAI - beyond Explainable AI. xxAI, 13200. Lecture Notes in Computer Science. Springer; 2020. https://doi.org/10.1007/978-3-031-04083-2_4. 2022.

[17] Kumar I, Scheidegger C, Venkatasubramanian S, Friedler S. Shapley residuals: quantifying the limits of the shapley value for explanations. Adv Neural Inf Process Syst 2021;34:26598–608.

[18] Létoffé O, Huang X, Marques-Silva J. Towards trustable SHAP scores. In: Proceedings of the AAAI conference on artificial intelligence. 39; 2025. p. 18198–208. https://doi.org/10.1609/aaai.v39i17.34002. 17.

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