Annals of Epidemiology xxx (xxxx) xxx

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.sciencedirect.com/journal/annals-of-epidemiology

Letter to the Editor

Towards reliable feature interpretation in machine learning-based longevity prediction

ARTICLE INFO

Keywords:
Extreme longevity prediction
Machine learning
Feature importance
Model interpretability
Statistical validation

Dor Atias et al. used machine learning (ML) models to predict extreme longevity and compared them with traditional methods [1]. Their work represents a significant contribution to the field, though several methodological aspects warrant further discussion. They employed three prediction models: logistic regression (LR), a traditional statistical method, and two machine learning (ML) models—generalized least absolute shrinkage and selection operator (LASSO) regression and extreme gradient boosting (XGBoost). They designed a modeling pipeline to predict extreme longevity using a cohort of approximately 10,000 men. Notably, XGBoost achieved the highest predictive performance with an ROC-AUC of 0.72, outperforming LASSO (0.71) and LR (0.69). Their analysis highlighted systolic blood pressure, smoking status, and a history of myocardial infarction as the most influential factors in predicting longevity.

While the high predictive accuracy of the XGBoost model is noteworthy, it does not necessarily ensure the reliability of its feature importance rankings. This disconnect between predictive performance and interpretive validity has been highlighted in numerous studies, as detailed in the supplementary material, which includes an extensive list of supporting references [2–4]. Notably, the combined use of tree-based algorithms like XGBoost and regularized linear models such as LASSO regression can undermine the stability and consistency of feature importance estimates [5–9]. Gradient boosting decision trees (GBDT), including XGBoost, are known to introduce biases in feature importance due to the nature of their tree-building process. Features selected for early splits are often overemphasized, regardless of their actual predictive value. Consequently, the resulting importance scores may reflect what optimizes model performance rather than what truly drives extreme longevity, potentially distorting the roles of factors such as systolic blood pressure, smoking status, and history of myocardial infarction.

These concerns are further amplified by the use of SHapley Additive exPlanations (SHAP) and similar feature attribution methods, which are commonly employed to interpret model outputs [10–13]. Because these explanations are derived directly from the model's predictions, they inherently reflect any biases present in the underlying algorithm. As a result, even models with comparable predictive accuracy may assign markedly different levels of importance to individual features, underscoring the instability of feature relevance across models. This issue of model dependence becomes particularly evident in the study's focus on non-linear thresholds derived from SHAP analysis. For instance, patterns such as a decline in predicted longevity when diastolic blood pressure exceeds 93 mmHg or the identification of an HDL level above 42 mg/dl as a favorable cutoff highlight the interpretive challenges of post hoc explanations. Although these thresholds may appear clinically meaningful, they could also represent algorithmic artifacts created by the optimal partitioning strategy of Gradient Boosting Decision Trees (GBDT) under the constraints of the dataset. Therefore, such findings should be interpreted with caution, and independent validation is essential to confirm their biological plausibility and robustness. These interpretative challenges are further exacerbated by the methodological limitations of the benchmark model itself, logistic regression [14–19]. As a parametric model, it assumes a strictly linear relationship between input variables and the target outcome—an assumption often inadequate for capturing the complex, non-linear interactions typical of epidemiological data.

Beyond these issues, a central challenge in interpreting ML models lies in validating feature importance in the absence of ground truth. Without a definitive reference, importance rankings remain highly model-dependent and susceptible to bias. This limitation is particularly evident in the present study, which involves complex, high-dimensional, and collinear features. Such characteristics not only hinder interpretability but also obscure the individual contributions of features, increasing the risk of overfitting and reducing the generalizability of findings. Moreover, the instability of importance measures is further exacerbated by the intricate structure of the feature space, making it difficult to draw robust, causally meaningful conclusions from the model outputs.

To address these methodological limitations and improve predictive reliability, a more comprehensive analytical framework is needed. This framework should account for the complexity of the data and incorporate methods that can capture non-linear patterns and reduce multicollinearity. Such a framework is particularly important for validating interaction effects reported in the original study, including the

DOI of original article: https://doi.org/10.1016/j.annepidem.2026.01.006.

https://doi.org/10.1016/j.annepidem.2026.01.005
Received 27 August 2025; Received in revised form 10 December 2025; Accepted 13 January 2026
Available online 15 January 2026

interpretation that a positive correlation between systolic blood pressure (SBP) and body mass index (BMI) may attenuate SBP's adverse impact on longevity. Determining whether this interaction reflects a genuine biological phenomenon or a model-driven artifact demands independent verification. Unsupervised techniques such as Feature Agglomeration and Highly Variable Gene Selection offer robust options for dimensionality reduction and feature prioritization, enabling the identification of stable associations [20,21]. In addition, non-parametric methods like Spearman's rho and Kendall's tau allow independent assessment of rank correlations among key variables (e.g., SBP, BMI, and longevity) without relying on model-specific assumptions [22,23]. These approaches not only improve interpretability but also strengthen reproducibility—an essential requirement for translating statistical findings into actionable insights in epidemiological research.

In conclusion, although machine learning techniques such as XGBoost and feature contribution analysis provide strong predictive capabilities, they are prone to biases that can limit interpretability, especially in complex applications like longevity prediction. To mitigate these limitations, it is important to combine ML with robust statistical methods that support dimensionality reduction and validation. This integrated approach enhances the reliability of predictive assessments and enables insights that are both accurate and grounded in biological and epidemiological reality. Future research should aim to develop frameworks that maintain a balance between predictive accuracy, interpretability, and biological relevance.

CRediT authorship contribution statement

Yoshiki Takahashi: Investigation. **Souichi Oka:** Writing – original draft, Conceptualization. **Yoshiyasu Takefuji:** Writing – review & editing, Supervision, Project administration.

Funding source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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.

Acknowledgements

We extend our sincere gratitude to Kiyo Yoshida and Ryota Ono of Science Park, Inc. for their invaluable assistance with the extensive literature review.

Data Availability

No new data were generated or analyzed in support of this research.

References

[1] Atias D, Ashri S, Goldbourt U, Benyamini Y, Gilad-Bachrach R, Hasin T, et al. Machine learning in epidemiology: an introduction, comparison with traditional methods, and a case study of predicting extreme longevity. Ann Epidemiol 2025; 110:23–33. https://doi.org/10.1016/j.annepidem.2025.07.024.

[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. https://doi.org/10.48550/arXiv.1801.01489.

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

[4] 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. Brief Bioinform 2024;25:bbae379. https://doi.org/10.1093/bib/bbae379.

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

[6] Adler AI, Painsky A. Feature importance in gradient boosting trees with cross-validation feature selection. Entropy 2022;24(5):687. https://doi.org/10.3390/e24050687.

[7] Alaimo Di Loro P, Scacciatelli D, Tagliaferri G. 2-step gradient boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy. Stat Methods Appl 2023;32:237–70. https://doi.org/10.1007/s10260-022-00643-4.

[8] Wüthrich K, Zhu Y. Omitted variable bias of Lasso-based inference methods: a finite sample analysis. Rev Econ Stat 2023;105(4):982–97. https://doi.org/10.1162/rest_a_01128.

[9] Jain R, Xu W. HDSI: High dimensional selection with interactions algorithm on feature selection and testing. PLoS One 2021;16(2):e0246159. https://doi.org/10.1371/journal.pone.0246159.

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

[11] 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.

[12] 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.

[13] 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.

[14] Dey D, Haque MS, Islam MM, Aishi UI, Shammy SS, Mayen MSA, Noor STA, Uddin MJ. The proper application of logistic regression model in complex survey data: a systematic review. BMC Med Res Methodol 2025;25:15. https://doi.org/10.1186/s12874-024-02454-5.

[15] Pinheiro-Guedes L, Martinho C, Martins MR. Logistic regression: limitations in the estimation of measures of association with binary health outcomes. Acta Med Port 2024;37(10):697–705. https://doi.org/10.20344/amp.21435.

[16] Wang T, Tang W, Lin Y, Su W. Semi-supervised inference for nonparametric logistic regression. Stat Med 2023;42(15):2573–89. https://doi.org/10.1002/sim.9737.

[17] van Maanen L, Katsimpokis D, van Campen AD. Fast and slow errors: logistic regression to identify patterns in accuracy–response time relationships. Behav Res Methods 2019;51:2378–89. https://doi.org/10.3758/s13428-018-1110-z.

[18] Rifada M, Chamidah N, Ningrum RA. Estimation of nonparametric ordinal logistic regression model using generalized additive models (GAM) method based on local scoring algorithm. AIP Conf Proc 2022;2668(1):070013. https://doi.org/10.1063/5.0111771.

[19] Work JW, Ferguson JG, Diamond GA. Limitations of a conventional logistic regression model based on left ventricular ejection fraction in predicting coronary events after myocardial infarction. Am J Cardiol 1989;64(12):702–7. https://doi.org/10.1016/0002-9149(89)90751-0.

[20] Zhang J, Wu X, Hoi SCH, Zhu J. Feature agglomeration networks for single stage face detection. Neurocomputing 2020;380:180–9. https://doi.org/10.1016/j.neucom.2019.10.087.

[21] Xie Y, Jing Z, Pan H, Xu X, Fang Q. Redefining the high variable genes by optimized LOESS regression with positive ratio. BMC Bioinforma 2025;26:104. https://doi.org/10.1186/s12859-025-06112-5.

[22] Okoye K, Hosseini S. Correlation tests in R: Pearson Cor, Kendall's Tau, and Spearman's Rho. In: Okoye K, Hosseini S, editors. R Programming: Statistical Data Analysis in Research. Springer Nature; 2024. p. 247–77. https://doi.org/10.1007/978-981-97-3385-9_12.

[23] Yu H, Hutson AD. A robust Spearman correlation coefficient permutation test. Commun Stat Theory Methods 2024;53:2141–53. https://doi.org/10.1080/03610926.2022.2121144.

Souichi Oka*, Yoshiki Takahashi

Science Park Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa 252-0029, Japan

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

E-mail address: takefuji@keio.jp.

* Corresponding author.

E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka), ytakahashi@sciencepark.co.jp (Y. Takahashi).