Check for updates

Diabetes, Obesity and Metabolism

WILEY

LETTER TO THE EDITOR

Feature Importance Bias and Competing Risks: Methodological Concerns in Ensemble Models for Diabetic Kidney Disease

Souichi Oka[1] | Kota Takemura[2] | Yoshiyasu Takefuji[2]

[1]Faculty of Data Science, Musashino University, Koto-ku, Tokyo, Japan | [2]Science Park Corporation, Zama, Kanagawa, Japan

Correspondence: Souichi Oka (s-oka@musashino-u.ac.jp)

Received: 26 May 2026 | **Revised:** 6 July 2026 | **Accepted:** 15 July 2026

Handling Editor: Richard Donnelly

To the Editor,

Tourkmani et al. developed a novel literature-informed ensemble framework, integrating Light Gradient Boosting Machine (LightGBM) and CoxBoost, to predict the three-year risk of diabetic kidney disease (DKD). Recognizing the substantial clinical and economic burden of DKD globally, their methodology utilized an extensive cohort of 18 742 patients with type 2 diabetes derived from a Saudi Arabian registry [1]. By strategically employing Bayesian multiple imputation to handle missing variables through the integration of external evidence and utilizing SHapley Additive exPlanation (SHAP) for model interpretability, they identified key predictors such as estimated glomerular filtration rate (eGFR) and albumin-creatinine ratio (ACR). The ensemble models demonstrated substantial performance, achieving an area under the receiver operating characteristic curve (AUROC) of 0.852 and a calibration slope of 0.98 across both temporal and multi-trial validation analyses. However, while their comprehensive approach demonstrates strong predictive capability and readiness for clinical implementation as a decision-support tool, the methodological utility and reliability of the resulting feature importance warrant further discussion.

This concern stems from underlying analytical pipeline artefacts that may inadvertently limit the clinical translatability and therapeutic actionability of such sophisticated risk models [2, 3]. Overreliance on global model accuracy metrics to justify variable relevance is a well-documented methodological challenge in medical machine learning. A substantial body of literature, comprising over 300 peer-reviewed articles, has consistently demonstrated that superior predictive performance does not necessarily equate to correct causal attribution [4]. Importantly, predictive performance, feature attribution and causal inference represent distinct analytical concepts. Measures of model accuracy evaluate how well outcomes are predicted, whereas feature attribution methods such as SHAP describe the contribution of variables to model predictions; neither should be interpreted as direct evidence of causal relationships. Machine learning algorithms are fundamentally designed to optimize loss minimization within a given dataset; thus, they may exploit statistical shortcuts, proxy variables or unmeasured confounders that merely correlate with the outcome, rather than capturing true pathophysiological relationships driving disease progression. Consequently, heavily relying on high-ranking features may not directly translate to causal or therapeutic inference without additional validation, particularly if a variable's algorithmic contribution to a particular prediction task is conflated with its actual biological aetiology.

Tree-based algorithms, notably the LightGBM component utilized by Tourkmani et al. (which carries a dominant weight of 64% in the final ensemble), tend to emphasize continuous variables and those possessing high cardinality [5]. Because continuous variables like baseline eGFR and ACR offer numerous potential thresholds for early node splits, they can accumulate substantial importance solely due to their mathematical structure. This architectural tendency is pronounced in high-dimensional clinical datasets characterized by complex multicollinearity [6]. In addition, greedy splitting algorithms frequently experience potential masking among correlated predictors. When strongly correlated clinical variables are present, the algorithm may

© 2026 John Wiley & Sons Ltd.

All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies.

disproportionately assign predictive importance to one variable while mathematically suppressing biologically equivalent or clinically relevant predictors. Within the proposed 25-feature 'Total Model' evaluated in this cohort, top-ranked predictors frequently share complex correlations with various included metabolic, demographic and clinical markers. Their dominant placement may therefore largely reflect the algorithm's internal partitioning mechanics and sensitivity to variable types rather than their independent prognostic value, necessitating careful consideration absent definitive ground truth regarding their direct causal roles.

Crucially, the current ensemble approach introduces methodological complexities regarding the statistical handling of competing risks, primarily mortality and kidney transplantation—within a longitudinal cohort of patients with type 2 diabetes. In this demographic, cardiovascular mortality represents a markedly prevalent competing event that can permanently preclude the clinical observation of DKD onset. The framework employs a dichotomous handling of identical clinical events: treating them as non-events in the binary LightGBM component, yet as censoring events in the CoxBoost component (weighted at 36%) [7]. This inconsistent approach differs from formal competing-risk approaches and may influence absolute risk estimates. By treating competing mortality as independent censoring, the survival component implicitly models a hypothetical population where death does not occur. While they note the use of Fine-Gray sensitivity analysis, the primary reliance on this independent censoring assumption within the main ensemble may affect absolute risk estimation rather than fully characterizing cumulative incidence under competing events. Additionally, the manuscript does not detail a formal assessment of proportional hazards assumptions. Although the models achieved excellent calibration in the reported validations, this dual handling of competing events can reduce the transportability of absolute risk estimates to distinct populations with higher baseline mortality, potentially distorting the features derived from this modelled risk landscape [8]. Future studies should consider formal competing-risk approaches, such as Fine–Gray subdistribution hazard models and cause-specific hazard models, to avoid potential bias arising from treating mortality and kidney transplantation as independent censoring events. These methods may provide more reliable estimates of cumulative DKD risk and predictor effects in populations with substantial competing-event burdens.

While SHAP provides model-based feature importance, its outputs remain inherently dependent on the underlying model and its assumptions. The very syntax of 'SHAP(model)' highlights that the explanation is conditional on the fitted model being explained and therefore necessarily inherits its structure, biases and data-dependent patterns. In the present framework, feature rankings may also reflect the influence of literature-informed priors introduced during Bayesian imputation rather than entirely independent algorithmic discovery. Accordingly, SHAP explanations should be interpreted as descriptions of model behaviour rather than independent evidence of pathophysiological importance or causality [9]. In the absence of public access to the dataset, we instead cite prior empirical studies showing that feature importance measures may not reliably capture true underlying relationships, even in well-performing models [10]. This does not critique the original study but highlights a general limitation of model-dependent attribution methods in complex clinical data.

To ensure broad clinical applicability and generalizability across diverse populations, future analytical frameworks must move beyond strictly model-dependent attribution techniques. Unsupervised models such as Feature Agglomeration and Highly Variable Gene Selection offer valuable approaches to dimensionality reduction, effectively grouping strongly correlated features or isolating the most informative predictors into clinically meaningful subsets [11, 12]. These algorithms proactively mitigate multicollinearity prior to the modelling phase and prevent the arbitrary masking of important clinical variables. Complementing these techniques with established non-parametric statistical methods, such as Spearman's rho and Kendall's tau, enables the detection of non-linear and monotonic associations that are strictly independent of algorithm architecture or hyperparameter tuning [13]. By systematically integrating these strategies, researchers can establish a more objective, transparent and structurally sound foundation for identifying meaningful clinical predictors, thereby improving interpretability and translatability in cardiovascular and renal machine learning research (Supporting Information).

Author Contributions

Souichi Oka: writing – original draft, conceptualization. **Kota Takemuras:** investigation. **Yoshiyasu Takefuji:** writing – review and editing, supervision, project administration.

Funding

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/dom.71140.

References

1. A. M. Tourkmani, T. J. Al-Harbi, A. A. Alghamdi, I. M. Youzghadli, F. S. Alosaimi, and A. Y. Azzam, "Literature-Informed Ensemble Machine Learning for Three-Year Diabetic Kidney Disease Risk Prediction in Type 2 Diabetes: Development, Validation, and Deployment of the PSMMC NephraRisk Model," *Diabetes, Obesity and Metabolism* 28 (2025): 1996–2026, https://doi.org/10.1111/dom.70385.

2. Z. C. Lipton, "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery," *ACM Queue* 16, no. 3 (2018): 31–57, https://doi.org/10.1145/3236386.3241340.

3. K. Lenhof, L. Eckhart, L. M. Rolli, and H. P. Lenhof, "Trust Me if You Can: A Survey on Reliability and Interpretability of Machine Learning

14631326, 0, Downloaded from https://dom-pubs.pericles-prod.literatumonline.com/doi/10.1111/dom.71140 by yoshiyasu takefuji - Keio University , Wiley Online Library on [21/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Approaches for Drug Sensitivity Prediction in Cancer," *Briefings in Bioinformatics* 25, no. 5 (2024): bbae379, https://doi.org/10.1093/bib/bbae379.

4. B. Bilodeau, N. Jaques, P. W. Koh, and B. Kim, "Impossibility Theorems for Feature Attribution," *Proceedings of the National Academy of Sciences of the United States of America* 121 (2024): e2304406120, https://doi.org/10.1073/pnas.2304406120.

5. A. I. Adler and A. Painsky, "Feature Importance in Gradient Boosting Trees With Cross-Validation Feature Selection," *Entropy* 24, no. 5 (2022): 687, https://doi.org/10.3390/e24050687.

6. J. Ugirumurera, E. A. Bensen, J. Severino, and J. Sanyal, "Addressing Bias in Bagging and Boosting Regression Models," *Scientific Reports* 14 (2024): 18452, https://doi.org/10.1038/s41598-024-68907-5.

7. N. Jiang, Y. Wu, and C. Li, "Limitations of Using COX Proportional Hazards Model in Cardiovascular Research," *Cardiovascular Diabetology* 23, no. 1 (2024): 219, https://doi.org/10.1186/s12933-024-02302-2.

8. J. Xue, Y. Chen, C. Xue, et al., "Limitations of Applying the COX Proportional Hazards Model to Glioma Studies," *Journal of Translational Medicine* 22, no. 1 (2024): 1156, https://doi.org/10.1186/s12967-024-05942-w.

9. X. Huang and J. Marques-Silva, "On the Failings of Shapley Values for Explainability," *International Journal of Approximate Reasoning* 171 (2024): 109112, https://doi.org/10.1016/j.ijar.2023.109112.

10. S. Oka and Y. Takefuji, "Revisiting AI Interpretability in Precision Oncology: Why Predictive Accuracy Does Not Ensure Stable Feature Importance," *Cancers* 18 (2026): 593, https://doi.org/10.3390/cancers18040593.

11. J. Zhang, X. Wu, S. C. H. Hoi, and J. Zhu, "Feature Agglomeration Networks for Single Stage Face Detection," *Neurocomputing* 380 (2020): 180–189, https://doi.org/10.1016/j.neucom.2019.10.087.

12. Y. Xie, Z. Jing, H. Pan, X. Xu, and Q. Fang, "Redefining the High Variable Genes by Optimized LOESS Regression With Positive Ratio," *BMC Bioinformatics* 26 (2025): 104, https://doi.org/10.1186/s12859-025-06112-5.

13. K. Okoye and S. Hosseini, "Correlation Tests in R: Pearson Cor, Kendall's Tau, and Spearman's Rho," in *R Programming: Statistical Data Analysis in Research*, ed. K. Okoye and S. Hosseini (Springer Nature, 2024), 247–277, https://doi.org/10.1007/978-981-97-3385-9_12.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.

14631326, 0, Downloaded from https://dom-pubs.pericles-prod.literatumonline.com/doi/10.1111/dom.71140 by yoshiyasu takefuji - Keio University , Wiley Online Library on [21/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License