Computational Methods in Applied Sciences

Tero Tuovinen
Pekka Neittaanmäki
Dietrich Knoerzer *Editors*

Challenges in Design Methods, Numerical Tools and Technologies for Sustainable Aviation, Transport and Industry

Commemorative Publication Dedicated to the 80th Jubilee of Prof. Jacques Periaux

ECCOMAS
European Community
on Computational Methods
in Applied Sciences

Springer

Chapter 16
Prevalence of AI Misapplications and Their Implications for Our Future

Yoshiyasu Takefuji

Abstract The emergence of generative AI has brought forth significant concerns regarding AI misapplications on a global scale. This paper presents an explainable AI framework, leveraging SHAP (SHapley Additive exPlanations), to illuminate potential societal impacts and risks associated with these technologies. We emphasize the critical importance of understanding the tools and methodologies that researchers worldwide have implemented and distributed. Furthermore, we propose a paradigm shift in research practices, advocating for responsible and ethical AI applications that benefit society at large. Despite the urgency of addressing AI misapplications, prominent scientific publications, including Nature, Science, and Cell, have not adequately addressed these concerns, suggesting a potential gap in editorial awareness regarding the broader implications of AI technologies. To bridge this gap in academic discourse, we recommend robust statistical approaches for evaluating AI systems, including nonlinear and nonparametric pairwise methods such as Spearman's correlation with p-values, Kendalls tau with p-values, Goodman–Kruskal gamma with p-values, Somers' D with p-values, and Hoeffding's D with p-values, while Mutual Information (MI) analysis examines complex multivariate interactions. This comprehensive approach aims to establish more rigorous standards for AI research and promote responsible innovation in the field.

Keywords Feature importance · Random forests · SHAP values · Biases · Statistical methods · Machine learning

16.1 Introduction

Many editors and editorial teams often underestimate the significant impact of biases introduced by AI algorithms, leading to flawed conclusions when fundamental AI principles are misunderstood. These biases are frequently introduced inadvertently

Y. Takefuji (✉)
Faculty of Data Science, Musashino University, Tokyo, Japan
e-mail: takefuji@keio.jp

T. Tuovinen et al. (eds.), *Challenges in Design Methods, Numerical Tools and Technologies for Sustainable Aviation, Transport and Industry*, Computational Methods in Applied Sciences 17, https://doi.org/10.1007/978-3-031-98675-8_16

by users who do not fully grasp the assumptions underlying AI tools, and any violation of these assumptions can significantly distort outcomes. Although experts in their respective fields, researchers, editors, and reviewers often employ AI technologies—such as machine learning predictions, feature importance analyses, and feature reduction—without a full appreciation of their implications.

This paper seeks to clarify key AI concepts to enhance research integrity, focusing on the significance of feature importance and the critical role of cross-validation. It is crucial to recognize that target prediction accuracy and feature importance accuracy are distinct; high target prediction accuracy does not necessarily equate to reliable feature importances. In fact, while cross-validation is valuable for evaluating prediction performance, it does not inherently improve the accuracy of feature importance assessments.

Researchers must grasp the critical role of ground truth values in AI applications. In supervised machine learning, ground truth data serves to validate target prediction accuracy; however, feature importance metrics lack such definitive reference points. As a result, different models employ various methods for computing feature importances, leading to discrepancies among their outputs. This inherent variability introduces significant biases in machine learning models—even those with high target prediction accuracy—which must be carefully considered when interpreting results. Over 100 peer-reviewed articles documented non-negotiable biases in feature importances from machine learning models.

The authors made substantial contributions to the study across multiple dimensions: conceptualizing the research framework, designing the methodology, analyzing data, writing and editing the manuscript, overseeing the research process, securing funding, and ensuring the accuracy and integrity of the findings. However, in several high-profile studies published in journals such as Nature, Science, and Cell, some supervisors lack a clear understanding of the limitations inherent in explainable AI tools like SHapley Additive exPlanations (SHAP). Additionally, many researchers do not adequately distinguish among machine learning predictions, feature importances, and cross-validation techniques.

The command 'explain = SHAP(model)' indicates that SHAP not only reflects but can also amplify existing biases in feature importance calculations. This issue is further compounded by the segmented roles within research teams, which can lead to a disconnect in sharing fundamental AI principles. Such disconnects may result in misleading outcomes and evaluation errors that have significant, adverse societal impacts.

This paper highlights that over 80% of journals reject articles that either emphasize bias or present conclusions influenced by it, while fewer than 20% of submissions are accepted. This trend can be attributed to the fact that many editors may not recognize bias as a critical issue.

28 peer-reviewed articles were published in prestigious journals across diverse fields [3, 23, 25, 34, 48, 35, 49, 36, 50, 51, 37, 52, 38, 39, 53, 54, 40, 41, 42, 55, 43, 56, 57, 44, 45, 46, 58] to support this paper.

16.2 Fundamental Theoretical Principles of Machine Learning

This paper delineates essential principles of artificial intelligence crucial for robust research methodology. While machine learning algorithms excel at predicting target outcomes with high accuracy, they often generate potentially biased feature importances, selections, and reductions. This bias stems from the fundamental challenge that feature importance calculations lack ground truth validation, unlike target predictions whose accuracy can be directly verified. Consequently, high predictive accuracy does not necessarily translate to reliable feature importance measurements. Statistical analysis addresses this challenge through three critical components: understanding data distribution, evaluating statistical relationships between variables, and validating statistical significance through p-values. Traditional robust methods, including Spearman's correlation, Kendall's tau, Goodman–Kruskal gamma, Somers' D, and Hoeffding's D, incorporate these components and provide statistical validation through p-values. However, these methods are limited to pairwise analysis, making them insufficient for understanding complex multivariate interactions. To address this limitation, Mutual Information (MI) analysis emerges as a powerful tool capable of capturing and quantifying intricate relationships among multiple variables simultaneously, thereby offering a more comprehensive understanding of feature interactions in AI systems.

In addition, cross-validation is a powerful tool that helps verify the accuracy of predictions. Unfortunately, it does not extend to evaluating the reliability of feature importances derived from machine learning models. This limitation can lead to misleading conclusions if researchers assume that high prediction accuracy automatically translates to trustworthy feature importances.

The integrity of research outcomes can be severely compromised when methodological assumptions are violated in AI applications. Specifically, the misapplication of linear methods to nonlinear data or parametric approaches to nonparametric distributions can lead to substantial distortions in results. These methodological misalignments, combined with inherent algorithmic biases, underscore a crucial point: biases and errors in AI systems are fundamentally human-induced phenomena, stemming from decisions made during model design, implementation, and interpretation. The root cause of these issues often lies in researchers' insufficient understanding of fundamental machine learning principles, leading to the introduction of significant biases and errors that can propagate through the entire analytical process. This emphasizes the critical need for robust theoretical grounding and appropriate methodological choices in AI research to ensure reliable and valid scientific conclusions.

Based on these AI principles, this paper presents examples from esteemed journals to illustrate how such biases and methodological errors can impact research findings. As diligent researchers, it is essential to pursue unbiased algorithms and robust methodologies. Relying on potentially biased calculations can lead to incorrect conclusions, adversely affecting the validity of data analyses and, ultimately, the broader implications of our work. Be aware that employing linear or parametric methods for analyzing unknown or biological data should be avoided, as such data typically exhibit nonlinear and nonparametric characteristics.

16.3 Case Studies of Biased Calculations

These biases can lead to outcomes that deviate significantly from true results. Importantly, this issue is not confined to isolated studies [21], it is a widespread concern affecting 34 articles published in Nature and other esteemed journals such as Cell, Nature Methods and others. Each of these articles addresses the bias issues stemming from feature importances in machine learning, which can lead to incorrect conclusions. Therefore, it is crucial to recognize that feature importances are model-specific and do not necessarily reflect true associations between the target variable and individual features. This highlights the necessity for careful consideration and mitigation of biases to ensure the reliability and validity of machine learning results. If left unaddressed, more researchers may rely on flawed feature importances, drawing inaccurate conclusions.

Schade et al. developed a classifier to predict sensitivity in triple-negative breast cancer, presenting a promising therapeutic approach for this aggressive tumor type [27]. Their classifier utilizes feature selection through Random Forest (RF) and Support Vector Machine (SVM) algorithms. However, feature selection with RF and SVM can introduce inherent biases due to their model-specific characteristics. Random Forest may prioritize features with greater variability or those characterized by complex interactions, potentially overlooking relevant but less prominent features. In contrast, SVM focuses on maximizing the margin between classes, which can skew feature selection toward those features that most significantly influence the decision boundary, thereby missing other important associations. Such biases can result in the misidentification of key features, compromising the accuracy and reliability of the classifier. The term “model-specific” indicates that different machine learning models yield distinct sets of selected features. Schade et al. illustrate this model-specific nature by demonstrating how various algorithms prioritize different features, even in the presence of true associations.

In another study, Shenhav et al. trained a gradient-boosted decision tree model to differentiate between children diagnosed with asthma at age three and healthy controls using data from nasal and gut microbiomes [29]. They employed tenfold cross-validation for training and testing and combined microbial trajectories with components from human milk, evaluating prediction accuracy using area under the receiver operating characteristic (auROC). Feature importance was assessed by ranking microbial taxa and human milk components based on their contributions to model performance. This study raises a critical alarm regarding the use of feature importance in machine learning applications due to inherent biases that could lead to incorrect conclusions. It emphasizes that models such as gradient-boosted decision trees (GBDT), when used to compute feature importances, may not accurately reflect true associations between the target variable and features, as these importances are model-specific and can mislead interpretations.

Almet et al. examined the top upstream transcription factors ranked by their regulatory effects on outflowing Sonic Hedgehog (SHH), utilizing random forest feature (Gini) importance to measure these effects [2]. The feature importances were calculated over 10 runs, with data presented as the mean ± standard deviation. The methods employed in this study can lead to incorrect conclusions.

Interpretable machine learning (IML) is essential for extracting meaningful insights from complex biological data. Chen et al. [8] identified three key pitfalls in the application of IML within biological contexts and offered strategies to mitigate these challenges. First, relying solely on a single IML method can yield biased assessments of feature importance. While cross-validation is effective for evaluating prediction accuracy, using multiple IML methods for cross-validation does not necessarily clarify target-feature relationships and may lead to misleading conclusions. Secondly, many IML techniques inadequately capture genuine associations between targets and features, often resulting in an overemphasis on irrelevant variables.

Even within the realm of safety science, the critical issue of bias has been overlooked by the editor. Lian et al.'s Data-Knowledge Hybrid-Driven Ensemble Feature Selection (DKH-EFS) method, while promising, relies on feature importances derived from Random Forest (RF) and Support Vector Utility Vector (SUV) algorithms, both of which are inherently biased. The RF algorithm tends to favor high-cardinality and high-variance features, while the SUV algorithm is sensitive to feature scaling and multicollinearity [19]. These inherent biases can result in the selection of misleading features for health assessments, ultimately jeopardizing the accuracy and reliability of the DKH-EFS method.

Dong et al. developed a sophisticated model to identify inefficient public open spaces (POSs) [10]. They employed a random forest approach, which constructs multiple decision trees through random sampling and the selection of random features. This powerful technique enables the model to effectively fit training samples and evaluate the significance of input features, ultimately generating predictions based on a voting or averaging mechanism. To quantify the importance of various indicators, they utilized the SHAP method. In this framework, a higher absolute SHAP value signifies greater relevance of the feature in the model's predictive process.

However, it is imperative to recognize that many researchers, including Dong et al., may overlook the critical distinction that machine learning feature importances do not necessarily represent true associations between the target variable and the features, due to the model-specific nature of these computations. This model-specificity implies that different algorithms yield varying feature importance scores, suggesting that the computed importances are not genuine associations but rather distortions influenced by the models' characteristics. In other words, random forest consistently and inherently suffers from biased feature importances [7, 22, 24, 33, 59].

This paper argues for a clear distinction between using machine learning for accurate prediction and employing it for feature importance calculations. While the use of machine learning for predicting target variables is valuable, relying on it for determining feature importance consistently leads to inherent biases and erroneous conclusions.

Additionally, although SHAP is widely embraced by researchers, its reliance on machine learning models makes it susceptible to the same biases inherent in those models. Consequently, when models produce skewed feature importances, SHAP inadvertently inherits these biases, leading to potentially misleading interpretations [5, 32]. Therefore, it is crucial for researchers to exercise caution when interpreting feature importances derived from machine learning techniques and to differentiate clearly between predictive and explanatory uses of these models.

Random forest is widely recognized for producing feature importances that are inherently biased. In fact, over 100 peer-reviewed articles have explored this issue in detail [7, 22, 24, 33, 59]. To address the shortcomings associated with biased feature importances, researchers should consider employing robust statistical methods [1, 14, 18, 30, 60], such as Spearman's correlation with p-values and Kendall's tau with p-values. These approaches provide a more reliable assessment of feature significance, independent of machine learning models and free from the biases that can distort interpretations. This paper advocates for employing robust statistical methods and discourages the use of feature importances from machine learning models. This paper addresses why random forest generates feature importance biases.

Understanding the biases associated with random forests and SHAP requires a closer look at their algorithmic structures and underlying mechanisms. Both methodologies have distinctive algorithms that contribute to the emergence of biased feature importances and interpretations.

Random forests are an ensemble learning method that constructs multiple decision trees during training. Each tree is trained on a random subset of the data, using the bootstrap aggregation method, which introduces randomness but also potential biases based on how samples are selected. Since each sample may favor particular instances, especially if the dataset has some inherent imbalance or noise, this randomness can be a source of bias in the feature importances reported by the model.

When building decision trees, random forests randomly select a subset of features to determine the best split at each node. This feature selection process can disproportionately favor certain features that yield better splits by chance. If a feature consistently appears in multiple trees due to this randomness, it may be assigned an inflated importance score, even if its contribution to prediction is not fundamentally strong. This phenomenon is compounded when features are correlated, as trees can erroneously attribute importance to one correlated feature over another based on random occurrences in the dataset.

To measure the "goodness" of a split within decisions, random forests we often use metrics like Gini impurity or entropy. These metrics evaluate how well a feature separates the classes, and if a feature provides significant splits in the majority class, it can lead to the model placing excessive importance on that feature. This situation creates a bias in the interpretation of feature relevance. Moreover, in imbalanced datasets, features that are related to the majority class can dominate the measure, while features relevant to the minority group may be overlooked altogether.

While it is true that random forests are generally less prone to overfitting compared to individual decision trees, they can still overfit to noise in the training data. Such overfitting skews feature importances, especially if the noise correlates with the target variable, thereby distorting the model's assessment of which features are truly important.

On the other hand, SHAP values are derived from cooperative game theory, treating features as players contributing to a prediction. The calculation of SHAP values is based on considering all possible combinations of the features, which can be computationally intensive. However, this approach does not inherently correct for biases that may exist in the underlying model. SHAP values are contingent on the predictions of the model they explain. If the model exhibits bias, such as inflated feature importances due to correlation or selection bias, the resulting SHAP values will reflect and potentially amplify these biases.

Moreover, SHAP provides a local approximation of feature contributions based on the area around a specific prediction. The method employs a linear model to approximate how each feature influences the prediction for that instance. However, if the underlying model is biased or exhibits complex interactions not captured in the model training, the SHAP values may misrepresent the true importance of those features in broader contexts, leading to skewed interpretations.

Finally, SHAP assumes that features contribute independently to the prediction outcome. In reality, interactions between features can complicate this assumption, which can lead to biases in the assessment of feature contributions. If a feature interacts with others in ways not taken into account during model training, the corresponding SHAP values can misrepresent its importance, further perpetuating any underlying biases.

In summary, from an algorithmic perspective, the biases in random forests arise from their ensemble approach, feature selection randomness, and the impact of noisy data on decision-making. In contrast, SHAP amplifies these biases through its dependence on model-specific outputs and assumptions regarding feature contributions. Recognizing these algorithmic intricacies is crucial for researchers aiming to accurately interpret feature importances and make informed decisions based on model outputs. By understanding the limitations of both random forests and SHAP, practitioners can navigate potential pitfalls in their analyses and consider complementing these methods with more robust statistical techniques to obtain clearer insights.

Three articles published in 2023 and 2024 in the International Journal of Information Management [9, 15, 16] highlight pitfalls associated with feature importances. These studies should reevaluate the associations between the target and features using Spearman's correlation with p-values, rather than relying on feature importances in machine learning. This week, two articles in the Nature Portfolio have been identified as having these pitfalls [61, 28].

Instead of investigating individual articles [9, 15, 16, 61, 28], this paper addresses the broader issue of why feature importances in machine learning are biased and do not represent true associations between the target and features. Feature importances are byproducts of machine learning models, where the primary goal is to improve prediction accuracy between the target and features, rather than to calculate associations between the target and individual features.

In other words, the problem lies in the fact that feature importances derived from machine learning models do not necessarily represent true associations between the target and features. These feature importances are model-specific and can induce biases. The following summarized reasons substantiate this claim:

1. Model-Specificity: Different algorithms may assign varying importance scores to the same features, leading to discrepancies in the perceived importance of variables [6, 31].
2. Correlation versus Causation: Feature importance measures the correlation between a feature and the target variable, but it doesn't establish a causal relationship [11, 26].
3. Feature Interactions: Machine learning models can capture complex interactions between features, which might not be fully accounted for in feature importance methods [8, 13].
4. Overfitting: Overfitting can inflate feature importance scores, leading to features being deemed more important than they actually are [4, 17].
5. Data Distribution: The distribution of data can influence feature importance, with features that are more informative for distinguishing between classes in a particular dataset having higher importance scores [12, 20].

Given the current prevalence of explainable AI tools like SHAP, it is prudent to temporarily rely on robust statistical methods, such as Spearman's correlation and Chi-squared tests, to assess associations between targets and features. These traditional statistical approaches come with well-established metrics, including p-values, which help in determining the significance of relationships without the inherent biases that can arise from certain machine learning algorithms.

In addition to these methods, it is essential to seek out algorithms that minimize bias, ensuring that our analyses are grounded in reliable and objective data interpretations. By prioritizing approaches that are free from bias, we can cultivate a more holistic and accurate understanding of complex systems. This shift not only enhances the integrity of our findings but also builds greater trust among stakeholders who rely on our results for decision-making.

Ultimately, while explainable AI tools are invaluable for interpreting complex models, the foundational robustness provided by established statistical methods will safeguard against potential misinterpretations. As we navigate the evolving landscape of data analysis, it will be crucial to balance modern techniques with traditional methodologies to foster comprehensive and unbiased computing practices.

16.4 Discussion

The presence of biases in machine learning studies is a pressing concern that significantly impacts the reliability of research outcomes. These biases arise from a variety of sources, including the model-specific nature of feature importance calculations and the misunderstandings surrounding fundamental AI principles. Many editors and editorial teams often underestimate the significant impact of biases introduced by AI algorithms, leading to flawed conclusions when these principles are misunderstood. These biases are frequently introduced inadvertently by researchers who do not fully grasp the assumptions underlying AI tools, and any violation of these assumptions can distort outcomes significantly.

A critical factor contributing to these biases is the failure to distinguish between target prediction accuracy and feature importance accuracy. While machine learning models may achieve high accuracy in predicting outcomes, this does not guarantee that the feature importances derived from these models are valid. For instance, models like Random Forest and Support Vector Machines may over-prioritize features based on the algorithms' inherent characteristics, resulting in inflated importance scores for certain features. Additionally, the training data can introduce biases if it contains imbalances or noise, further distorting the feature importance outputs. Thus, researchers must understand that failures to maintain methodological assumptions—such as applying linear methods to nonlinear data—can exacerbate these issues.

Moreover, while cross-validation is a valuable technique for assessing prediction performance, it does not inherently enhance the accuracy of feature importance analyses. This misconception can lead researchers to draw erroneous conclusions, believing that an accurate model necessarily produces reliable feature importances. As a result, discrepancies among feature importance outputs from various models become significant, further complicating interpretations and increasing the likelihood of flawed conclusions.

The issue is further compounded by the segmented roles within research teams, which can create a disconnect in sharing fundamental AI principles. Supervisors and researchers alike, despite being experts in their respective fields, may lack a clear understanding of the limitations inherent in explainable AI tools like SHapley Additive exPlanations (SHAP). Such disconnects can lead to misleading outcomes and evaluation errors that may have significant adverse societal impacts.

This paper seeks to clarify key AI concepts to enhance research integrity by emphasizing the importance of ground truth values and the critical role of feature importance and cross-validation. In supervised machine learning, ground truth data validates target prediction accuracy; however, feature importance metrics lack definitive reference points, making their reliability questionable. Over 100 peer-reviewed articles have documented non-negotiable biases present in feature importances derived from machine learning models, highlighting the urgency of addressing these concerns.

It is notable that over 80% of journals reject articles that emphasize bias or present conclusions influenced by it, while fewer than 20% of submissions are accepted. This trend suggests that many editors may not recognize bias as a critical issue, which can perpetuate flawed analyses and misinterpretations across published research.

Furthermore, this paper outlines several fundamental principles of artificial intelligence essential for sound research practices. Relying on biased calculations can lead to incorrect conclusions, negatively affecting the validity of data analyses and the broader implications of research findings. For example, employing linear or parametric methods to analyze biological data—which typically exhibits nonlinear and nonparametric characteristics—can introduce significant distortion in results.

In conclusion, understanding and addressing the biases inherent in machine learning methodologies is vital for preserving research integrity. Through careful engagement with these principles, researchers can improve the accuracy of their findings and avoid common pitfalls related to feature importance and predictive modeling. By prioritizing robust methodologies and a comprehensive understanding of AI tools, the research community can work toward producing reliable, impactful outcomes that genuinely reflect the underlying data associations.

References

1. Aguilar-Elena R, Agún-González JJ (2024) Chi-square automatic interaction detection (CHAID) analysis of the use of safety goggles and face masks as personal protective equipment (PPE) to protect against occupational biohazards. J Biosaf Biosecurity 6(2):125–133. https://doi.org/10.1016/j.jobb.2024.05.001
2. Almet AA, Tsai YC, Watanabe M et al (2024) Inferring pattern-driving intercellular flows from single-cell and spatial transcriptomics. Nature Method. https://doi.org/10.1038/s41592-024-02380-w
3. Arif M, Takefuji Y (2025) Why AI image generators cannot afford to be blind to racial bias. AI & Soc. https://doi.org/10.1007/s00146-025-02258-1
4. Babyak MA (2004) What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 66(3):411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9
5. Bilodeau B, Jaques N, Koh PW, Kim B (2024) Impossibility theorems for feature attribution. Proc Natl Acad Sci USA 121(2):e2304406120. https://doi.org/10.1073/pnas.2304406120
6. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
7. Chen J, Ooi LQR, Tan TWK, Zhang S, Li J, Asplund CL, Eickhoff SB, Bzdok D, Holmes AJ, Yeo BTT (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
8. Chen V, Yang M, Cui W et al (2024) Applying interpretable machine learning in computational biology: Pitfalls, recommendations, and opportunities for new developments. Nat Methods 21(12):1454–1461. https://doi.org/10.1038/s41592-024-02359-7
9. Căpățină A, Patel NJ, Mitrov K, Cristea DS, Micu A, Micu A-E (2024) Elevating students' lives through immersive learning experiences in a safe metaverse. Int J Information Management, 75:102723.https://doi.org/10.1016/j.ijinfomgt.2023.102723

10. Dong X, Zhang X, Jing Y, Zhou Q, Bai L, Du S (2024) Does every public open space (POS) contribute to sustainable city development? An assessment of inefficient POS in Beijing. Sustain Cities Soc: 105980. https://doi.org/10.1016/j.scs.2024.105980
11. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451
12. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182. https://doi.org/10.1162/153244303322753616
13. Hawkins DM (2004) The problem of overfitting. J Chem Inf Comput Sci 44(1):1–12. https://doi.org/10.1021/ci0342472
14. 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(B):123633. https://doi.org/10.1016/j.eswa.2024.123633
15. Joung J, Kim H (2023) Interpretable machine learning-based approach for customer segmentation for newproduct development from online product reviews. Int J Information Management, 70:102641.https://doi.org/10.1016/j.ijinfomgt.2023.102641
16. King KK, Wang B (2023) Diffusion of real versus misinformation during a crisis event: a big data-driven approach. Int J Information Management, 71:102390.https://doi.org/10.1016/j.ijinfomgt.2021.102390
17. Kuhn M, Johnson K (2019) Feature engineering and selection: A practical approach for predictive models. CRC Press. https://doi.org/10.1201/9781315108230
18. Li X, Ma Y, Zhou Q, Zhang X (2024) Sparse large-scale high-order fuzzy cognitive maps guided by Spearman correlation coefficient. Appl Soft Comput 167(A):112253. https://doi.org/10.1016/j.asoc.2024.112253
19. Lian Z, Zhou ZJ, Hu CH, Wang J, Zhang CC, Zhang CL (2024) A health assessment method with attribute importance modeling for complex systems using belief rule base. Reliab Eng Syst Saf 251:110387. https://doi.org/10.1016/j.ress.2024.110387
20. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774. https://doi.org/10.5555/3295222.3295230
21. Lynch CJ, Elbau IG, Ng T et al (2024) Frontostriatal salience network expansion in individuals in depression. Nature. https://doi.org/10.1038/s41586-024-07805-2
22. Nguyen TT, Huang JZ, Nguyen TT (2015) Unbiased feature selection in learning random forests for high-dimensional data. Sci World J 2015:471371. https://doi.org/10.1155/2015/471371
23. Nguyen MH, Takefuji Y (2025) Reassessing predictive modeling for emergency department return in COVID-19 patients. Am J Emerg Med 43. https://doi.org/10.1016/j.ajem.2025.01.009
24. Oh S (2022) Predictive case-based feature importance and interaction. Inf Sci 593:155–176. https://doi.org/10.1016/j.ins.2022.02.003
25. Pan H, Takefuji Y (2025) Enhancing feature importance analysis with Spearman's correlation with p-values: Recommendations for improving PHLF prediction. Eur J Surg Oncol 51(7):109687. https://doi.org/10.1016/j.ejso.2025.109687
26. Pearl J (2009) Causality: Models, reasoning, and inference. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161
27. Schade AE, Perurena N, Yang Y et al (2024) AKT and EZH2 inhibitors kill TNBCs by hijacking mechanisms of involution. Nature. https://doi.org/10.1038/s41586-024-08031-6
28. Shen M, Chen M, Chen Y, et al (2024) Mitophagy related diagnostic biomarkers for coronary in-stent restenosis identified using machine learning and bioinformatics. Scientific Reports, 14:24137. https://doi.org/10.1038/s41598-024-74862-y
29. Shenhav L, Fehr K, Reyna ME et al (2024) Microbial colonization programs are structured by breastfeeding and guide healthy respiratory development. Cell 187(19):5431–5452.e20. https://doi.org/10.1016/j.cell.2024.07.022
30. Shi X, Xu M, Du J (2023) Max-sum test based on Spearman's footrule for high-dimensional independence tests. Comput Stat Data Anal 185:107768. https://doi.org/10.1016/j.csda.2023.107768
31. Shmueli G (2010) To explain or to predict? Stat Sci 25(3):289–310. https://doi.org/10.1214/10-STS330

32. Singhal A, Neveditsin N, Tanveer H, Mago V (2024) Toward fairness, accountability, transparency, and ethics in AI for social media and health care: Scoping review. JMIR Med Inform 12:e50048. https://doi.org/10.2196/50048
33. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8(1):25. https://doi.org/10.1186/1471-2105-8-25
34. Takefuji Y (2024a) Addressing bias in feature importances derived from XGBoost. Br J Anaesth 133(3):351–359. https://doi.org/10.1016/j.bja.2024.11.033
35. Takefuji Y (2024c) Evaluating feature importance biases in logistic regression: Recommendations for robust statistical methods. Eur J Intern Med. https://doi.org/10.1016/j.ejim.2024.11.022
36. Takefuji Y (2024e) Mitigating feature importance bias in regression models for clinical decision-making. Am J Obstet Gynecol. https://doi.org/10.1016/j.ajog.2024.12.010
37. Takefuji Y (2024h) Reply to the editor. Clin Nutr. https://doi.org/10.1016/j.clnu.2024.11.031
38. Takefuji Y (2024j) Unveiling hidden biases in machine learning feature importance. J Energy Chem 102:49–51. https://doi.org/10.1016/j.jechem.2024.10.032
39. Takefuji Y (2025a) Addressing bias in feature importance: A hybrid approach for risk prediction in prognostic survival models. JCO Precis Oncol. https://doi.org/10.1200/PO-24-00785
40. Takefuji Y (2025d) Critical evaluation of feature importance assessment in FFNN-based models for predicting Kamlet-Taft parameters. Green Chem Eng. https://doi.org/10.1016/j.gce.2025.01.003
41. Takefuji Y (2025e) Model-specific feature importances: Distinguishing true associations from target-feature relationships. J Affect Disord 369:390–391. https://doi.org/10.1016/j.jad.2024.10.019
42. Takefuji Y (2025f) Reevaluating feature importances in machine learning models for schizophrenia and bipolar disorder: The need for true associations. Brain Behav Immun 124:123–124. https://doi.org/10.1016/j.bbi.2024.11.036
43. Takefuji Y (2025h) Reevaluating feature importance in machine learning models for CO2 photoreduction: A statistical perspective. Appl Catal B 368:125145. https://doi.org/10.1016/j.apcatb.2025.125145
44. Takefuji Y (2025k) Critical evaluation of feature importance assessment in proteomic analysis using skin microdialysis. J Allergy Clin Immunol. https://doi.org/10.1016/j.jaci.2024.12.1096
45. Takefuji Y (2025l) Challenges in feature importance interpretation: Analyzing LSTM–NN predictions in battery material flotation. J Ind Inf Integr 45:100809. https://doi.org/10.1016/j.jii.2025.100809
46. Takefuji Y (2025m) Enhancing machine learning in gas–solid interaction analysis: Addressing feature selection and dimensionality challenges. Coord Chem Rev 534:216583. https://doi.org/10.1016/j.ccr.2025.216583
47. Takefuji Y (2025o) Reevaluating feature importance in machine learning: Concerns regarding SHAP interpretations in the context of the EU Artificial Intelligence Act. Water Res 280:123514. https://doi.org/10.1016/j.watres.2025.123514
48. Takefuji Y (2024b) Chi-Squared and P-values vs. machine learning feature selection. Ann Oncol. https://doi.org/10.1016/j.annonc.2024.10.013
49. Takefuji Y (2024d) Mitigating biases in feature selection and importance assessments in predictive models using LASSO regression. Oral Oncol 159, Article 107090. https://doi.org/10.1016/j.oraloncology.2024.107090
50. Takefuji Y (2024f) Reassessing feature importance biases in machine learning models for infection analysis. J Infect 89(6), Article 106357. https://doi.org/10.1016/j.jinf.2024.106357
51. Takefuji Y (2024g) Reevaluating statistical methods in metabolomic studies: A case for Spearman's correlation. Molecular Plant 17(1). https://doi.org/10.1016/j.molp.2024.12.014
52. Takefuji Y (2024i) Unveiling feature importance biases in linear regression: Implications for protein-centric cardiovascular research. Atherosclerosis, Article 119049. https://doi.org/10.1016/j.atherosclerosis.2024.119049

53. Takefuji Y (2025b) Addressing feature importance biases in machine learning models for early diagnosis of type 1 Gaucher disease. J Clin Epidemiol 178, Article 111619. https://doi.org/10.1016/j.jclinepi.2024.111619
54. Takefuji Y (2025c) Beyond XGBoost and SHAP: Unveiling true feature importance. J Hazard Mater 488, Article 137382. https://doi.org/10.1016/j.jhazmat.2025.137382
55. Takefuji Y (2025g) Reevaluating feature selection in phase field models for battery performance: A call for robust statistical approaches. Energy Storage Mater 75, Article 104060. https://doi.org/10.1016/j.ensm.2025.104060
56. Takefuji Y (2025i) Reevaluating feature importance in machine learning for food authentication: Addressing bias and enhancing methodological rigor. Trends Food Sci & Technol, Article 104853. https://doi.org/10.1016/j.tifs.2024.104853
57. Takefuji Y (2025j) Beyond principal component analysis: Enhancing feature reduction in electronic noses through robust statistical methods. Trends Food Sci & Technol, 104919. https://doi.org/10.1016/j.tifs.2025.104919
58. Takefuji Y (2025n) Reevaluating feature importance in gas–solid interaction predictions: A call for robust statistical methods. Coord Chem Rev 534, Article 216584. https://doi.org/10.1016/j.ccr.2025.216584
59. Thakur D, Biswas S (2024) Permutation importance based modified guided regularized random forest in human activity recognition with smartphone. Eng Appl Artif Intell 129:107681. https://doi.org/10.1016/j.engappai.2023.107681
60. Vierra A, Razzaq A, Andreadis A (2023) Chapter 28—Categorical variable analyses: Chi-square, fisher exact, and mantel-haenszel. In: Eltorai AEM, Bakal JA, Newell PC, Osband AJ (eds) Handbook for designing and conducting clinical and translational research, pp 171–175. Academic Press. https://doi.org/10.1016/B978-0-323-90300-4.00095-1
61. Wu Y, Shi Z, Zhou X, et al (2024) scHiCyclePred: a deep learning framework for predicting cell cycle phases from single-cell Hi-C data using multi-scale interaction information. Communications Biology, 7:923.https://doi.org/10.1038/s42003-024-06626-3