Water Research 280 (2025) 123514

Contents lists available at ScienceDirect

Water Research

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

Reevaluating feature importance in machine learning: concerns regarding SHAP interpretations in the context of the EU artificial intelligence act

Yoshiyasu Takefuji [*]

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

ARTICLE INFO

Keywords
Machine learning
SHAP
Feature importance
EU Artificial intelligence act
Statistical methods
Bias mitigation

ABSTRACT

This paper critically examines the analysis conducted by Maußner et al. on AI analysis, particularly their interpretation of feature importances derived from various machine learning models using SHAP (SHapley Additive exPlanations). Although SHAP aids in interpretability, it is subject to model-specific biases that can misrepresent relationships between variables. The paper emphasizes the lack of ground truth values in feature importance assessments and calls for careful consideration of statistical methodologies, including robust nonparametric approaches. By advocating for the use of Spearman's correlation with p-values and Kendall's tau with p-values, this work aims to strengthen the integrity of findings in machine learning studies, ensuring that conclusions drawn are reliable and actionable.

Maußner et al. conducted a legal analysis of the EU Artificial Intelligence Act and the Ethics Guidelines for Trustworthy AI (Maußner et al., 2025). They showcased the concept of explainable AI through various machine learning models, including Linear Least Squares Regression (LS), Decision Tree Regression (DT), K-Nearest Neighbors Regression (KNN), Long Short-Term Memory (LSTM), and Random Forest Regression (RF). Their findings revealed that each model with SHAP produced unique feature importances (Maußner, 2025).

This paper recognizes the thorough legal assessment of the EU Artificial Intelligence Act and the Ethics Guidelines for Trustworthy AI conducted by Maußner et al. However, it raises significant concerns about their interpretation of feature importances derived from machine learning models using SHAP (SHapley Additive exPlanations). The model-specific nature of these importance metrics can lead to misleading or erroneous conclusions.

While the primary goal of machine learning is to generate accurate predictions based on known ground truth values, the feature importances produced by these models lack corresponding ground truth references, complicating their validation. Despite Maußner et al.'s claims of "robustness" and "trust," they did not adequately address the potential distortions in feature importance assessments derived from SHAP. Their findings indicated that different models yield varying feature importances, raising concerns that conclusions drawn from these metrics may be fundamentally flawed. This paper questions why they did not recognize the possibility of erroneous conclusions in their analyses.

Moreover, although Maußner et al. are experts in water research, they may not fully grasp the complexities of algorithmic calculations and the biases that can arise from models, particularly regarding SHAP. This suggests a potential disconnect between their domain expertise and the underlying computational methodologies.

The issue of non-negligible bias in machine learning models is well-established, with over 100 peer-reviewed articles documenting substantial biases in feature importance assessments (Fisher, 2019; Giancesco, 2018; Strobl, 2007). There are several bias mitigation methods, but none can completely eliminate biases (Altmann, 2010). Although SHAP (SHapley Additive exPlanations) can be a valuable tool for interpretability, it is critically dependent on the underlying model. As a result, SHAP can inherit and even magnify biases embedded in that model due to its design (i.e., explain=SHAP(model)) (Bilodeau, 2024; Cross, 2024; Momenzadeh, 2022). This necessitates caution when interpreting feature importances and drawing conclusions solely based on SHAP outputs.

The universal challenge in feature importance estimation stems from the fact that, unlike target prediction in supervised learning—where well-defined ground truth values exist—there is no equivalent ground truth for the contributions of individual features. Instead, methods such as permutation importance, gradient-based approaches, and SHAP rely solely on the trained model's internal logic to infer feature significance. Consequently, these methods inherently introduce biases because each employs distinct assumptions and limitations that may skew the

* Corresponding author at: Professor, Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan.
E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.watres.2025.123514
Received 25 December 2024; Received in revised form 3 March 2025; Accepted 17 March 2025
Available online 18 March 2025

interpretation of a feature's role in the model's predictions (Bilodeau, 2024; Cross, 2024; Momenzadeh, 2022; Huang, 2024; Kumar, 2021; Lones, 2024; Molnar, 2022). In this context, it is also noteworthy that Maußner et al. demonstrated that different models inherently generate distinct feature importance estimates, further highlighting the variability introduced by methodological choices.

It is important to note that target prediction accuracy and feature importance accuracy are two fundamentally different issues. While high target prediction accuracy may indicate that a model performs well in its primary prediction task, it does not guarantee that the derived feature importances are reliable (Fisher, 2019; Lipton, 2018). Feature importance methods are not validated against an external benchmark for true causality; rather, they remain subject to biases arising from the model's internal structure and the peculiarities of the training data.

Evidence from >100 peer-reviewed studies indicates that significant biases in feature importance estimates are prevalent across a wide range of fields. These biases can result from several factors, including imbalances or peculiarities in the training data, multicollinearity among features, and complex interactions that are not sufficiently disentangled by conventional feature importance methods. Even when water demand prediction models are built on high-quality data—such as measurements from calibrated flow sensors—the absence of an independently validated "true" importance measure remains an inherent limitation. This observation implies that while factors like dataset size and measurement precision may influence the magnitude of bias, they do not resolve the fundamental challenge of accurately estimating feature contributions without an external standard.

To illustrate, consider a water demand forecasting model that uses sensor data from flow meters together with environmental variables, such as temperature and precipitation. If subtle measurement errors or unaccounted confounders cause the model to overemphasize the significance of temperature, then techniques like SHAP—which depend on the model's internally derived relationships—will reflect and potentially amplify this bias. Such distortions can mislead decision-making by causing an over-reliance on temperature forecasts while undervaluing other critical variables, such as occupancy or economic activity. This example, analogous to documented challenges in medical or health-related applications, underscores that the theoretical underpinnings governing biases in feature importance estimation remain consistent across diverse data domains.

This paper advocates for the adoption of bias-free, robust statistical methods, such as Spearman's correlation (Eden, 2022; Yu, 2024) and Kendall's tau (Ouachene, 2024; Wang, 2021), both of which provide p-values and are effective in handling nonlinear relationships in a nonparametric context. Maußner et al. should reevaluate their findings using these robust methods to enhance the integrity and reliability of their outcomes.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

Funding

This research has no fund.

Availability of data and material

Not applicable.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original draft, Validation, Investigation, Conceptualization.

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.

Data availability

No data was used for the research described in the article.

References

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26 (10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134.

Bilodeau, B., Jaques, N., Koh, P.W., Kim, B., 2024. Impossibility theorems for feature attribution. In: Proceedings of the National Academy of Sciences of the United States of America, 121, e2304406120. https://doi.org/10.1073/pnas.2304406120.

Cross, J.L., Choma, M.A., Onofrey, J.A., 2024. Bias in medical AI: implications for clinical decision-making. PLOS Digital Health 3 (11), e0000651. https://doi.org/10.1371/journal.pdig.0000651.

Eden, S.K., Li, C., Shepherd, B.E., 2022. Nonparametric estimation of Spearman's rank correlation with bivariate survival data. Biometrics 78 (2), 421–434. https://doi.org/10.1111/biom.13453.

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.

Gianfrancesco, M.A., Tamang, S., Yazdany, J., Schmajuk, G., 2018. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern. Med. 178 (11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763.

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

Kumar, I., Scheidegger, C., Venkatasubramanian, S., Friedler, S., 2021. Shapley residuals: quantifying the limits of the Shapley value for explanations. Adv. Neural Inf. Process Syst. 34, 26598–26608.

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

Lones, M.A., 2024. Avoiding common machine learning pitfalls. Patterns 5 (10), 101046. https://doi.org/10.1016/j.patter.2024.101046.

Maußner, C., Oberascher, M., Autengruber, A., Kahl, A., Sitzenfrei, R., 2025. Explainable artificial intelligence for reliable water demand forecasting to increase trust in predictions. Water Res. 268 (B), 122779. https://doi.org/10.1016/j.watres.2024.122779.

Molnar, C., et al., 2022. General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K. R., Samek, W. (Eds.), xxAI - Beyond Explainable AI, xxAI - Beyond Explainable AI, 13200. Springer, p. 4. https://doi.org/10.1007/978-3-031-04083-2_4.

Momenzadeh, A., Shamsa, A., Meyer, J.G., 2022. Bias or biology? Importance of model interpretation in machine learning studies from electronic health records. JAMIA Open, 5 (3). https://doi.org/10.1093/jamiaopen/ooac063. Article ooac063.

Ouachene, N., Kiessé, T.S., Corson, M.S., 2024. Using conditional Kendall's tau estimation to assess interactions among variables in dairy-cattle systems. Agric. Syst. 220. https://doi.org/10.1016/j.agsy.2024.104089. Article 104089.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform. 8, 25. https://doi.org/10.1186/1471-2105-8-25.

Wang, J.H., Chen, Y.H., 2021. Network-adjusted Kendall's tau measure for feature screening with application to high-dimensional survival genomic data. Bioinform. 37 (15), 2150–2156. https://doi.org/10.1093/bioinformatics/btab064.

Yu, H., Hutson, A.D., 2024. A robust Spearman correlation coefficient permutation test. Commun. Stat. Theory Methods 53 (6), 2141–2153. https://doi.org/10.1080/03610926.2022.2121144.