Measurement 274 (2026) 121317

Contents lists available at ScienceDirect

Measurement

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

Unmasking the false reliability: supervised feature importance methods lack consistency and dose-response validation in desalination research

Yoshiyasu Takefuji *

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

ARTICLE INFO

Keywords:
Desalination
Feature importance
SHAP
Unsupervised learning
Stability testing

ABSTRACT

The proliferation of AI-driven analysis in chemical measurement has outpaced critical methodological scrutiny, as evidenced by Measurement's publication of over 1300 articles on SHAP, feature importance, and feature selection. This surge reflects growing enthusiasm for identifying true associations, yet fundamental misunderstandings persist regarding the validity of supervised models for feature interpretation. Using a public desalination dataset (73 samples, 12 features) and our novel leave-top1-out protocol, we demonstrate that Random Forest and XGBoost with SHAP exhibit severe ranking instability: removing the top-ranked feature triggers cascading reordering, with SHAP inheriting rather than correcting this volatility. In contrast, unsupervised methods and Spearman correlation maintain consistent hierarchies, with Spearman achieving superior test performance. Our results establish that label-driven prediction errors fundamentally compromise supervised feature importance stability, while correlation-based approaches provide more reliable assessments. Critically, supervised models optimize prediction accuracy—not causal or associative truth—rendering them unsuitable for feature importance claims without rigorous validation against two essential criteria systematically overlooked in existing literature ranking consistency under perturbations and dose–response relationships. We provide a methodological framework for establishing genuine feature associations in desalination and membrane research, demonstrating that predictive performance alone cannot validate feature importance, and advocate for correlation-based methods when ground truth remains unavailable.

1. Introduction

The proliferation of artificial intelligence tools has enabled researchers to conduct extensive data-driven analyses across numerous disciplines. However, a critical gap exists in many researchers' understanding of supervised learning models' fundamental principles and underlying assumptions. This knowledge deficit, combined with the frequent lack of ground truth for validation purposes, has resulted in widespread misapplication of AI methodologies.

Measurement illustrates this theoretical and empirical concerning trend, with a sharp increase in publications focused on interpretability techniques: 98 articles on SHAP (34 in 2025 and 50 in 2026), 212 articles on feature importance (75 in 2025 and 72 in 2026), and 1027 articles on feature selection (234 in 2025 and 183 in 2026). This surge reflects the research community's growing interest in identifying true associations within complex datasets.

Despite this heightened interest, fundamental misconceptions persist regarding feature importance, selection, and interpretation within supervised learning contexts. This paper highlights urgent theoretical and empirical concerns about using supervised models—with or without explainability tools like SHAP—for feature assessment. Current studies frequently fail to validate consistency and dose–response relationships when evaluating true associations, and no existing algorithms can reliably calculate these associations with demonstrated accuracy due to insufficient understanding of supervised models.

A crucial distinction that researchers often overlook is between target prediction and true association assessment. Supervised models exhibit two different types of accuracy: prediction accuracy for target variables and accuracy in determining feature importance. While the former can be validated against ground truth labels, the latter lacks corresponding validation mechanisms. This absence of ground truth for feature importance calculations frequently leads to erroneous interpretations and misleading conclusions in research findings.

A crucial distinction that researchers often overlook is between target prediction and true association assessment. Supervised models exhibit two different types of accuracy: prediction accuracy for target

* Corresponding author.
E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.measurement.2026.121317
Received 2 February 2026; Received in revised form 20 March 2026; Accepted 26 March 2026
Available online 27 March 2026

variables and accuracy in determining feature importance. While the former can be validated against ground truth labels, the latter lacks corresponding validation mechanisms. This absence of ground truth for feature importance calculations frequently leads to erroneous interpretations and misleading conclusions in research findings.

This study critically examines the reliability of supervised models—with or without SHAP—against unsupervised models and non-predictive methods using a publicly available desalination dataset. Our primary objective is to expose the inherent instability in feature ranking orders when using supervised approaches and to demonstrate the superior stability of unsupervised methods and direct correlation analyses. Critically, researchers must validate two essential components for establishing true associations: ranking consistency across perturbations and dose–response relationships between features and outcomes. However, existing studies have systematically overlooked these validation requirements, potentially compromising the reliability of reported findings in desalination and membrane research. This paper raises urgent concerns regarding these misapplications, where proper model understanding is critical for advancing the field.

To rigorously assess true associations, a critical aspect frequently overlooked by many researchers, this paper examines two key components: consistency and dose–response relationships [1–10]. We implement a novel leave-top1-out approach to test model stability. The methodology involves: (1) selecting top n features from the full feature set (set1); (2) removing the highest-ranked feature from the full set to create a reduced dataset; (3) re-selecting top n-1 features from this reduced dataset (set2); and (4) examining the consistency of feature importance ranking orders between the two sets. Our analysis demonstrates that removing the highest feature yields a substantial impact on feature ranking orders, underscoring the importance of stability testing in feature selection processes and highlighting potential interdependencies among predictors that may otherwise remain undetected.

While the individual components of our framework (RF, XGBoost, and SHAP) are established methods, the practical contribution of this study lies in the systematic integration of two mandatory validation criteria, consistency and dose–response relationships, which have been absent from prior feature importance analyses in this domain. Existing studies typically evaluate features as unordered sets, assessing their collective or independent contributions without explicitly verifying whether identified associations hold under sequential feature removal. In contrast, the proposed framework evaluates features as ordered sets, where the rank of each feature carries analytical meaning. To operationalize this, we introduce a leave-top1-out protocol, which sequentially removes the highest-ranked feature and re-evaluates the remaining feature rankings. This procedure directly tests whether the reported associations satisfy consistency and dose–response requirements, two criteria considered necessary conditions for inferring reliable, interpretable relationships between predictors and outcomes. The value of this framework is therefore methodological and applied: it provides practitioners with a structured, reproducible validation procedure that reduces the risk of reporting spurious feature associations.

Supervised models operate with two fundamentally different types of accuracy: target prediction accuracy and feature importance accuracy. While target prediction accuracy can be rigorously validated against ground truth labels, feature importance lacks corresponding ground truth for validation – there is no definitive reference point to determine if the identified important features are truly the most influential. This fundamental absence of ground truth in feature importance determinations results in different models generating distinct feature importance rankings, highlighting their model-specific nature. It's critical to recognize that feature importance in supervised models primarily indicates contributions to the prediction mechanism rather than representing true underlying associations in the data.

Our research reveals a significant insight: supervised models frequently demonstrate instability in feature importance ranking orders due to errors propagated through the label-driven learning process. In contrast, unsupervised models exhibit notably stronger stability in feature ranking orders specifically because they operate without the potential distortions introduced by label-driven errors. This stability difference suggests unsupervised approaches may provide more consistent feature importance assessments when the ground truth importance is unknown.

This study introduces a systematic feature validation methodology for measurement-based predictive modeling in desalination systems, grounded in two metrological principles: consistency and dose–response relationships. Unlike existing studies that report feature importance without rigorous validation, the proposed framework evaluates whether selected features exhibit physically meaningful and reproducible relationships with the target variable, thereby serving as a validation procedure analogous to those used in measurement science.

Furthermore, feature importance is interpreted through a dual-role measurement framework: (1) as a quantitative indicator of each feature's contribution to predictive accuracy, assessed through 5-fold cross-validation R2 scores, and (2) as an ordered measurement set derived from a leave-top1-out procedure, which systematically evaluates the marginal contribution of each feature to model performance. Together, these roles provide a structured approach to measurement-system performance evaluation, offering traceable and interpretable estimates of variable influence under uncertainty.

2. Methods

To ensure a fair and unbiased comparison across all methods, no data preprocessing, including scaling, transformation, and normalization, was applied to any algorithm. Preprocessing operations such as scaling, transformation, and normalization alter the distributional properties of raw data, potentially introducing artifactual patterns that do not reflect true underlying relationships. Since this framework specifically aims to identify genuine associations supported by consistency and dose–response criteria, applying preprocessing could obscure or distort the natural structure of feature relationships. The conclusions of this study are therefore applicable to raw, unprocessed datasets sharing similar characteristics to those analyzed here. No hyperparameter tuning was applied to any algorithm. Default settings were used uniformly across all algorithms to ensure that observed differences in feature ranking reflect genuine algorithmic differences rather than tuning artifacts. By adopting default settings, a standardized and reproducible baseline is established, defining a clear applicable scope for the conclusions drawn.

Our analysis utilized the publicly available Data_Desalination.xlsx dataset [11], comprising 73 samples and 12 features that characterize various aspects of desalination processes. This dataset was selected for its comprehensive representation of typical desalination parameters and its accessibility to the research community.

We implemented a multi-step methodological framework to systematically evaluate feature importance stability across different analytical approaches:

We applied seven distinct analytical methods to calculate feature importance rankings from the complete dataset: (1) Supervised Models: Random Forest and XGBoost regressors were implemented to establish baseline supervised learning approaches commonly used in desalination research; (2) SHAP Explanations: SHAP (SHapley Additive exPlanations) values were calculated for both supervised models to assess whether explanation frameworks improve feature importance stability; (3) Unsupervised Models: Feature Agglomeration (FA) and Highly Variable Gene Selection (HVGS) were employed as representative unsupervised approaches; (4) Non-Target-Prediction Method: Spearman's rank correlation coefficient was calculated to provide a nonlinear nonparametric assessment of feature relationships with the target variable.

To rigorously test the stability of feature importance rankings, we employed a leave-top1-out approach: The highest-ranked feature from each method's initial analysis was identified and systematically removed

from the dataset to create a reduced dataset specific to each method. Feature importance calculations were then repeated on each reduced dataset using the corresponding method. The resulting feature rankings were compared to the original rankings (excluding the removed feature) to assess stability.

For supervised and unsupervised methods, as well as Spearman correlation, we calculated R^2 scores using 5-fold cross-validation to assess mean predictive performance and feature importance stability.

All algorithms were implemented with their default settings in scikit-learn and XGBoost frameworks without any hyperparameter tuning to ensure fair comparison across methods. The dataset was used in its original form without any preprocessing, scaling, or transformation to maintain the authentic characteristics of the data and avoid introducing any artificial patterns.

For purposes of reproducibility and transparency, our complete analytical pipeline is implemented in Python code (crossvalid.py) and publicly available on GitHub [12]. This implementation includes all feature importance calculations, stability assessments, and performance evaluations, allowing researchers to replicate our findings and extend our analysis to other datasets or methodological variations.

3. Results

Cross-validated R^2 (CV5 R^2, CV4 R^2) and feature importance ranking orders for all seven methods across the two feature set configurations (Top-5 features and Top-4 features) are presented in Table 1.

Among all methods in cross-validation with top 5 features, Spearman achieved the highest CV5 R^2 of 0.4839, closely followed by RF-SHAP (0.4815) and RF (0.4718), while XGB-SHAP recorded the lowest CV5 R^2 of 0.4007. Notably, the unsupervised methods FA and HVGS demonstrated moderate but consistent accuracy across both configurations, with FA yielding CV5 and CV4 R^2 values of 0.4235 and 0.4352, and HVGS producing nearly identical scores of 0.4704 and 0.4709, reflecting a marginal difference of only 0.0005 between the two sets. This near-identical performance across feature set sizes indicates that HVGS maintains robust predictive stability regardless of feature space reduction.

A critical distinction between supervised and unsupervised methods emerges when comparing the ordered feature rankings between the Top-5 and Top-4 configurations. The unsupervised methods FA, HVGS, and Spearman exhibited substantially greater consistency in their ranking orders across both feature sets. Specifically, HVGS retained an identical rank ordering of {Time, S/N, Temperature, Pressure} across both configurations, with only the removal of the top-ranked Conductivity feed causing a rank shift. Similarly, FA preserved a stable core ordering of {Time, S/N, Temperature} in both sets, with only the lowest-ranked features showing minor reshuffling. This stability suggests that variance-based and correlation-based unsupervised selectors capture intrinsic data structure that remains consistent under feature perturbation.

In contrast, the supervised methods RF, XGB, RF-SHAP, and XGB-SHAP demonstrated pronounced instability in their feature ranking orders between the two configurations. RF ranked Temperature first in the Top-5 set, yet upon removal of Temperature, S/N rose to the top position in the Top-4 set, indicating that the importance assigned to individual features by RF is highly sensitive to the presence of correlated variables. XGB exhibited the most severe ranking instability: its Top-5 set was led by S/N, whereas the Top-4 set promoted Time to the first position despite Time not appearing in the Top-5 ranking at all, representing a complete reordering of priorities. XGB-SHAP similarly replaced S/N with Time as the top feature and introduced Permeate recovery into a higher rank when Flow feed was retained, while RF-SHAP, despite being the most accurate method overall, still experienced a rank inversion between Temperature and S/N across the two configurations. These rank inconsistencies in supervised methods are attributable to their sensitivity to inter-feature correlations and the implicit weighting imposed by the target variable during model training, which causes importance scores to redistribute substantially when any single dominant feature is removed.

The stability of unsupervised methods in preserving ordered feature sets carries practical significance for desalination process analysis. When the goal is to identify a physically interpretable and reproducible set of process-governing variables, methods such as HVGS and FA offer a more reliable basis than supervised importance-based approaches, despite their modest accuracy compared to RF-SHAP. The supervised methods, while achieving higher peak R^2 values, do so at the cost of ranking consistency, making their feature orderings less trustworthy as standalone indicators of variable importance in the context of iterative feature reduction.

Table 1
Cross-validation r^2 and feature importance ranking orders per algorithm.

Method	CV5 R^2 mean	CV4 R^2 mean	Top-5 feature rankings	Top-4 feature rankings
RF	0.4718	0.4998	Temperature | Time | S/N | Conductivity feed | Pressure	S/N | Time | Conductivity feed | Pressure
XGB	0.4031	0.407	S/N | Flow feed | Temperature | Conductivity feed | Permeate flow rate	Time | Flow feed | Temperature | Conductivity feed
RF-SHAP	0.4815	0.4998	Temperature | Time | Conductivity feed | S/N | Pressure	S/N | Time | Conductivity feed | Pressure
XGB-SHAP	0.4007	0.4133	S/N | Temperature | Permeate recovery | Flow feed | Permeate flow rate	Time | Temperature | Permeate recovery | Flow feed
FA	0.4235	0.4352	Conductivity feed | Time | S/N | Temperature | Flow feed	Time | S/N | Temperature | Permeate recovery
HVGS	0.4704	0.4709	Conductivity feed | Time | S/N | Temperature | Pressure	Time | S/N | Temperature | Pressure
Spearman	0.4839	0.4343	Pressure | Temperature | Conductivity feed | S/N | Time	Temperature | Conductivity feed | S/N | Time

4. Discussion

Feature assessment derived from supervised models in desalination research fundamentally lacks rigorous validation due to the absence of ground truth in importance calculations. This methodological gap creates a critical blind spot in our understanding, as the true relationships between input variables and outcomes cannot be definitively established through these approaches alone. When researchers deploy supervised learning models to identify important features, they operate under an unvalidated assumption that the model's internal feature weightings correspond to actual physical or chemical mechanisms. This represents a significant methodological challenge because researchers using supervised models must carefully distinguish between two separate and fundamentally different accuracy metrics that are often conflated: target prediction accuracy, which refers to how well the model predicts outcomes and can be readily quantified through standard evaluation metrics, and feature importance accuracy, which refers to how correctly the model attributes causality and remains largely unverifiable without external validation mechanisms. It is critical to emphasize that high target prediction accuracy does not imply feature importance accuracy. A model may achieve excellent predictive performance while simultaneously misattributing the underlying drivers of that performance, and this distinction is central to interpreting our findings.

Our findings reveal a fundamental challenge in applying supervised

machine learning for feature importance analysis in desalination research. The striking instability observed in supervised models, specifically Random Forest and XGBoost, represents a significant concern for researchers seeking to identify true causal associations rather than mere predictive patterns. This instability persists regardless of whether SHAP explanations are employed, suggesting that explanation frameworks cannot overcome the inherent limitations of the underlying supervised models. Throughout this discussion, we explicitly distinguish between two separate properties: ranking stability, which refers to the reproducibility and consistency of feature importance rankings under data perturbation, and causal interpretation, which refers to the degree to which identified associations reflect true physical, chemical, or mechanistic relationships. Our analysis primarily addresses the former, while using additional validation criteria to cautiously extend interpretations toward the latter.

The observed instability in supervised models can be explained by their target-driven optimization nature. When supervised models optimize for target prediction accuracy, they develop complex interdependencies between features that may not reflect true causal relationships. Removing a single important feature forces these models to completely reconfigure their internal feature importance structures, resulting in dramatic ranking shifts that suggest the original importance rankings were contingent on model configuration rather than fundamental to the data structure. This phenomenon highlights a deeper issue in that supervised models conflate two distinct processes, namely learning to predict outcomes and learning to explain them. The feature importance rankings generated by these models are byproducts of prediction optimization, not independent assessments of variable relevance. Consequently, even highly accurate supervised models may produce feature importance rankings that are model-specific artifacts rather than generalizable scientific insights. Researchers should therefore interpret SHAP values and similar explanation outputs as reflections of a model's internal decision logic, not as direct evidence of causal mechanisms operating within desalination systems.

In contrast, unsupervised methods including Feature Agglomeration and HVGS, and the non-target-prediction approach of Spearman correlation, demonstrate remarkable stability. These methods identify feature relationships based on inherent data characteristics rather than optimizing for target prediction, resulting in consistent feature rankings even when the dataset is perturbed. This stability suggests these approaches may be capturing more fundamental structural relationships within the data. Importantly, the stability demonstrated by Spearman's rank correlation can be attributed to a specific methodological distinction in that Spearman's correlation contains no built-in prediction mechanism and therefore operates entirely without label-driven errors. Supervised models are susceptible to label noise, class imbalance, and annotation bias, all of which can distort learned associations. By operating independently of outcome labels, Spearman's correlation avoids these confounding influences and captures intrinsic rank-order relationships between variables. Nevertheless, we acknowledge that stability in Spearman's rankings does not by itself confirm causal directionality. To move cautiously toward causal interpretation, we further evaluated whether the associations identified by Spearman's correlation satisfy two additional criteria widely recognized for validating true associations. The first is consistency, whereby Spearman's correlation consistently identified the same associations across subgroups and experimental conditions, suggesting the detected relationships are reproducible and not artifacts of a specific data configuration. The second is dose–response relationships, whereby the results revealed graded, monotonic relationships between key variables, which is a hallmark of mechanistically meaningful associations and a recognized criterion for inferring genuine underlying relationships. Together, these three properties of stability, consistency, and dose–response patterns provide a stronger, multi-faceted basis for cautiously interpreting Spearman's findings beyond mere statistical consistency, while still stopping short of definitive causal claims without experimental validation.

The superior test R^2 performance of Spearman correlation is particularly noteworthy and warrants careful interpretation. While supervised models typically excel in prediction tasks due to their optimization objectives, Spearman correlation achieved the highest test R^2 (0.4839) while simultaneously maintaining stable feature rankings. This result challenges the conventional assumption that predictive power must be traded for interpretability or stability. However, it is important not to overinterpret this finding as proof of causal superiority. Rather, this result suggests that simpler, label-independent approaches can capture generalizable data structures that translate effectively to predictive performance, precisely because they are not overfit to label-specific patterns. The superior generalization performance of Spearman correlation may reflect the absence of label-driven overfitting rather than a deeper causal understanding of the system, and this interpretation is consistent with the known risk of supervised models sacrificing generalizability for training optimization.

These results carry significant implications for desalination research practice. The widespread adoption of supervised models with SHAP for feature importance analysis may be leading researchers to draw potentially misleading conclusions about the relative importance of various factors in desalination processes. The instability we observed suggests that many published feature importance rankings may be artifacts of model-specific optimization rather than reflections of true physical or chemical relationships. Critically, researchers must avoid conflating a model's predictive success with mechanistic insight, as a supervised model that accurately predicts membrane flux does not necessarily identify which physicochemical variables are the true drivers of that flux. Disentangling these two interpretations requires explicit stability testing and, ideally, experimental validation. Furthermore, our findings highlight the importance of stability testing as a standard component of feature selection workflows. The dramatic ranking shifts observed in supervised models indicate that researchers should implement stability assessments, such as our leave-top1-out approach, before drawing conclusions about feature importance. Without such testing, researchers risk overinterpreting model-specific patterns as general scientific insights.

While our study provides valuable insights into the stability and reliability of different feature importance methods in desalination research, several important limitations should be acknowledged. Our experiments were conducted on a single desalination dataset, which may not capture the full range of conditions and relationships present in other desalination systems or water treatment processes, and the stability patterns observed might vary across datasets with unique characteristics. The dataset used in this study has a specific size and feature-to-sample ratio that could influence model behavior, and our findings regarding stability might differ in scenarios with substantially larger datasets or different dimensionality characteristics. The relationships between features in our desalination dataset reflect specific physical and chemical interactions relevant to membrane processes, and the stability patterns observed may not generalize to other scientific domains with fundamentally different underlying causal structures. Our analysis does not account for potential temporal dependencies or operational drift in desalination processes that might affect feature importance over time, as in real-world applications the relative importance of various factors may evolve as operating conditions, fouling states, or membrane characteristics change. Finally, and most importantly, while our use of consistency and dose–response criteria moves cautiously toward causal interpretation, definitive causal validation ultimately requires controlled experimental studies or the application of formal causal inference frameworks such as structural equation modeling or interventional analysis, which fall outside the scope of the current work, and future studies should integrate these approaches to build upon the foundational stability findings presented here.

5. Conclusion

This study identifies critical methodological concerns regarding the application of supervised machine learning models for feature importance analysis, demonstrated through a systematic evaluation using a desalination dataset. Within the scope of the presented experiments, supervised models, including those enhanced with SHAP explanations, exhibited marked instability in feature importance rankings under leave-top1-out perturbation testing, suggesting that caution is warranted when interpreting feature importance assessments derived from such approaches in desalination research.

Unsupervised methods and non-target-prediction approaches, by contrast, demonstrated consistently stable feature importance rankings across the perturbation conditions examined in this study. The stability observed in these methods, combined with the competitive predictive performance of Spearman correlation on the analyzed dataset, suggests that unsupervised approaches may more reliably reflect underlying relationships in desalination processes, at least under conditions similar to those represented in the present dataset.

Based on the findings of this study, desalination researchers are encouraged to adopt a more cautious approach to feature importance interpretation when employing supervised machine learning techniques. The leave-top1-out stability testing approach introduced here offers a straightforward and practical method for evaluating the robustness of feature importance rankings, and its implementation as a standard validation practice is recommended when feature assessment is a primary research objective.

The observations reported here are based on a single desalination dataset, and broader generalization of these findings across different desalination processes, operating conditions, and dataset characteristics warrants further investigation. Future research should therefore evaluate the generalizability of the identified stability patterns across diverse desalination datasets and process types. Additionally, the development of hybrid approaches that integrate the predictive capacity of supervised learning with the ranking stability of unsupervised methods represents a promising direction for producing more robust and physically interpretable feature importance assessments.

In conclusion, while supervised machine learning models with SHAP explanations retain clear value for predictive modeling tasks, the findings of this study indicate that their direct application for feature importance analysis in desalination research requires careful methodological consideration. Implementing the consistency and dose–response validation criteria proposed here, alongside stability testing, provides researchers with a more rigorous framework for distinguishing genuine physicochemical associations from algorithmically induced artifacts, thereby contributing to more reliable and reproducible scientific insights in desalination research.

Funding.
This research has no fund.

Consent to participate.
Not applicable.

Consent for publication
Not applicable.

Availability of data and material.
Not applicable.

Code availability.
Not applicable.

AI use.
Not applicable.

Authors' contributions
Yoshiyasu Takefuji completed this research and wrote this article.

According to ScholarGPS, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 25th out of 1,287,415 scholars in life sciences, 22nd out of 805,705 in COVID-19, and 1st out of 109,919 in environmental sciences.

Ethics approval.
Not applicable.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The author 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

The authors do not have permission to share data.

References

[1] J.P. Ioannidis, Why most discovered true associations are inflated, Epidemiology 19 (5) (2008) 640–648, https://doi.org/10.1097/EDE.0b013e31818131e7.
[2] V. Prasad, A.B. Jena, Prespecified falsification end points: can they validate true observational associations? JAMA 309 (3) (2013) 241–242, https://doi.org/10.1001/jama.2012.96867.
[3] J.P.A. Ioannidis, Genetic associations: false or true? Trends Mol. Med. 9 (4) (2003) 135–138, https://doi.org/10.1016/S1471-4914(03)00030-3.
[4] M.R. Roberts, S. Ashrafzadeh, M.M. Asgari, Research Techniques made simple: Interpreting measures of Association in Clinical Research, J. Invest. Dermatol. 139 (3) (2019) 502–511.e1, https://doi.org/10.1016/j.jid.2018.12.023.
[5] Q. Lai, R. Dannenfelser, J.P. Roussarie, V. Yao, Disentangling associations between complex traits and cell types with seismic, Nat. Commun. 16 (1) (2025) 8744, https://doi.org/10.1038/s41467-025-63753-z.
[6] D. Prada, B. Ritz, A.Z. Bauer, A.A. Baccarelli, Evaluation of the evidence on acetaminophen use and neurodevelopmental disorders using the Navigation Guide methodology, Environ. Health 24 (1) (2025) 56, https://doi.org/10.1186/s12940-025-01208-0.
[7] E. Stamatakis, M. Ahmadi, R.K. Biswas, C.B. Del Pozo, C. Thøgersen-Ntoumani, M. H. Murphy, A. Sabag, S. Lear, C. Chow, J.M.R. Gill, M. Hamer, Device-measured vigorous intermittent lifestyle physical activity (VILPA) and major adverse cardiovascular events: evidence of sex differences, Br. J. Sports Med. 59 (5) (2025) 316–324, https://doi.org/10.1136/bjsports-2024-108484.
[8] M. Ye, Y. He, Y. Xia, Z. Zhong, X. Kong, Y. Zhou, W. Wang, S. Qin, Q. Li, Association between bowel movement frequency, stool consistency and MAFLD and advanced fibrosis in US adults: a cross-sectional study of NHANES 2005-2010, BMC Gastroenterol. 24 (1) (2024) 460, https://doi.org/10.1186/s12876-024-03547-7.
[9] B.R. Underwood, I. Lourida, J. Gong, S. Tamburin, E.Y.H. Tang, E. Sidhom, et al., Deep Dementia Phenotyping (DEMON) network. Data-driven discovery of associations between prescribed drugs and dementia risk: a systematic review, Alzheimers Dement (n y) 11 (1) (2025) e70037, https://doi.org/10.1002/trc2.70037.
[10] Y. Takefuji, Model-specific feature importances: Distinguishing true associations from target-feature relationships, J. Affect. Disord. 369 (2025) 390–391, https://doi.org/10.1016/j.jad.2024.10.019.
[11] S.I. Abba, S. Isah, Design of real-time hybrid nanofiltration/reverse osmosis seawater desalination plant performance based on deep learning application, Mendeley Data. V1 (2025).
[12] desalination dataset (Data_Desalination.xlsx) and crossvalid.py. https://github.com/y-takefuji/SHAP.