Chemosphere 396 (2026) 144846

Contents lists available at ScienceDirect

Chemosphere

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

When high accuracy misleads: Stability limits of supervised feature importance in QSAR biodegradation

Yoshiyasu Takefuji

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

HIGHLIGHTS

- 107 Chemosphere biodegradation papers in 2025; ML widely applied.
- Two accuracies matter: target prediction vs feature-importance reliability.
- Feature importance lacks ground truth; rankings are model-specific and unstable.
- Unsupervised methods yield more stable rankings with competitive accuracy.
- Reproducible QSAR study (1055 comps, 41 feats); code in GitHub 'biodeg.py'.

ARTICLE INFO

Keywords:
QSAR biodegradation
Feature importance stability
Supervised vs unsupervised selection
Interpretability in materials science
Model-induced bias

ABSTRACT

Supervised machine learning excels at target prediction but can mischaracterize structure–biodegradability associations when feature importance is treated as ground truth. Using the QSAR Biodegradation dataset (1055 chemicals; 41 descriptors), we compare targeted supervised models (random forest, XGBoost, logistic regression), unsupervised methods (feature agglomeration, highly variable gene selection), and non-targeted supervised approaches (Spearman correlation). We evaluate cross-validated accuracy and ranking stability via a top-10 selection protocol and a leave-top-1-out perturbation. XGBoost attains the highest accuracy (0.8569) yet exhibits ranking instability; random forests are similarly unstable. In contrast, unsupervised and non-targeted supervised methods achieve strong accuracy (≈0.819–0.849) with perfect stability. Results caution against equating high predictive accuracy with reliable feature importance and support stability-aware, label-agnostic selection for interpretable materials science.

1. Introduction

One hundred and seven Chemosphere articles on biodegradation have been published to date in 2025 (Chand et al., 2025). Across this literature, supervised machine learning is typically used for two purposes: predicting target outcomes and estimating feature importance. However, because many materials scientists are not specialists in machine learning and may overlook algorithmic errors, biases, and confounding, these studies can mischaracterize the strength and nature of associations between chemical structure features and biodegradation.

This paper raises significant alarms regarding the use of supervised models for feature importance analysis. The concerns stem from three critical limitations: the absence of ground truth in feature importance calculations, the fact that feature importance metrics reflect contributions to prediction rather than true associations with outcomes, and label-driven errors that can propagate through supervised models. These limitations frequently lead to erroneous interpretations and misguided conclusions. Using a publicly available QSAR Biodegradation dataset, we demonstrate that feature importance rankings derived from supervised models suffer from considerable instability. We compare these results with more robust alternatives, including unsupervised approaches such as feature agglomeration and highly variable gene selection, as well as non-target-prediction methods like Spearman's correlation. Our findings highlight the need for caution when interpreting feature importance from predictive models.

Researchers should recognize that target supervised machine learning models entail two different notions of accuracy: target prediction accuracy and the reliability of feature-importance estimates. Target prediction accuracy can be validated against ground-truth labels (e.g., with models such as random forests), but feature importance values lack

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.chemosphere.2026.144846
Received 1 September 2025; Received in revised form 19 December 2025; Accepted 22 January 2026
Available online 29 January 2026

its ground truth for direct validation. This paper acknowledges the strong success of supervised models in target prediction, yet raises substantial theoretical and empirical concerns about using them for feature selection or importance attribution, given the absence of ground truth and the model specific nature of importance scores. The latter means different algorithms can yield markedly different feature importance rankings on the same dataset.

Cairone et al. (2025) evaluated several supervised models. To enhance interpretability, the authors analyzed the best-performing model using permutation feature importance (PFI)—a method that ranks features by their contribution to predicting the response variable—alongside standard performance metrics, including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2).

To illustrate these issues in a materials science context, this paper analyzes the QSAR Biodegradation benchmark dataset comprising 1055 chemicals and 41 features, and evaluates the effectiveness and consistency of feature selection and feature importance across diverse algorithms, including targeted supervised models (random forest, XGBoost, logistic regression), unsupervised methods (feature agglomeration, highly variable gene selection), and non-target supervised approaches (Spearman's correlation with p-values). We select the top 10 features from the full set and assess performance via cross-validated accuracy, interpreting higher accuracy as indicative of better feature selection. Next, we remove the single highest-ranked feature from the full set to create a reduced dataset, then reselect the top 9 features to assess the stability of feature ranking. The goal is to identify algorithms that deliver high predictive accuracy while maintaining robust stability in feature rankings.

Over 300 peer-reviewed studies have shown that feature importances from supervised models are inherently biased and often yield misleading interpretations because they are model-specific (Fisher et al., 2019; Steiner and Kim, 2016; Nalenz et al., 2024; Nazer et al., 2023; Ugirumurera et al., 2024; Alaimo Di Loro et al., 2023; Adler and Painsky, 2022). In methods such as random forests, feature importance reflects each feature's contribution to prediction, not its true association with the underlying phenomenon (Dunne et al., 2023; Huti et al., 2023; Loecher, 2024; Nalenz et al., 2024; Nguyen et al., 2015; Salles et al., 2021; Smith et al., 2024; Strobl et al., 2007; Wallace et al., 2023; Zarei et al., 2021). This model dependence also induces instability in feature rankings across resamples or folds. By contrast, unsupervised approaches tend to produce more stable rankings because they do not rely on labels and thus avoid label-driven bias or error. For example, unsupervised feature agglomeration leverages correlation structure and variance to assess feature relevance, and highly variable gene selection prioritizes features based on variance alone. This paper shows that unsupervised models yield more stable feature rankings than supervised models, while maintaining strong cross-validation accuracy.

Our leave-top1-out approach represents a complementary stability assessment that focuses specifically on ordered sets rather than the non-ordered feature sets typically examined in conventional methods. While traditional stability assessments evaluate whether features consistently appear across different data samples, they generally do not preserve or evaluate the stability of feature ranking orders under systematic perturbations (Ioannidis, 2008; Ioannidis, 2003; Lai et al., 2025; Prada et al., 2025; Prasad and Jena, 2013; Roberts et al., 2019; Stamatakis et al., 2025; Takefuji, 2025; Underwood et al., 2025; Ye et al., 2024).

By examining how the removal of highest-ranked features affects the subsequent ordering of remaining features, this paper introduces a method to quantify the strength of impact on feature importance stability. This ordered-set analysis provides insights into feature dependencies that may not be captured when only considering feature inclusion/exclusion without attention to ranking orders.

We acknowledge that this approach inherently advantages univariate methods over multivariate approaches. The stability differences observed in our experiments largely reflect the algorithmic characteristics of the methods, as multivariate models naturally reorder variables after removing a top predictor due to their design to capture feature interactions.

The leave-top1-out approach should be viewed as a complementary diagnostic tool rather than a replacement for conventional stability assessments. While it offers unique insights into ranking preservation under feature perturbation, it has inherent limitations in making broader claims about method reliability or identification of causal relationships without ground truth information. This ordered-set stability analysis contributes to the computational validation toolkit by examining a dimension of methodological behavior that complements existing approaches, while recognizing the context-dependent nature of feature importance stability and the need for multiple validation strategies when evaluating feature selection methods.

2. Methods

We employed one of publicly available datasets for purposes of reproducibility and transparency. The QSAR Biodegradation dataset (#1494 on OpenML) contains molecular descriptors for 1055 chemicals with binary classification of biodegradability. Its purpose is to enable prediction of environmental degradation rates based on chemical structure alone. This dataset supports environmental protection by helping identify compounds that persist versus those that degrade quickly in nature. With 41 features representing molecular properties such as topology, quantum characteristics, and constitutional indices, it serves multiple stakeholders: regulators assessing chemical safety, researchers developing eco-friendly compounds, and industry professionals screening new products. The dataset reduces reliance on expensive laboratory testing and supports green chemistry principles. As a benchmark in both machine learning and chemoinformatics, it demonstrates how computational methods can effectively bridge chemical structures and environmental behaviors.

To assess target accuracy, 5-fold cross-validation performance, and the stability of feature rankings, we examine three categories of models: targeted supervised models (including random forest, XGBoost, and logistic regression), unsupervised models (such as feature agglomeration and highly variable gene selection), and non-targeted supervised approaches (such as Spearman's correlation with p-values).

Throughout all assessments, we maintained the default parameter settings for each algorithm. For XGBoost specifically, this means utilizing an ensemble of 100 trees with a maximum tree depth of 6, as per the standard configuration of the XGBClassifier with 5-fold cross-validation. These baseline settings were preserved to ensure fair comparison across methods without introducing optimization bias. For stability testing, we employed a position-based metric to evaluate consistency in feature ranking orders. This approach involved comparing the expected positions of the remaining 9 features against their actual positions after perturbation. By quantifying the deviation between expected and observed feature rankings, we could systematically assess the resilience of different feature importance methods to variations in the dataset.

Table 1
Cross-validation accuracy and stability in feature rankings.

Algorithm	Top 10 Features CV Mean Accuracy	Stability in feature rankings
Random Forest	0.84455	Unstable: 0/9
XGBoost	0.856872	Unstable: 3/9
Logistic Regression	0.752607	Unstable: 1/9
Feature Agglomeration	0.849289	Stable:9/9
HVGS	0.818957	Stable:9/9
Spearman Correlation	0.823697	Stable:9/9

3. Results

Table 1 shows the results of mean cross-validation accuracy and stability. Across targeted supervised models, XGBoost achieved the highest cross-validated mean accuracy at 0.8569 but showed instability in feature rankings (unstable in 3/9 tests), while Random Forest followed with 0.8446 accuracy and was highly unstable (0/9 stability), and Logistic Regression trailed with 0.7526 accuracy and low stability (1/9). In contrast, unsupervised models demonstrated both strong accuracy and consistent feature importance: Feature Agglomeration reached 0.8493 accuracy with perfect stability (9/9), Spearman Correlation delivered 0.8237 accuracy with perfect stability (9/9), and HVGS posted 0.8190 accuracy with perfect stability (9/9). For non-targeted supervised models, stability was uniformly high (9/9 across the board), with accuracies spanning roughly 0.819–0.849, indicating that while targeted supervised approaches may slightly edge out in peak accuracy (notably XGBoost), they can suffer from notably less stable feature rankings compared with the consistently stable unsupervised and non-targeted supervised methods. For purposes of reproducibility and transparency, Python code, biodeg.py, is publicly available at GitHub (2025)

4. Discussion

These findings have direct implications for environmental risk assessment, where understanding which chemical properties drive biodegradability is crucial for regulatory compliance and green chemistry initiatives. When screening novel compounds, unstable feature rankings could lead to misguided molecular design strategies or inconsistent prioritization of environmental hazards. Regulatory agencies relying on predictive models to evaluate chemicals under frameworks need assurance that identified structural alerts and physicochemical drivers are robust across datasets, not artifacts of model selection. The stability advantages of unsupervised methods shown here could enhance reproducibility in QSAR biodegradation models, supporting more reliable identification of persistent pollutants and facilitating the development of environmentally benign alternatives in accordance with green chemistry principles.

While previous studies identified critical issues with feature importance accuracy in supervised models, they often failed to propose concrete solutions for calculating true associations based on fundamental principles like consistency and dose-response relationships. Our proposed stability testing methodology directly addresses this gap by systematically evaluating how removing the highest-ranked feature impacts the ordering of remaining features. This approach inherently accommodates both consistency principles and dose-response relationships by measuring whether feature importance maintains proportional rankings when the feature space is perturbed. The results clearly demonstrate that supervised models frequently suffer from instability in feature ranking orders due to label-driven errors and model-specific biases, whereas unsupervised models and non-target-prediction methods like Spearman correlation exhibit substantially stronger stability due to their independence from outcome-related noise. This distinction provides practitioners with a more reliable foundation for identifying features with genuine associations rather than merely predictive utility.

Our results highlight a key trade-off between peak predictive performance and the reliability of feature interpretation across modeling paradigms. While targeted supervised models such as XGBoost achieved the highest cross-validated mean accuracy, they exhibited markedly lower stability in feature rankings across resamples. In contrast, unsupervised and non-targeted supervised approaches delivered slightly lower—but competitive—accuracy alongside near-perfect stability in feature importance. For applications where interpretability, reproducibility, and biological or domain plausibility of features matter, this stability advantage is often decisive.

There is now substantial evidence across hundreds of studies that feature importances from supervised learning are model- and data-dependent, and thus prone to bias and instability. Label-driven bias means supervised models optimize for prediction with respect to observed labels, which may contain noise, measurement error, or confounding; as a result, learned importance reflects how features help reduce predictive loss, not necessarily their true associations with the underlying generative process. Model-specific attribution further compounds this, because importance scores such as Gini importance in random forests or gain in gradient boosting are tied to a model's functional form and training dynamics; different algorithms (and even different settings within the same algorithm) can yield divergent rankings from the same data, even when predictive accuracy is similar, undermining portability of feature interpretations across studies. Sampling variability and multicollinearity exacerbate instability: with correlated predictors, small changes in the training set can flip split choices in tree-based models or alter coefficients in linear models with regularization, thereby reshuffling importance rankings. Supervised models may also assign high importance to features that capture interactions or higher-order effects that aid prediction, even if those features are not primary drivers of the outcome, while clinically or biologically meaningful features can receive low importance if their effects are partially masked by correlated surrogates.

By contrast, the unsupervised and non-targeted supervised methods we evaluated demonstrated strikingly consistent feature rankings, with only modest differences in accuracy relative to the best targeted supervised model. Label-free criteria drive feature scoring by intrinsic data structure—correlation, variance, and redundancy—rather than by contingent patterns in the outcome variable, removing label-induced noise and reducing overfitting to idiosyncrasies of a particular outcome or sample split.

Methods such as Feature Agglomeration explicitly leverage correlation structures to group redundant features and emphasize representative variables, stabilizing rankings because correlated features are treated systematically rather than competitively. Variance-based selection approaches like HVGS prioritize features with consistently high dispersion across samples, a property that tends to be more stable under resampling than label-conditioned importance. Even when labels are present, pipelines that decouple feature scoring from model idiosyncrasies—such as using correlation or mutual information screens prior to modeling, or applying model-agnostic stability selection—can preserve high stability while maintaining competitive accuracy.

The empirical pattern observed here—XGBoost achieving the highest mean accuracy (0.8569) with moderate instability, Random Forest attaining slightly lower accuracy (0.8446) with severe instability, and Logistic Regression lagging in both (0.7526 accuracy, low stability)—underscores that optimizing solely for accuracy can be misleading when the downstream goal includes interpretation or feature prioritization.

Meanwhile, unsupervised methods (Feature Agglomeration at 0.8493, Spearman Correlation at 0.8237, HVGS at 0.8190, each with 9/9 stability) and non-targeted supervised methods (9/9 stability across roughly 0.819–0.849 accuracy) offer a more dependable basis for inference, biomarker discovery, or hypothesis generation. When interpretability is paramount—such as for identifying candidate biomarkers, policy-relevant drivers, or actionable features—methods with demonstrated ranking stability should be prioritized, even at a small cost in peak accuracy, because unstable rankings risk irreproducible findings and misallocated validation resources. If targeted supervised models are used for their predictive advantages, they should be complemented with stability diagnostics: perform bootstrapped or cross-validated ranking stability analyses, examine sensitivity to correlated feature pruning, and report confidence sets of important features rather than single ordered lists. It is also prudent to use model-agnostic or label-free feature scoring as a foundation—techniques like Feature Agglomeration, Spearman Correlation, and HVGS can provide a stable shortlist that can then be validated or refined with supervised models—while proactively

addressing multicollinearity via clustering correlated features, group-wise selection, or dimensionality reduction to mitigate arbitrary rank swapping among surrogates. When feasible, permutation-based, conditional, or causal-oriented importance measures can reduce some model-specific biases and better reflect unique contributions, though they remain susceptible to sampling variability.

These findings should be interpreted with several caveats. Stability was assessed under a specific resampling design and feature space; alternative cohorts, noise levels, or feature-engineering choices may shift the balance between accuracy and stability. Accuracy differences, while statistically meaningful, may not translate to material improvements in deployment; future work should therefore assess calibration, fairness, and decision-relevant utility alongside stability. Incorporating domain knowledge—such as pathway or group constraints for genes—may further enhance both stability and interpretability by aligning feature selection with underlying structure. Finally, causal feature attribution remains an open challenge; extending these evaluations to counterfactual or interventional settings could help separate predictive surrogates from true drivers. In sum, the results reinforce a critical message: feature importance from supervised models reflects predictive utility under a specific model and sample, not necessarily true association. For robust, reproducible feature prioritization—especially in high-dimensional, correlated settings—unsupervised or non-targeted supervised strategies, coupled with explicit stability assessment, provide a more reliable pathway.

5. Conclusion

This study makes several significant contributions to the field of feature importance analysis and model interpretability. First, we empirically demonstrate the critical trade-off between predictive performance and feature ranking stability across supervised and unsupervised modeling approaches. While targeted supervised models like XGBoost achieved marginally higher accuracy, they exhibited substantially lower stability in feature rankings compared to unsupervised and non-targeted methods, which maintained near-perfect stability with competitive accuracy.

Second, we provide systematic evidence of how label-driven bias in supervised models compromises the reliability of feature importance. When models optimize for prediction against noisy or confounded labels, importance scores reflect predictive utility rather than true associations with underlying phenomena. This distinction is crucial for applications where feature interpretability guides scientific discovery or decision-making.

Third, our results highlight how model-specific attribution mechanisms further undermine the portability of feature interpretations across studies. Different algorithms—and even parameter changes within the same algorithm—can yield divergent feature rankings despite similar predictive performance, compounded by sampling variability and multicollinearity among predictors.

Perhaps most importantly, we demonstrate that unsupervised and non-targeted methods offer a more dependable foundation for inference and hypothesis generation by leveraging intrinsic data structures rather than contingent outcome patterns. Feature Agglomeration, Spearman Correlation, and highly variable gene selection achieved remarkable stability with only modest accuracy trade-offs relative to the best supervised models.

These findings have profound implications for fields where interpretability is paramount, such as biomarker discovery, policy analysis, and clinical decision support. We recommend that researchers prioritize methods with demonstrated ranking stability when interpretation matters, complemented by rigorous stability diagnostics when using supervised models for their predictive advantages.

Consent to participate

Not applicable.

Ethics approval

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The author has no permission to share data.

Code availability

Python code is publicly available at GitHub.

AI use

Not applicable.

Funding

This research has no fund.

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

The authors do not have permission to share data.

References

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

Alaimo Di Loro, P., Scacciatelli, D., Tagliaferri, G., 2023. 2-step gradient boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy. Stat. Methods Appl. 32, 237–270. https://doi.org/10.1007/s10260-022-00643-4.

Cairone, S., et al., 2025. Enhancing process monitoring and control in novel carbon capture and utilization biotechnology through artificial intelligence modeling: an advanced approach toward sustainable and carbon-neutral wastewater treatment. Chemosphere 376, 144299. https://doi.org/10.1016/j.chemosphere.2025.144299.

Chand, S., et al., 2025. Sustainable synthesis and multifunctional applications of biowaste-derived carbon nanomaterials and metal oxide composites: a review. Chemosphere 385, 144540. https://doi.org/10.1016/j.chemosphere.2025.144540.

Dunne, R., Reguant, R., Ramarao-Milne, P., Szul, P., Sng, L.M.F., Lundberg, M., Twine, N. A., Bauer, D.C., 2023. Thresholding Gini variable importance with a single-trained random forest: an empirical Bayes approach. Comput. Struct. Biotechnol. J. 21, 4354–4360. https://doi.org/10.1016/j.csbj.2023.08.033.

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.

GitHub, 2025. Biodeg.py. https://github.com/y-takefuji/biodegradation/blob/main/biodeg.py.

Huti, M., Lee, T., Sawyer, E., King, A.P., 2023. An investigation into race bias in random forest models based on breast DCE-MRI derived radiomics features. Clinical Image Based Procedure Fairness AI Med Imaging Ethical Philos Issues Med Imaging 14242, 225–234. https://doi.org/10.1007/978-3-031-45249-9_22.

Ioannidis, J.P., 2008. Why most discovered true associations are inflated. Epidemiology 19 (5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7.

Ioannidis, J.P.A., 2003. Genetic associations: false or true? Trends Mol. Med. 9 (4), 135–138. https://doi.org/10.1016/S1471-4914(03)00030-3.

Lai, Q., Dannenfelser, R., Roussarie, J.P., Yao, V., 2025. Disentangling associations between complex traits and cell types with seismic. Nat. Commun. 16 (1), 8744. https://doi.org/10.1038/s41467-025-63753-z.

Loecher, M., 2024. Debiasing SHAP scores in random forests. AStA Advances in Statistical Analysis 108, 427–440. https://doi.org/10.1007/s10182-023-00479-7.
Nalenz, M., Rodemann, J., Augustin, T., 2024. Learning de-biased regression trees and forests from complex samples. Mach. Learn. 113, 3379–3398. https://doi.org/10.1007/s10994-023-06439-1.
Nazer, L.H., Zatarah, R., Waldrip, S., et al., 2023. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digital Health 2 (6), e0000278. https://doi.org/10.1371/journal.pdig.0000278.
Nguyen, T.T., Huang, J.Z., Nguyen, T.T., 2015. Unbiased feature selection in learning random forests for high-dimensional data. Sci. World J., 471371 https://doi.org/10.1155/2015/471371.
Prada, D., Ritz, B., Bauer, A.Z., Baccarelli, A.A., 2025. Evaluation of the evidence on acetaminophen use and neurodevelopmental disorders using the Navigation Guide methodology. Environmental health : a global access science source 24 (1), 56. https://doi.org/10.1186/s12940-025-01208-0.
Prasad, V., Jena, A.B., 2013. Prespecified falsification end points: can they validate true observational associations? JAMA 309 (3), 241–242. https://doi.org/10.1001/jama.2012.96867.
Roberts, M.R., Ashrafzadeh, S., Asgari, M.M., 2019. Research techniques made simple: interpreting measures of Association in clinical research. J. Invest. Dermatol. 139 (3), 502–511.e1. https://doi.org/10.1016/j.jid.2018.12.023.
Salles, T., Rocha, L., Gonçalves, M., 2021. A bias-variance analysis of state-of-the-art random forest text classifiers. Advances in Data Analysis and Classification 15, 379–405. https://doi.org/10.1007/s11634-020-00409-4.
Smith, H.L., Biggs, P.J., French, N.P., et al., 2024. Lost in the Forest: encoding categorical variables and the absent levels problem. Data Min. Knowl. Discov. 38, 1889–1908. https://doi.org/10.1007/s10618-024-01019-w.
Stamatakis, E., Ahmadi, M., Biswas, R.K., Del Pozo Cruz, B., Thøgersen-Ntoumani, C., Murphy, M.H., Sabag, A., Lear, S., Chow, C., Gill, J.M.R., Hamer, M., 2025. Device-measured vigorous intermittent lifestyle physical activity (VILPA) and major adverse cardiovascular events: evidence of sex differences. Br. J. Sports Med. 59 (5), 316–324. https://doi.org/10.1136/bjsports-2024-108484.
Steiner, P.M., Kim, Y., 2016. The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. J. Causal Inference 4 (2), 20160009. https://doi.org/10.1515/jci-2016-0009.
Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinf. 8, 25. https://doi.org/10.1186/1471-2105-8-25.
Takefuji, Y., 2025. 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.
Ugirumurera, J., Bensen, E.A., Severino, J., Sanyal, J., 2024. Addressing bias in bagging and boosting regression models. Sci. Rep. 14 (1), 18452. https://doi.org/10.1038/s41598-024-68907-5.
Underwood, B.R., Lourida, I., Gong, J., Tamburin, S., Tang, E.Y.H., Sidhom, E., et al., Deep Dementia Phenotyping (DEMON) Network, 2025. Data-driven discovery of associations between prescribed drugs and dementia risk: a systematic review. Alzheimer's Dementia 11 (1), e70037. https://doi.org/10.1002/trc2.70037.
Wallace, M.L., Mentch, L., Wheeler, B.J., Lyons, M., Reichmann, W.M., 2023. Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. BMC Med. Res. Methodol. 23 (1), 144. https://doi.org/10.1186/s12874-023-01965-x.
Ye, M., He, Y., Xia, Y., Zhong, Z., Kong, X., Zhou, Y., Wang, W., Qin, S., Li, Q., 2024. 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), 460. https://doi.org/10.1186/s12876-024-03547-7.
Zarei, M., Najarchi, M., Mastouri, R., 2021. Bias correction of global ensemble precipitation forecasts by Random forest method. Earth Science Informatics 14, 677–689. https://doi.org/10.1007/s12145-021-00577-7.