



Short communication

## Ensuring reliable feature importance in food chemistry AI

Yoshiyasu Takefuji<sup>\*</sup>

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

## ARTICLE INFO

## Keywords:

Feature importance bias  
Machine learning reliability  
Food chemistry data analysis  
Model assumption violations  
Multifaceted analytical approaches

## ABSTRACT

Food chemistry's rapid AI adoption (2060 AI articles; 415 in 2025) spans machine learning, logistic regression, random forests, and XGBoost. Yet a skills gap in supervised learning fuels misinterpretation: Models optimize prediction, not true associations, and feature importances lack ground-truth validation. High accuracy does not ensure reliable attributions; importances are model- and data-biased. Using a microplastic–cancer case, we show parametric logistic regression on nonlinear data distorts inference. We propose a standards-based pipeline: Unsupervised structure discovery (e.g., feature agglomeration, highly variable feature selection), nonparametric association tests (spearman with  $p$ -values), and explicit stability audits of rankings. This multifaceted approach mitigates label-driven bias, improves robustness, and aligns AI insights with mechanistic understanding, supporting credible risk assessment and safer application of AI in food chemistry

### 1. Introduction

Food Chemistry has witnessed an unprecedented surge in artificial intelligence research, publishing 2060 AI-related articles, with 415 appearing in 2025 alone. This trend extends to other computational methodologies, including 799 machine learning studies (385 in 2025), 113 articles employing logistic regression (47 in 2025), 392 papers utilizing random forest techniques (167 in 2025), and 64 studies applying XGBoost methodologies (38 in 2025). This publication pattern reflects the food chemistry community's rapidly growing adoption of advanced computational approaches for addressing complex food chemical challenges.

Despite this computational revolution, a significant knowledge gap persists among many domain experts who lack fundamental understanding of supervised machine learning principles. This gap has led to widespread AI misapplications in food chemistry research, where sophisticated tools are employed without adequate comprehension of their underlying mechanisms. The situation is particularly problematic because supervised machine learning models exhibit two distinct types of accuracy: target prediction accuracy and feature importance reliability. While the former can be objectively validated against ground truth values (labels), the latter lacks corresponding validation benchmarks, making the reliability of feature importances difficult to assess systematically.

This absence of ground truth for feature importances leads to two critical insights that are often overlooked. First, feature importances

derived from supervised models are inherently biased, reflecting contributions to prediction rather than true causal relationships between variables. Second, high target prediction accuracy does not guarantee reliable feature importances due to label-driven errors that can significantly compromise interpretability. To illustrate these issues, this paper presents a practical example applying parametric logistic regression to nonparametric nonlinear data from a real-world study examining microplastic exposure and cancer health risk. This case exemplifies how mismatches between model assumptions and data characteristics can fundamentally undermine the validity of feature importance interpretations in environmental health assessments.

This paper raises urgent concerns regarding supervised machine learning applications in hazardous materials research. Researchers must recognize that no algorithm can definitively calculate true associations between variables with complete certainty. Relying exclusively on a single methodological approach is inherently inadequate, necessitating multifaceted analytical strategies in all cases. Furthermore, data analysis outcomes are reliable only when the underlying assumptions of the chosen models are satisfied—a fundamental requirement frequently neglected in practical applications. Violating these assumptions can severely distort results and lead to misleading conclusions, potentially resulting in flawed risk assessments or ineffective remediation strategies. Consequently, researchers should conduct comprehensive stability analyses on feature importance rankings, acknowledging that importances derived from supervised models are model-specific and may vary substantially across different algorithms. This validation approach

<sup>\*</sup> Corresponding author at: Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan.

E-mail address: [takefuji@keio.jp](mailto:takefuji@keio.jp).

<https://doi.org/10.1016/j.foodchem.2026.148515>

Received 23 November 2025; Received in revised form 3 February 2026; Accepted 15 February 2026

Available online 18 February 2026

0308-8146/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

ensures that findings related to hazardous materials are both scientifically rigorous and practically applicable.

This paper showcases a single example of parametric logistic regression against nonparametric nonlinear data analysis. Xu et al. investigated the relationship between microplastics in human feces and colorectal cancer risk, revealing striking findings (Xu et al., 2025). Their multivariable logistic regression analysis demonstrated that individuals in the highest exposure quartile had a substantially elevated adjusted odds ratio of 11.3 (95% CI: 6.77–19.5,  $p$  for trend <0.01) compared to those in the lowest quartile. A restricted cubic spline analysis indicated a nonlinear dose-response relationship between microplastic exposure and cancer risk. Notably, their stratified analyses suggested this association was particularly pronounced among females and individuals who frequently consumed spicy or high-fat foods (Xu et al., 2025).

When parametric methods like logistic regression are applied to nonparametric data that violate these assumptions, or when linear methods are applied to inherently nonlinear relationships, multiple statistical outputs including coefficient estimates, standard errors, test statistics,  $p$ -values, confidence intervals, and odds ratios can be substantially distorted (Dey et al., 2025; Pinheiro-Guedes et al., 2024; Wang et al., 2023; Osborne, 2015; van Maanen et al., 2019; Work et al., 1989; Zulfadhli et al., 2024; Akturk et al., 2025; Rifada et al., 2022; Suliyanto et al., 2020; Wibowo et al., 2021; Steyerberg et al., 2011; Özkale, 2016). This fundamental mismatch between analytical method and data characteristics potentially undermines the validity of the study's striking conclusions about microplastic exposure and colorectal cancer risk.

While supervised models possess two distinct types of accuracy: target prediction accuracy and feature importance reliability. While target prediction accuracy can be validated against ground truth labels, feature importances lack its ground truth for accuracy validation. Feature importances derived from supervised models are inherently biased or skewed (Adler & Painsky, 2022; Alaimo Di Loro et al., 2023; Dunne et al., 2023; Fisher et al., 2019; Huti et al., 2023; Nalenz et al., 2024; Nazer et al., 2023; Nguyen et al., 2015; Salles et al., 2021; Smith et al., 2024; Steiner & Kim, 2016; Strobl et al., 2007; Ugirumurera et al., 2024; Wallace et al., 2023; Zarei et al., 2021).

This paper challenges the effectiveness of feature selection across supervised models, unsupervised models and non-targeted methods to ensure reliable feature importance assessment. This paper reveals that multifaceted approaches such as unsupervised models and non-targeted methods produce reliable feature importance assessment rather than solely relying on supervised models due to label-driven errors and biases.

## 2. Methods

Due to the absence of algorithms to accurately calculate true associations between variables, this paper employs multifaceted approaches including unsupervised models such as feature agglomeration (FA) and highly variable gene selection (HVGS), and followed by non-targeted nonlinear nonparametric statistical methods such as Spearman's correlation with  $p$ -values instead of solely relying on parametric logistic regression.

Due to the absence of datasets from Xu et al., this paper examines the effectiveness of feature selection across supervised models such as Random Forest and XGBoost, unsupervised models like feature agglomeration (FA) and highly variable feature selection (HVFS), and non-parametric correlation methods such as Spearman's correlation. We utilize a public dataset on microplastic consumption in various foods from 1990 to 2018, comprising 723 instances and 21 features (Kaggle, 2025). The study identifies the top 5 features from the full feature set and employs cross-validation to evaluate the effectiveness of each algorithm, with higher cross-validation accuracy indicating superior feature selection. To assess stability in feature importance rankings, we remove the highest-ranked feature from the full dataset, create a reduced dataset, and then re-select the top 4 features, comparing the

consistency of selections between the original and reduced datasets.

## 3. Results

Our analysis reveals significant differences in feature selection performance across methods. As shown in Table 1, Feature Agglomeration (FA) and Highly Variable Feature Selection (HVFS) demonstrated identical performance with the highest cross-validation accuracy (0.9347), selecting the same features in identical order. Random Forest achieved the second-highest accuracy (0.9255), followed by Spearman's correlation (0.9138) and XGBoost (0.9129). Notably, all methods identified "total\_milk" as the most important feature in the full dataset, while "non-starchy\_vegetables" and "fruits" consistently appeared in the top selections across all methods. When evaluating stability after removing the top feature, FA and HVFS maintained perfect consistency (100% stability) in their rankings, preserving all four secondary features in identical order. In contrast, supervised methods showed greater variability, with Random Forest and XGBoost introducing "potatoes" and "eggs" in their reduced feature sets, indicating lower stability in feature importance rankings.

For purposes of reproducibility and transparency, Python code, mp.py is publicly available at GitHub (GitHub, 2025).

## 4. Discussion

The study by Xu et al. focused on a total of 258 CRC patients and 493 healthy controls containing no food type features while this paper examines diverse food types. Individual dietary variations, such as those related to lactose intolerance or other dietary restrictions, could alter microplastic exposure patterns. Our analysis identifies milk as a primary source based on average population consumption patterns, though this would naturally differ for non-milk consumers, highlighting the importance of considering dietary subgroups in future research.

The unprecedented proliferation of artificial intelligence methodologies in food chemistry, as evidenced by Food Chemistry's publication of numerous AI-related articles, represents both a remarkable opportunity and a significant challenge for the field. Our findings reveal a fundamental disconnect between this computational revolution and the methodological understanding required to apply these techniques appropriately, particularly in food chemistry where conclusions directly impact human health, safety, and regulatory policy.

Our analysis demonstrates that supervised models inherently suffer from instability in feature-importance rankings due to label-driven errors, while unsupervised models and non-targeted methods exhibit

**Table 1**  
cross-validation accuracy and stability in feature ranking orders.

model	CV mean accuracy	Top 5 ranking orders	Top 4 ranking orders
random forest	0.9255	total_milk, fruits, non-starchy_vegetables, other_starchy_vegetables, unprocessed_red_meats	eggs, potatoes, fruits, non-starchy_vegetables
XGBoost	0.9129	total_milk, fruits, non-starchy_vegetables, other_starchy_vegetables, eggs	eggs, potatoes, fruits, non-starchy_vegetables
FA	0.9347	total_milk, non-starchy_vegetables, fruits, refined_grains, other_starchy_vegetables	non-starchy_vegetables, fruits, refined_grains, other_starchy_vegetables
HVGS	0.9347	total_milk, non-starchy_vegetables, fruits, refined_grains, other_starchy_vegetables	non-starchy_vegetables, fruits, refined_grains, other_starchy_vegetables
Spearman	0.9138	total_milk, fruits, potatoes, unprocessed_red_meats, non-starchy_vegetables	fruits, potatoes, unprocessed_red_meats, non-starchy_vegetables

substantially stronger stability. This disparity stems from a critical limitation of supervised approaches: they optimize for prediction accuracy rather than capturing true variable associations. The perfect consistency (100% stability) demonstrated by Feature Agglomeration (FA) and Highly Variable Feature Selection (HVFS) in our experiment stands in stark contrast to the variability observed in Random Forest and `XGBoost`, which introduced entirely new features (“potatoes” and “eggs”) when the dataset was modified—highlighting their vulnerability to dataset perturbations.

This instability is particularly concerning when examining complex exposure–outcome relationships in food chemistry, such as contaminant levels versus health outcomes, where parametric logistic regression is often applied to inherently nonlinear relationships. While such analyses may reveal striking associations (e.g., high odds ratios for top exposure quartiles), our findings suggest that conclusions can be significantly compromised by methodological limitations. Nonlinear dose–response patterns, identifiable via restricted cubic splines, further underscore the mismatch between linear assumptions of logistic regression and the complex reality of food data, potentially leading to distorted coefficient estimates, misleading *p*-values, and ultimately unreliable conclusions about risk.

The food chemistry research community faces a critical methodological challenge: while prediction accuracy can be objectively validated against known outcomes, feature importance—often the primary interest when identifying key drivers of quality, safety, or nutritional outcomes—lacks corresponding ground-truth validation. This creates a scenario where researchers may achieve high predictive accuracy while drawing spurious conclusions about which variables actually matter. Our finding that unsupervised approaches outperformed supervised methods in both accuracy (0.9347 vs. 0.9255) and stability highlights the urgent need for methodological recalibration in food data analysis.

We propose a systematic approach to assess and enhance feature-importance reliability: researchers should evaluate stability by selecting top features from the full dataset, removing the highest-ranked feature, re-selecting features from the reduced dataset, and comparing the consistency between selections. Existing stability methods often focus on prediction consistency rather than the ranking of feature importance—a critical distinction when seeking mechanistic understanding rather than merely improving predictive accuracy. This straightforward validation process can reveal whether feature importances genuinely reflect underlying data structures or are artifacts of the modeling approach, providing a practical framework for distinguishing robust feature relationships from model-dependent statistical artifacts that may mislead food risk assessments.

This study reveals a troubling paradox in food chemistry AI applications: despite rapid growth in methodological sophistication, fundamental concerns about validity remain underaddressed. The inherent limitations of supervised machine learning—particularly the absence of ground-truth validation for feature importances—demand urgent attention. Our analysis demonstrates that reliance on supervised models alone, especially parametric approaches like logistic regression for nonparametric data, can lead to distorted conclusions with potentially serious implications for food safety assessments, risk management strategies, and regulatory decisions.

For food chemistry researchers seeking to leverage computational methods, our findings underscore the necessity of multifaceted analytical strategies that incorporate unsupervised learning and nonparametric methods alongside supervised techniques. Only through such methodological triangulation can we ensure that growing computational capabilities translate into genuine scientific insights rather than misleading artifacts of analysis. As artificial intelligence continues to transform food chemistry, methodological rigor must advance in parallel with computational sophistication to ensure reliable conclusions about contaminants, quality attributes, and their impacts.

Due to the absence of universal algorithms, to validate true associations, consistency and dose-response relationships assessment is

mandatory in food chemistry.

## 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 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>
- Akturk, B., Beyaztas, U., Shang, H. L., et al. (2025). Robust functional logistic regression. In *19. Advances in Data Analysis and Classification* (pp. 121–145). <https://doi.org/10.1007/s11634-023-00577-z>
- 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. *Statistical Methods and Applications*, 32, 237–270. <https://doi.org/10.1007/s10260-022-00643-4>
- Dey, D., Haque, M. S., Islam, M. M., Aishi, U. I., Shabby, M. S. A., et al. (2025). The proper application of logistic regression model in complex survey data: A systematic review. *BMC Medical Research Methodology*, 25, Article 15. <https://doi.org/10.1186/s12874-024-02454-5>
- Dunne, R., Reguant, R., Ramarao-Milne, P., Szul, P., Sng, L. M. F., Lundberg, M., ... Bauer, D. C. (2023). Thresholding Gini variable importance with a single-trained random forest: An empirical Bayes approach. *Computational and Structural Biotechnology Journal*, 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. *Journal of Machine Learning Research*, 20, 177.
- GitHub. (2025). mp.py. <https://github.com/y-takefuji/microplastic/blob/main/mp.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. In *Clinical image based procedure fairness AI med imaging ethical Philos issues med imaging*. 14242 pp. 225–234. [https://doi.org/10.1007/978-3-031-45249-9\\_22](https://doi.org/10.1007/978-3-031-45249-9_22)
- Kaggle. (2025). Microplastic dataset: processed\_microplastics.csv. <https://www.kaggle.com/datasets/jayeshmohanani/dataset-for-microplastic-consumption-in-food-items>
- van Maanen, L., Katsimpokis, D., & van Campen, A. D. (2019). Fast and slow errors: Logistic regression to identify patterns in accuracy–response time relationships. *Behavior Research Methods*, 51, 2378–2389. <https://doi.org/10.3758/s13428-018-1110-z>
- Nalenz, M., Rodemann, J., & Augustin, T. (2024). Learning de-biased regression trees and forests from complex samples. *Machine Learning*, 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), Article 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. *The Scientific World Journal*, 2015, Article 471371. <https://doi.org/10.1155/2015/471371>
- Osborne, J. (2015). A practical guide to testing assumptions and cleaning data for logistic regression. In *A practical guide to testing assumptions and cleaning data for logistic regression* (pp. 84–130). SAGE Publications, Ltd. <https://doi.org/10.4135/9781483399041.n4>
- Özkale, M. R. (2016). Iterative algorithms of biased estimation methods in binary logistic regression. *Statistical Papers*, 57, 991–1016. <https://doi.org/10.1007/s00362-016-0780-9>
- Pinheiro-Guedes, L., Martinho, C., & Martins, M. R. O. (2024). Logistic regression: Limitations in the estimation of measures of association with binary health outcomes. *Acta Médica Portuguesa*, 37(10), 697–705. <https://doi.org/10.20344/amp.21435>
- Rifada, M., Chamidah, N., & Ningrum, R. A. (2022). Estimation of nonparametric ordinal logistic regression model using generalized additive models (GAM) method based on local scoring algorithm. *AIP Conference Proceedings*, 2668(1), Article 070013. <https://doi.org/10.1063/5.0111771>
- 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 Mining and Knowledge Discovery*, 38, 1889–1908. <https://doi.org/10.1007/s10618-024-01019-w>
- Steiner, P. M., & Kim, Y. (2016). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of Causal Inference*, 4(2), 20160009. <https://doi.org/10.1515/jci-2016-0009>
- Steyerberg, E. W., Schemper, M., & Harrell, F. E. (2011). Logistic regression modeling and the number of events per variable: Selection bias dominates. *Journal of Clinical Epidemiology*, 64(12), 1464. <https://doi.org/10.1016/j.jclinepi.2011.06.016>
- 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, 25. <https://doi.org/10.1186/1471-2105-8-25>
- Ugirimurera, J., Bensen, E. A., Severino, J., & Sanyal, J. (2024). Addressing bias in bagging and boosting regression models. *Scientific Reports*, 14(1), 18452. <https://doi.org/10.1038/s41598-024-68907-5>
- 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 Medical Research Methodology*, 23(1), 144. <https://doi.org/10.1186/s12874-023-01965-x>
- Wang, T., Tang, W., Lin, Y., & Su, W. (2023). Semi-supervised inference for nonparametric logistic regression. *Statistics in Medicine*, 42(15), 2573–2589. <https://doi.org/10.1002/sim.9737>
- Wibowo, W., Amelia, R., Octavia, F. A., & Wilantari, R. N. (2021). Classification using nonparametric logistic regression for predicting working status. *AIP Conference Proceedings*, 2329(1), Article 060032. <https://doi.org/10.1063/5.0043598>
- Work, J. W., Ferguson, J. G., & Diamond, G. A. (1989). Limitations of a conventional logistic regression model based on left ventricular ejection fraction in predicting coronary events after myocardial infarction. *American Journal of Cardiology*, 64(12), 702–707. [https://doi.org/10.1016/0002-9149\(89\)90751-0](https://doi.org/10.1016/0002-9149(89)90751-0)
- Xu, J., Qu, J., Jin, H., & Mao, W. (2025). Associations between microplastics in human feces and colorectal cancer risk. *Journal of Hazardous Materials*, 495, Article 139099. <https://doi.org/10.1016/j.jhazmat.2025.139099>
- 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>
- Zulfadhli, M., Budiantara, I. N., & Ratnasari, V. (2024). Nonparametric regression estimator of multivariable Fourier series for categorical data. *MethodsX*, 13, Article 102983. <https://doi.org/10.1016/j.mex.2024.102983>