Food Chemistry 494 (2025) 146171

Contents lists available at ScienceDirect

Food Chemistry

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

Letter to the Editor

Addressing Bias in machine learning feature importance for food quality assessment

ARTICLE INFO

Keywords:
Almond deterioration
Machine learning
Random Forest
SHapley additive exPlanations
Feature importance
Statistical validation

ABSTRACT

Li et al. (2025) highlighted Random Forest's (RF) high accuracy and SHapley Additive exPlanations (SHAP)-derived feature importance for almond deterioration. However, concerns persist regarding the reliability of these interpretations, as high predictive accuracy doesn't guarantee valid feature rankings due to inherent biases in tree-based models, further amplified by SHAP's model dependency. To mitigate this, integrating robust statistical methods such as Spearman's rho, Kendall's tau, Total correlation and Effective transfer entropy is crucial for unbiased assessment. This combined approach ensures a more reliable evaluation of key indicators. Future research should prioritize methodologies combining machine learning with rigorous statistical validation for more interpretable and trustworthy insights in complex biological systems. This integrated approach holds significant promise for improving the reliability of feature importance evaluations, leading to more trustworthy insights applicable to food science and chemistry fields.

Letter to the Editor:

Li et al. (2025) conducted a study, "Unraveling almonds deterioration using whole-cell biosensor coupled with machine learning approaches and SHAP interpretation," which presents intriguing findings warranting further discussion. Their work addresses critical research gaps in food quality assessment, specifically the need for rapid, non-destructive testing methods capable of capturing changes in chemical composition related to food deterioration (e.g., almonds) beyond traditional laboratory-based chemical analysis. Implicitly, their contribution also extends to the broader challenge of developing robust feature algorithms for complex chemical data, such as that derived from hyperspectral imaging. Their methodological contributions primarily include the development of a novel whole-cell biosensor array and its integration with machine learning (ML) for real-time monitoring and enhancing interpretability through SHAP. Their study aimed to develop a real-time method to monitor almond quality using a whole-cell biosensor array with various ML algorithms, including Linear Discriminant Analysis (LDA), Logistic Regression (LR), Partial Least Squares Discriminant Analysis (PLS-DA), Support Vector Machine (SVM), and Random Forest (RF). Reporting that Support Vector Machine (97.5 % accuracy) and Random Forest (100 % accuracy) outperformed linear models, they focused on interpreting the feature importance of the RF model using SHAP. Their analysis of the top 10 features highlighted pspA2, pspA1, rpoS2, and katG9 as particularly influential, with higher values of these correlating with undeteriorated almonds and lower values with deteriorated ones. However, relying primarily on RF and its SHAP interpretation to identify key features introduces potential methodological biases, representing a notable limitation.

Li et al. (2025) have presented a novel method for monitoring almond quality; however, their paper raises a critical concern regarding the interpretation of feature importances derived from ML models and SHAP analysis. Their study emphasized the high predictive accuracy of the RF model before proceeding to its SHAP interpretation. While their work contributes to transparency in understanding model predictions, it is crucial to acknowledge that high predictive accuracy does not inherently confirm the reliability of feature importance (Lipton, 2018; Musolf et al., 2022). While Li et al.'s work seeks to identify key factors for almond deterioration through feature importance, the reported high prediction accuracy of RF might inadvertently imply the reliability of the subsequent SHAP results. As over 300 previous studies have pointed out, achieving high prediction accuracy does not ensure that feature importance interpretations are trustworthy (Lenhof et al., 2024; Mandler & Weigand, 2024; Potharlanka & Bhat, 2024; Steiner & Kim, 2016; Wood et al., 2024). Additional discussion and pertinent literature are provided in the supplementary material.

Tree-based models like RF, while powerful for prediction, can exhibit biases in feature importance calculations. RF's internal structure and splitting logic can lead to skewed feature importance assessments, often favoring variables utilized early in the tree construction (Mohamed Huti et al., 2023; Salles et al., 2021; Ugirumurera et al., 2024). Although Li et al. (2025) employed the Mantel test to evaluate the association between volatile compounds and promoters, providing evidence for the biological relevance of specific sensor responses, this statistical analysis operates independently of the machine learning model's feature importance assessment. Consequently, the inherent biases of RF and its SHAP interpretation present a potential concern, as the model's learned relationships may not fully align with the biological correlations identified through the Mantel test, potentially leading to a skewed interpretation of key deterioration indicators.

Additionally, SHAP values, while intended to elucidate feature importance, are intrinsically tied to the model they interpret, potentially reflecting or amplifying the model's inherent biases (Bilodeau et al., 2024; Fisher et al., 2019; Huang & Marques-Silva, 2024; Kumar et al., 2021; Lones, 2024; Molnar et al., 2022). The function of 'explain = SHAP(model)' underscores this dependency. SHAP values can overestimate feature importance, especially for features with high variability

https://doi.org/10.1016/j.foodchem.2025.146171
Received 29 April 2025; Received in revised form 10 August 2025; Accepted 28 August 2025
Available online 29 August 2025

or many discrete categories—properties that inherently inflate SHAP magnitudes and may skew interpretation. As SHAP relies on the model's output for its explanations, its inherent vulnerability to model biases demands careful interpretation of the results and assessment of the analysis's reliability. While other feature analysis methods exist, such as Permutation Importance, LIME, and Integrated Gradients, offering varying computational costs and approaches to interpretability, they are also susceptible to model bias. Despite the existence of these simpler alternative methods, SHAP is widely utilized as one of the most consistent frameworks due to its comprehensive game-theoretic principles and versatility. However, it is important to recognize that all these methods, including SHAP, are susceptible to model bias, necessitating careful application to ensure reliable interpretations.

The challenge of validating feature importance is compounded by the lack of ground truth, as different models can yield varying rankings due to their distinct methodologies and inherent biases. In Li et al.'s study, this issue is amplified by the biosensor array's high dimensionality, comprising 72 bioluminescent intensity readings. This large number of variables, coupled with the potential for collinearity among these sensor units, can obscure the true relationships within the data. Consequently, relying solely on the Random Forest model and its SHAP interpretation makes it difficult to confidently pinpoint the actual determinants of almond deterioration, as the identified feature importances may be influenced by the model's specific biases and the complex interplay of the 72 sensor readings (Touw et al., 2013).

These factors not only complicate the isolation of individual sensor effects but also diminish the perceived importance of predictive sensors within the biosensor array. This complexity extends beyond theoretical considerations, manifesting in practical applications where researchers may find it challenging to identify the true determinants of almond deterioration, potentially leading to conclusions about underlying mechanisms that should be interpreted with caution. Furthermore, the high dimensionality of the 72 sensor readings increases the risk of overfitting, causing models to capture noise rather than genuine signals and to emphasize spurious sensors in importance measures. This is particularly problematic in biological and agricultural research, which inherently involves noisy and complex datasets, requiring robust models that can distinguish between meaningful signals and random fluctuations. Additionally, the complexity of the biosensor array renders importance measures highly sensitive to minor changes in data or model configurations, which may affect their stability and reliability. This instability can complicate the interpretation of model-derived insights and hinder the development of consistent and reproducible research findings regarding the key indicators of almond deterioration.

Addressing these limitations necessitates a robust analytical framework that considers data characteristics, inter-variable statistical dependencies, and rigorous validation. Effective modeling hinges on a thorough grasp of data distribution patterns. Probing intricate variable associations, particularly via non-parametric techniques, is paramount. Furthermore, confirming the statistical significance of findings through hypothesis testing and *p*-value analysis is vital to avoid misleading conclusions. Instead of relying exclusively on machine learning models and SHAP for identifying key features, we propose a synergistic approach that incorporates impartial, resilient statistical methods, such as Spearman's rho and Kendall's tau, complemented by p-value evaluation (Okoye & Hosseini, 2024; Yu & Hutson, 2024). These non-parametric tools are especially adept at characterizing monotonic relationships. For more complex dependencies, including non-monotonic collinearity and interactions, alternative non-parametric methods like Total correlation and Effective transfer entropy offer valuable insights (Caserini & Pagnottoni, 2022; Kerby et al., 2024; Tserkis et al., 2025; Umeki et al., 2025). To enhance interpretability, future research should adopt robust feature engineering strategies tailored to the biological nature of the data. Approaches such as feature agglomeration (FA) or highly variable gene selection (HVGS) offer biologically informed dimensionality reduction, helping to preserve meaningful variation while minimizing noise (Xie et al., 2025; Zhang et al., 2020). These methods can improve model stability and yield more reliable insights from high-dimensional biosensor data. Prioritizing these statistical principles will substantially bolster the credibility and dependability of feature importance assessments, particularly in critical domains like non-destructive food quality assessment, enabling more robust identification of deterioration indicators and guiding the development of advanced techniques such as hyperspectral imaging combined with reliable feature algorithms. To further explore uncertainty and variability in feature rankings, researchers might consider constructing confidence intervals from RF ensemble outputs, which can complement core statistical methods by offering valuable insights into potential overfitting, especially in high-dimensional contexts.

In conclusion, while machine learning techniques like Random Forest and SHAP provide powerful tools for prediction and feature interpretation, they are not without inherent biases. In complex biological and agricultural domains like almond quality assessment, a more robust understanding of the underlying factors requires the integration of unbiased statistical methods and rigorous validation. This combined approach, leveraging the strengths of both machine learning and statistical analysis, is essential for achieving more accurate, reliable, and interpretable insights into the key indicators of almond deterioration. Future research should focus on developing and applying such integrated methodologies to enhance our understanding of complex biological systems.

CRediT authorship contribution statement

Souichi Oka: Writing – original draft, Conceptualization. **Takuma Yamazaki:** Investigation. **Yoshiyasu Takefuji:** Writing – review & editing, Supervision, Project administration.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2025.146171.

Data availability

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

References

Bilodeau, B., Jaques, N., Koh, P. W., & Kim, B. (2024). *Impossibility theorems for feature attribution. Proceedings of the National Academy of Sciences, 121, article e2304406120.* https://doi.org/10.1073/pnas.2304406120

Caserini, N. A., & Pagnottoni, P. (2022). Effective transfer entropy to measure information flows in credit markets. *Statistical Methods & Applications, 31*, 729–757. https://doi.org/10.1007/s10260-021-00614-1

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, Article 177.* https://doi.org/10.48550/arXiv.1801.01489

Huang, X., & Marques-Silva, J. (2024). On the failings of Shapley values for explainability. *International Journal of Approximate Reasoning, 171, Article 109112.* https://doi.org/10.1016/j.ijar.2023.109112

Kerby, T., White, T., & Moon, K. R. (2024). *Learning local higher-order interactions with total correlation. Proceedings of the 2024 IEEE 34th international workshop on machine* *learning for signal processing (MLSP), 1–6*. https://doi.org/10.1109/MLSP58920.2024.10734758

Kumar, I., Scheidegger, C., Venkatasubramanian, S., & Friedler, S. (2021). Shapley residuals: Quantifying the limits of the Shapley value for explanations. *Advances in Neural Information Processing Systems, 34*, 26598–26608.

Lenhof, K., Eckhart, L., Rolli, L. M., & Lenhof, H. P. (2024). Trust me if you can: A survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. *Briefings in Bioinformatics, 25(5), Article bbae379*. https://doi.org/10.1093/bib/bbae379

Li, Q., Chen, S., Han, J., Li, B., Wu, L., & Li, J. (2025). Unraveling almonds deterioration using whole-cell biosensor coupled with machine learning approaches and SHAP interpretation. *Food Chemistry, 484, Article 144392*. https://doi.org/10.1016/j.foodchem.2025.144392

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

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

Mandler, H., & Weigand, B. (2024). A review and benchmark of feature importance methods for neural networks. *ACM Computing Surveys, 56, Article 318*. https://doi.org/10.1145/3679012

Mohamed Huti, T. L., Sawyer, E., & King, A. P. (2023). An investigation into race Bias in random Forest models based on breast DCE-MRI derived Radiomics features. In *, Vol. 14242. Clinical image based procedures. Fairness, AI in medical imaging, ethical and philosophical issues in medical imaging* (pp. 225–234). Springer. https://doi.org/10.1007/978-3-031-45249-9_22.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., … Bischl, B. (2022). General pitfalls of model-agnostic interpretation methods for machine learning models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K. R. Müller, & W. Samek (Eds.), *xxAI - beyond explainable AI (pp. 4). Springer*. https://doi.org/10.1007/978-3-031-04083-2_4

Musolf, A. M., Musolf, J., & Hopfensitz, M. (2022). What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. *Human Genetics, 141*(9), 1515–1528. https://doi.org/10.1007/s00439-021-02402-z

Okoye, K., & Hosseini, S. (2024). Correlation tests in R: Pearson Cor, Kendall's tau, and spearman's rho. In K. Okoye, & S. Hosseini (Eds.), *R programming: Statistical data analysis in research* (pp. 247–277). Springer Nature. https://doi.org/10.1007/978-981-97-3385-9_12.

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

Steiner, P. M., & Kim, Y. (2016). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of Causal Inference, 4, Article 20160009*. https://doi.org/10.1515/jci-2016-0009

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. F. T. (2013). Data mining in the life sciences with random forest: A walk in the park or lost in the jungle. *Briefings in Bioinformatics, 14*, 315–326. https://doi.org/10.1093/bib/bbs034

Tserkis, S., Assad, S. M., Lam, P. K., & Narang, P. (2025). Quantifying total correlations in quantum systems through the Pearson correlation coefficient. *Physics Letters A, 543, Article 130432*. https://doi.org/10.1016/j.physleta.2025.130432

Ugirumurera, J., Bensen, E. A., Severino, J., & Sanyal, J. (2024). Addressing bias in bagging and boosting regression models. *Scientific Reports, 14, Article 18452*. https://doi.org/10.1038/s41598-024-68907-5

Umeki, N., Kabashima, Y., & Sako, Y. (2025). Evaluation of information flows in the RAS-MAPK system using transfer entropy measurements. *eLife, 14, Article e104432*. https://doi.org/10.7554/eLife.104432

Wood, D., Papamarkou, T., Benatan, M., & Allmendinger, R. (2024). Model-agnostic variable importance for predictive uncertainty: An entropy-based approach. *Data Mining and Knowledge Discovery, 38*, 4184–4216. https://doi.org/10.1007/s10618-024-01070-

Xie, Y., Jing, Z., Pan, H., Xu, X., & Fang, Q. (2025). Redefining the high variable genes by optimized LOESS regression with positive ratio. *BMC Bioinformatics, 26, Article 104*. https://doi.org/10.1186/s12859-025-06112-5

Yu, H., & Hutson, A. D. (2024). A robust spearman correlation coefficient permutation test. *Communications in Statistics - Theory and Methods, 53*(6), 2141–2153. https://doi.org/10.1080/03610926.2022.2121144

Zhang, J., Wu, X., Hoi, S. C. H., & Zhu, J. (2020). Feature agglomeration networks for single stage face detection. *Neurocomputing, 380*, 180–189. https://doi.org/10.1016/j.neucom.2019.10.087

Souichi Oka[a,*,1], Takuma Yamazaki[a,2], Yoshiyasu Takefuji[b,3]
[a] *Science Park Corporation, 3-24-9 Iriya-Nishi Zama-shi, Kanagawa 252-0029, Japan*
[b] *Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan*

[*] Corresponding author.
E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka), tyamazaki@sciencepark.co.jp (T. Yamazaki), takefuji@keio.jp (Y. Takefuji).

[1] 0009-0000-4840-5232.

[2] 0009-0009-3178-3590.

[3] 0000-0002-1826-742x.