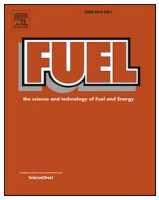
Fuel 424 (2026) 139411

Contents lists available at ScienceDirect

Fuel

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



Letter to the editor

Limits of SHAP feature importance in XGBoost–SOL Surrogates for catalytic Cracking: Correlation-Driven Bias, stability Diagnostics, and the Need for unsupervised validation

To the Editor.

Li et al. (2026) developed an XGBoost-based surrogate modeling framework integrated with a structure-oriented lumping (SOL) kinetic model to optimize direct crude-oil catalytic cracking processes. Their study reconstructed a molecular-level representation of crude oil and trained an XGBoost surrogate to predict product yields under varying operating and catalyst conditions. Strong predictive performance was reported, and SHapley Additive exPlanations (SHAP) were used to rank key operating variables and to support mechanistic interpretations regarding cracking severity and product distribution [1]. However, researchers are not familiar with algorithm-induced errors and underlying assumptions of machine learning tools whereas supervised models lack ground truth for accuracy validation in feature importance calculations, leading to erroneous conclusions. This paper raises urgent alarms regarding the use of supervised models for feature importance or selection assessments.

While the reported predictive accuracy was high, the interpretability claims warrant careful consideration. As with other supervised learning pipelines, the surrogate model achieved target-prediction accuracy, but feature-importance reliability lacks a ground-truth reference. Consequently, SHAP-based rankings may reflect how the model exploits statistical dependencies in the data rather than identifying stable physical or chemical determinants. This concern is particularly relevant in process systems where operating variables are correlated and may act as interchangeable proxies without affecting predictive performance [2–6].

These limitations become more evident when the mechanism by which SHAP generates explanations is examined. SHAP explanations are computed strictly with respect to a trained predictive model, as expressed by the functional relationship explain = SHAP(model). In this setting, SHAP quantifies how individual inputs contribute to the model's output rather than whether those inputs represent true mechanistic drivers. As a result, any proxy relationships or structural biases present in the trained surrogate are directly propagated into the SHAP attributions. In the presence of correlated variables, importance may be redistributed arbitrarily among interchangeable inputs, and small perturbations in data or model structure may lead to substantial shifts in ranking. Although SHAP satisfies theoretical properties such as local accuracy and consistency under specific assumptions, these guarantees apply only to the internal behavior of the model being explained and do not ensure physical or chemical validity [7–11].

To evaluate whether feature importance reflects true associations rather than algorithmic artifacts, two criteria are particularly important: consistency and dose–response behavior. Consistency requires that importance rankings remain coherent under perturbation, while dose–response relationships require systematic and proportional changes in predicted outcomes as feature values vary. Because the modeled relationship can be expressed as $y = f(x) = f(x_1, \ldots\ldots, x_n)$, verification of these properties should not rely solely on supervised model outputs. A practical diagnostic is a leave-top-1-out stress test, in which the highest-ranked variable is removed and the remaining features are re-ranked. Predictable, proportionate shifts support genuine associations, whereas large or erratic reordering suggests instability driven by correlation structure or model dependence [12–15].

To illustrate this, we applied a leave-top-1-out procedure on a publicly available concrete strength dataset [16] with no missing values and real-valued variables, using Strength as the target variable. The top eight features were first selected from the full dataset, after which the highest-ranked feature was removed and the top seven features were reselected from the reduced dataset. Feature ranking orders and cross-validation performance were compared across Random Forest, XGBoost, Feature Agglomeration (FA), Highly Variable Gene Selection (HVGS), and Spearman correlation, without scaling or transformation. The feature ranking stability and cross-validation results are summarized in Table 1.

The results show clear differences in feature importance stability. Feature Agglomeration and HVGS exhibit stable behavior, where removal of the highest-ranked feature leads to a monotonic shift in ranking, preserving the relative order of remaining features, with new features entering only at lower ranks. Spearman correlation shows relatively stable behavior with minor changes in ordering. In contrast, XGBoost demonstrates unstable behavior, where ranking order changes and features such as FineAggregate shift position or newly appear after removal of the top feature.

Cross-validation results support these observations. After removal of the highest-ranked feature, XGBoost shows a notable reduction in predictive performance, whereas HVGS and Feature Agglomeration maintain high performance. In contrast, Random Forest and Spearman also show substantial reductions after feature removal. This indicates that unsupervised methods retain predictive capability under perturbation, while supervised approaches depend more strongly on specific features.

In summary, while the proposed XGBoost–SOL surrogate framework demonstrates strong predictive utility for process optimization, caution is warranted when interpreting SHAP-derived feature importance as evidence of mechanistic significance. To address these limitations and strengthen interpretability, we recommend augmenting the pipeline with unsupervised feature-stability methods such as feature agglomeration (FA) and highly variable gene selection (HVGS), followed by non-targeted nonlinear nonparametric association tests such as Spearman correlation with p-values [17–19]. These approaches operate independently of labels, helping to avoid label-driven errors and reduce model-specific distortions. Combining these unsupervised and nonparametric methods with supervised modeling would provide a more reliable and

https://doi.org/10.1016/j.fuel.2026.139411
Received 15 January 2026; Received in revised form 4 April 2026; Accepted 6 April 2026
Available online 8 April 2026

Table 1
Cross-validation performance and feature ranking stability under leave-top-1-out perturbation.

Method	CV10 Accuracy	CV9 Accuracy	Top5 features of Top8	Top4 features of Top7
Random Forest	0.9072	0.3955	AgeInDays, Cement, Water, Slag, Superplast	Cement, Water, FineAgg, Slag
XGBoost	0.9318	0.4241	AgeInDays, Cement, Water, FlyAsh, Superplast	Cement, Water, FlyAsh, FineAgg
Feature Agglomeration (FA)	0.9068	0.9046	FineAgg, CoarseAgg, Cement, Slag, AgeInDays	CoarseAgg, Cement, Slag, AgeInDays
HVGS (Variance)	0.9067	0.8637	Cement, Slag, FineAgg, CoarseAgg, FlyAsh	Slag, FineAgg, CoarseAgg, FlyAsh
Spearman	0.9065	0.3954	AgeInDays, Cement, Superplast, Water, CoarseAgg	Cement, Superplast, Water, CoarseAgg

physically grounded understanding of which variables truly contribute to product-yield outcomes.

Statement of Financial Support.

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.

Data availability

The authors do not have permission to share data.

References

[1] Li Z, Qian X, Liu T, et al. Enhancing engineering strategies for direct catalytic cracking of crude oil via molecular-level kinetic reaction networks: a machine learning–enabled multi-objective optimization framework. Fuel 2026;413:138273. https://doi.org/10.1016/j.fuel.2026.138273.
[2] Wu L. A review of the transition from Shapley values and SHAP values to RGE. Statistics 2025;1:23. https://doi.org/10.1080/02331888.2025.2487853.
[3] Bilodeau B, Jaques N, Koh PW, Kim B. Impossibility theorems for feature attribution. PNAS 2024;121(2):e2304406120. https://doi.org/10.1073/pnas.2304406120.
[4] Huang X, Marques-Silva J. On the failings of Shapley values for explainability. Int J Approx Reason 2024;171:109112. https://doi.org/10.1016/j.ijar.2023.109112.
[5] Hooshyar D, Yang Y. Problems with SHAP and LIME in interpretable AI for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. IEEE Access 2024;12:137472–90. https://doi.org/10.1109/ACCESS.2024.3463948.
[6] Lones MA. Avoiding common machine learning pitfalls Patterns 2024;5(10):101046. https://doi.org/10.1016/j.patter.2024.101046.
[7] Molnar C, et al. General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W, eds. xxAI – Beyond Explainable AI. Lecture Notes in Computer Science. Vol 13200. Springer; 2022:4–17. doi:10.1007/978-3-031-04083-2_4.
[8] Kumar I, Scheidegger C, Venkatasubramanian S, Friedler S. Shapley residuals: quantifying the limits of the Shapley value for explanations. Adv Neural Inf Proces Syst 2021;34:26598–608. https://doi.org/10.48550/arXiv.2106.04242.
[9] Létoffé O, Huang X, Marques-Silva J. Towards trustable SHAP scores. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2025;39(17):18198–18208.
[10] Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin Transl Sci 2024;17(11):e70056. https://doi.org/10.1111/cts.70056.
[11] Coupland H, Scheidwasser N, Katsiferis A, et al. Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. BMC Public Health 2025;25(1):1520. https://doi.org/10.1186/s12889-025-22705-4.
[12] Sherkati A, Saffar Soflaei S, Darroudi S, et al. Association of serum levels and intakes of sodium and potassium with hypertension in the MASHAD cohort study population: a cross-sectional study. J Health Popul Nutr 2025;44(1):184. https://doi.org/10.1186/s41043-025-00919-x.
[13] Zhao Q, Xu J, Shi Z, et al. Genome-wide pleiotropy analysis reveals shared genetic associations between type 2 diabetes mellitus and subcortical brain volumes. Research (Wash DC) 2025;8:0688. https://doi.org/10.34133/research.0688.
[14] Schmidt RJ, Steeves M, Bayrak-Toydemir P, et al. ClinGen Low Penetrance/Risk Allele Working Group. Recommendations for risk allele evidence curation, classification, and reporting. Genet Med 2024;26(3):101036. https://doi.org/10.1016/j.gim.2023.101036.
[15] Yang L, Sadler MC, Altman RB. Genetic association studies using disease liabilities from deep neural networks. Am J Hum Genet 2025;112(3):675–92. https://doi.org/10.1016/j.ajhg.2025.01.019.
[16] Yeh IC. Modeling of strength of high-performance concrete using artificial neural networks. Cem Concr Res 1998;28(12):1797–808.
[17] Zhang J, Wu X, Hoi SCH, Zhu J. Feature agglomeration networks for single stage face detection. Neurocomputing 2020;380:180–9. https://doi.org/10.1016/j.neucom.2019.10.087.
[18] Xie Y, Jing Z, Pan H, et al. Redefining highly variable genes by optimized LOESS regression with positive ratio. BMC Bioinf 2025;26:104. https://doi.org/10.1186/s12859-025-06112-5.
[19] Eden SK, Li C, Shepherd BE. Nonparametric estimation of Spearman's rank correlation with bivariate survival data. Biometrics 2022;78(2):421–34. https://doi.org/10.1111/biom.13453.

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

[*] Corresponding author at: Suyam Ghale, Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan.
E-mail addresses: g2550002@stu.musashino-u.ac.jp (S. Ghale), souichi.oka@sciencepark.co.jp (S. Oka), takefuji@keio.jp (Y. Takefuji).