e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Volume 10, Issue 1, **90-99.**
https://e-journal.hamzanwadi.ac.id/index.php/edumatic
DOI: 10.29408/edumatic.v10i1.33509

Optimizing XGBoost Performance through Recursive Feature Elimination for Methanol Conversion Prediction

Ibnu Richo Kurniawan [1], Muhammad Akrom [1,*], Novianto Nur Hidayat [1], Muhammad Naufal [1]

[1] Universitas Dian Nuswantoro, Indonesia
[*] Corresponding author: Muhammad Akrom, Universitas Dian Nuswantoro, Indonesia
✉ m.akrom@dsn.dinus.ac.id

Received: 15 December 2025 | Revised: 12 January 2026 | Accepted: 1 March 2026 | Published: 15 March 2026

Abstract
The strong nonlinear interaction between catalytic properties and operating conditions complicates accurate space time yield modeling in thermocatalytic carbon dioxide hydrogenation, especially when redundant descriptors are included. Although XGBoost is widely used for predictive tasks, the influence of feature redundancy on generalization and interpretability in carbon dioxide to methanol systems remains insufficiently examined. This study investigates the integration of Recursive Feature Elimination with XGBoost using 639 experimental observations derived from copper based catalysts. Reducing the feature set from fifteen to eight variables improves generalization performance, as indicated by lower prediction error and higher explained variance. The retained variables correspond to key catalytic and operational parameters, including reaction temperature, pressure, and copper content, aligning with established kinetic and mechanistic principles. These results show that eliminating redundant descriptors stabilizes cross validated performance and reduces training complexity without sacrificing predictive accuracy. The reduced model concentrates predictive weight on kinetically relevant variables, providing a clearer quantitative representation of the parameters that govern space time yield in carbon dioxide hydrogenation.

Keywords: feature selection; methanol; recursive feature eliminiation (rfe); xgboost

To cite this article: Kurniawan, I. R., Akrom, M. F., Hidayat, N. N., & Naufal, M. (2026). Optimizing XGBoost Performance through Recursive Feature Elimination for Methanol Conversion Prediction. *Edumatic: Jurnal Pendidikan Informatika*, *10*(1), 90–99. https://doi.org/10.29408/edumatic.v10i1.33509

INTRODUCTION

The conversion of carbon dioxide (CO_2) into methanol via thermocatalytic hydrogenation has been widely recognized as a promising route toward sustainable chemical production and carbon utilization (Khalil et al., 2025). From a reactor engineering perspective, system performance is most effectively quantified using space-time yield (STY), which captures the efficiency of reactant conversion relative to catalyst mass and reactor volume. Beyond its experimental relevance, STY represents a scientifically meaningful target variable for reaction modeling because it integrates kinetic behavior, mass transfer effects, and catalyst utilization efficiency (Suvarna et al., 2022). However, STY optimization remains challenging not only due to the strong nonlinear coupling between operating conditions and catalytic properties, but also because data-driven STY models are particularly susceptible to overfitting arising from highly correlated experimental parameters. As a consequence, both experimental exploration

and predictive modeling of STY are constrained by complexity, redundancy, and limited interpretability (Aklilu & Bounahmidi, 2024).

Machine learning has therefore emerged as a powerful framework for modeling catalytic reaction systems characterized by multivariate and nonlinear interactions. Among existing algorithms, Extreme Gradient Boosting (XGBoost) has demonstrated strong predictive capability and robustness when applied to structured experimental datasets (Shi et al., 2024; Yao et al., 2022). XGBoost and related tree-based ensemble models have also been widely adopted in other domains due to their ability to handle heterogeneous and correlated variables (Lamens & Bajorath, 2025; Shaik et al., 2024). Previous studies have reported high prediction accuracy for STY in CO_2 hydrogenation using XGBoost (Suvarna et al., 2022). Nevertheless, most implementations emphasize numerical accuracy while treating XGBoost as a predictive black box, in which all available experimental descriptors are retained without systematic assessment of feature relevance or redundancy. As a result, model evaluation is often limited to performance metrics, with minimal discussion of feature interpretability, physical meaning, or generalization robustness (Schwaller et al., 2022).

Indiscriminate expansion of high dimensional feature spaces in scientific modeling may obscure dominant physicochemical relationships and artificially enhance apparent performance through noise driven fitting. The concern becomes particularly pronounced in catalytic datasets, where operating variables are often intrinsically correlated due to experimental design constraints. Previous studies report that feature reduction can enhance robustness, stability, and interpretability by suppressing redundant or weakly informative descriptors (Barbieri et al., 2024; Lamens & Bajorath, 2025; Lee et al., 2022). Beyond predictive accuracy, feature selection has also contributed to strengthening robustness and structured decision making in engineering contexts (Bernal et al., 2025). Nevertheless, its application in carbon dioxide hydrogenation modeling has rarely been positioned as a scientific analytical instrument. Existing investigations have not systematically evaluated whether reducing feature dimensionality can maintain predictive fidelity while simultaneously revealing the chemically dominant parameters governing STY. Such an omission constrains the progression of machine learning models from purely predictive tools toward frameworks capable of supporting mechanistic interpretation and transferable scientific understanding.

Robust data handling and evaluation strategies are therefore essential when modeling STY using experimental datasets. Preprocessing techniques such as feature standardization and outlier filtering have been shown to improve numerical stability and suppress noise-driven learning, particularly for non-normal experimental engineering data (Mallikharjuna Rao et al., 2023). Likewise, the choice of evaluation metrics plays a critical role in assessing model reliability. Root Mean Squared Error (RMSE) is sensitive to large prediction errors in reactor productivity indicators, while the coefficient of determination (R^2) quantifies the proportion of systematic experimental variance captured by the model (Zhu et al., 2022). RFE and related feature selection strategies have been extensively validated as effective tools for improving generalization performance and stabilizing learning in high-dimensional and nonlinear systems, as demonstrated across omics analysis, environmental modeling, geospatial classification, and neural network–based feature ranking frameworks (Barzani et al., 2024; Benjamin et al., 2023; Chen et al., 2025; Ding et al., 2024). When integrated with physically meaningful descriptors, feature selection can therefore function not only as a technical optimization step, but also as a scientific instrument for isolating dominant structure–performance relationships in catalytic reaction systems.

In this context, the present study systematically investigates the integration of RFE with the XGBoost algorithm for predicting STY in CO_2-to-methanol hydrogenation using experimental data derived from Cu-based catalysts. This work establishes a structured RFE–XGBoost modeling framework that mitigates feature redundancy and improves model

generalization, while simultaneously demonstrating that feature reduction enhances interpretability by concentrating learning capacity on chemically meaningful variables. In addition, the proposed approach provides a predictive tool that supports experimental design and data-driven optimization of catalytic reactors. Through this perspective, feature selection is positioned not merely as a technical preprocessing step, but as a scientific instrument for revealing dominant structure–performance relationships in catalytic reaction systems.

METHOD

This study adopts a computational quantitative experimental design, as illustrated in Figure 1, to evaluate the impact of feature redundancy on STY prediction in carbon dioxide hydrogenation to methanol. The framework consists of dataset preparation, preprocessing, dual modeling pathways with and without RFE, XGBoost modeling with hyperparameter optimization, and performance evaluation. This design isolates the effect of feature dimensionality on generalization, overfitting, and interpretability without introducing mechanistic assumptions.

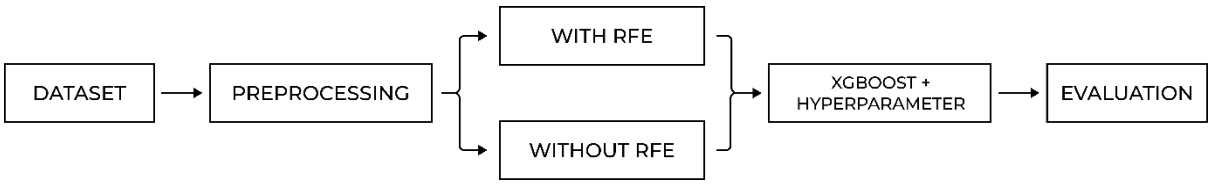

Figure 1. Methodological framework for feature selection and methanol conversion prediction modeling

The dataset was obtained from thermocatalytic CO_2 hydrogenation experiments over copper-based catalysts and includes only methanol-producing conditions. It consists of 639 samples with sixteen numerical variables, comprising one target variable (STY) and fifteen explanatory variables. The data cover representative operating conditions, including reaction temperature, pressure, gas composition, and reactor parameters. Although derived from a single experimental source, the dataset exhibits strong internal consistency and relevance to STY as a reactor performance indicator. This limitation is acknowledged; however, the dataset remains suitable for systematically evaluating feature reduction under controlled and chemically meaningful operating regimes.

Data preprocessing was conducted to ensure numerical stability and physical consistency. All features were standardized using z-score normalization. Outliers were identified using the Interquartile Range method and removed rather than modeled. Approximately a small fraction of samples (<5%) was excluded. These outliers corresponded to extreme STY values associated with atypical operating conditions and did not represent the dominant kinetic regime of CO_2 hydrogenation. Their removal reduced variance inflation and prevented noise-driven fitting that could distort feature ranking and generalization.

The dataset was split into training and testing subsets using an 80:20 ratio. To prevent data leakage, all preprocessing steps, including scaling and feature selection, were performed exclusively on the training data. The test set remained completely unseen and was used only for final performance evaluation.

Feature selection was performed using RFE with XGBoost as the base estimator. RFE is particularly suitable for catalytic systems due to correlated operating variables and physical redundancy among descriptors such as temperature, pressure, and space velocity. Recursive ranking was conducted within five-fold cross-validation, and the optimal number of features was determined by minimizing cross-validated RMSE. Feature selection stability was assessed

through consistency of selected variables across folds, yielding a reduced feature set of eight chemically meaningful variables.

Modeling was conducted using the Extreme Gradient Boosting algorithm with a squared error objective function. Hyperparameter tuning was performed after feature selection using grid search over defined ranges of n_estimators, max_depth, learning_rate, subsample, and colsample_bytree. Regularization and early stopping were applied to mitigate overfitting and ensure robust convergence (Lv et al., 2025).

Model performance was evaluated using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). RMSE captures sensitivity to large prediction errors in reactor productivity metrics, while R^2 quantifies the proportion of experimental variance explained by the model. Statistical significance of performance differences between baseline and RFE-based models was assessed using paired cross-validated t-tests (Shao et al., 2023).

RESULTS AND DISCUSSION

Results

The dataset used in this study consists of 639 observations obtained from thermocatalytic CO_2 hydrogenation experiments for methanol production. The dataset includes fifteen explanatory variables representing catalyst composition and operating conditions, along with one target variable in the form of STY. The wide variation in experimental parameters enables the identification of nonlinear response behavior, but also introduces variability associated with unstable or non-representative operating regimes. Therefore, data preprocessing was performed prior to modeling to improve numerical consistency and suppress noise that could distort learning behavior and feature attribution.

All variables were standardized using StandardScaler to ensure comparable scaling and prevent magnitude dominance during model training. In addition, potential outliers were detected using the Interquartile Range method to identify extreme experimental deviations. This preprocessing step was intended to stabilize variance structure and reduce the influence of anomalous observations on feature learning. By controlling both scale imbalance and extreme dispersion, the dataset becomes more suitable for robust modeling. The distributions before and after outlier treatment are presented in Figure 2 and Figure 3, respectively.

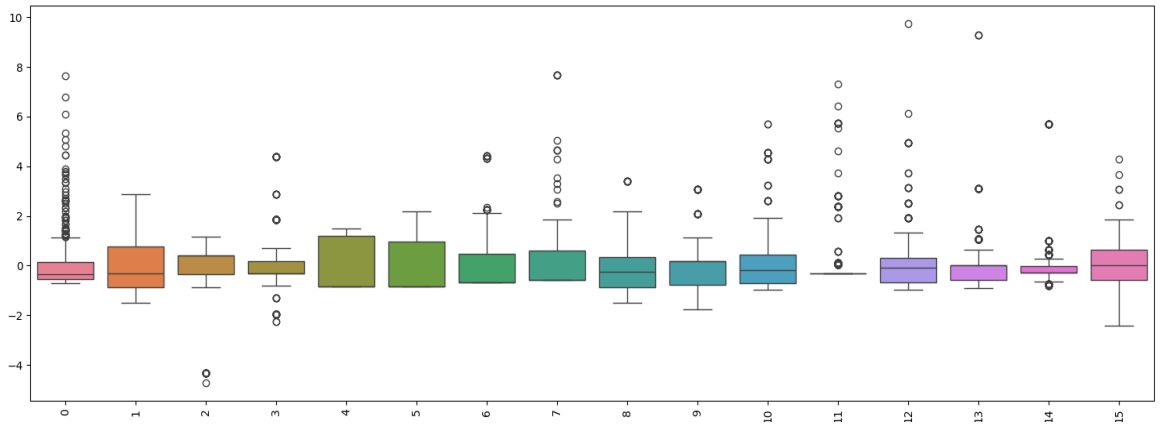

Figure 2. Data before outlier treatment

As observed in Figure 2, several variables exhibit pronounced tails and extreme values that deviate significantly from the main data cluster. These extreme observations indicate operating conditions outside the dominant experimental regime, potentially associated with unstable catalytic behavior or non-representative reactor states. After outlier removal, Figure 3 shows a clear reduction in variance dispersion, with tighter interquartile ranges across most features. Importantly, the overall shape and central tendency of the STY distribution are

preserved, demonstrating that data cleaning removes anomalous conditions without distorting the primary structure of the reaction performance data. This confirms that the retained dataset remains representative of the stable catalytic operating range governing methanol synthesis

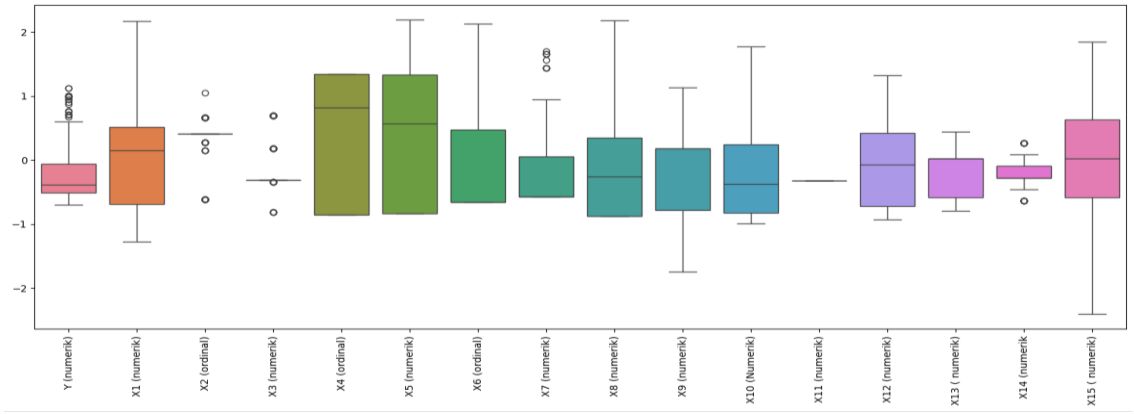

Figure 3. Data after outlier treatment

The predictive performance of the baseline XGBoost model using all fifteen input variables is summarized in Table 1. This configuration represents the full feature space without dimensional reduction. As such, it reflects the algorithm's predictive capacity under maximal informational input. The baseline therefore serves as a reference point for evaluating subsequent improvements in generalization and interpretability resulting from feature selection.

Table 1. Baseline model evaluation

Evaluation Metric	Value
RMSE	0.1183
R^2	0.9082

As detailed in Table 1, the baseline model attains an RMSE of 0.1183 and an R^2 of 0.9082, indicating that a substantial proportion of STY variability is captured even without feature selection. Within the context of reactor productivity modeling, this level of error remains moderate rather than trivial, as deviations of this magnitude may correspond to practically meaningful differences in predicted performance. When positioned against previously reported machine learning approaches for STY prediction, the obtained accuracy aligns with commonly observed results using full feature configurations. These findings therefore establish a quantitative reference for determining whether subsequent improvements originate from structurally meaningful feature reduction rather than statistical fluctuation.

The recursive feature elimination procedure yielded a reduced descriptor configuration that balances dimensional efficiency and predictive stability. Through iterative removal of weakly informative variables, the model converged toward a compact subset with consistent contribution across validation folds. This reduction minimizes redundancy within the experimental space. The refined feature set therefore represents a more structured formulation of the variables governing STY prediction.

As reflected in Table 2, reaction temperature, reaction pressure, gas hourly space velocity, calcination temperature, catalyst mass, copper content, and promoter related descriptors dominate the predictive structure. The exclusion of several initially included variables indicates the presence of physical and informational redundancy within the full multidimensional configuration. The concentration of importance around operating severity and active phase composition highlights the central kinetic and transport mechanisms

controlling STY formation. Collectively, these findings indicate that feature selection serves not only as a dimensional reduction technique but also as a systematic means of isolating the physically meaningful structure embedded in the experimental dataset.

Table 2. Selected features using rfe

Label	Feature Name	Importance
X1	Cu Content (%)	0.082988
X6	Type of Promoter 1	0.171853
X7	Promoter 1 Content	0.062099
X8	Calcination Temperature [K]	0.241223
X12	GHSV [cm3 h-1 gcat-1]	0.144334
X13	Catalyst Mass [g]	0.098015
X14	Reaction Pressure [MPa]	0.139217
X15	Reaction Temperature [K]	0.060270

Table 3. Hyperparameter tuning

Parameter Name	Value
colsample_byte	1
eta	0.1
eval_metric	mse
max_depth	4
n_estimators	500
subsample	0.7

The finalized hyperparameter configuration, as presented in Table 3, reflects a structurally balanced model characterized by moderate tree depth and controlled subsampling rates. Such settings constrain unnecessary model complexity while preserving sufficient flexibility to capture nonlinear interactions inherent in catalytic reaction data. The absence of extreme or overly aggressive parameter choices indicates that the observed performance enhancement is not driven by over parameterization. Instead, the improvement is more plausibly attributed to refined feature representation, which strengthens structural coherence and generalization capacity.

Table 4. RMSE and R^2 result for each fold

Fold	RMSE	R^2
1	0.0436	0.9902
2	0.0448	0.9873
3	0.0228	0.9966
4	0.0199	0.9977
5	0.0278	0.9959
Average	0.0318	0.9936

The cross-validation results summarized in Table 4 show consistently low RMSE values and high R^2 scores across folds, indicating strong internal consistency within the dataset. Because cross validation is conducted within the same experimental domain, the reported metrics primarily reflect in domain stability rather than external generalization. The slightly higher performance observed during cross validation compared with the independent test set is therefore methodologically expected and does not suggest information leakage. Collectively,

these findings confirm that the model exhibits stable learning behavior across data partitions sharing similar operating conditions.

The predictive results presented in Table 5 indicate that the reduced feature model achieves an RMSE of 0.1012 and an R^2 of 0.9328 on unseen data, outperforming the baseline configuration despite relying on fewer descriptors. This reduction in prediction error represents a meaningful enhancement in STY estimation accuracy and lowers uncertainty when forecasting reactor performance under new operating conditions. From a modeling standpoint, the improvement strengthens confidence in data driven reactor optimization by narrowing the predictive error margin while preserving generalization robustness.

Table 5. Evaluation of the xgboost model with rfe

Evaluation Metric	Value
RMSE	0.1012
R^2	0.9328

The empirical findings obtained in this study indicate that dimensional refinement via Recursive Feature Elimination enhances predictive stability and computational efficiency while maintaining accuracy. The consistency between the retained descriptors and chemically meaningful parameters suggests that the reduced configuration preserves the dominant structural relationships governing the catalytic system. This coherence reinforces the validity of the feature selection process as more than a technical adjustment, establishing the reduced model as a credible basis for subsequent scientific interpretation and mechanistic discussion.

Discussion

The present study demonstrates that data preprocessing, particularly outlier removal, plays a critical scientific role in stabilizing machine learning based modeling of catalytic reaction systems. While the Interquartile Range method effectively suppresses extreme deviations associated with anomalous experimental conditions, it is important to recognize that outliers are not universally uninformative. In catalytic systems operating near kinetic boundaries, such as diffusion limited regimes, catalyst deactivation onset, or thermodynamic equilibrium constraints, extreme observations may encode meaningful transition behaviour (Schwaller et al., 2022). Consequently, the current preprocessing strategy is most appropriate for modeling dominant and stable operating regimes rather than rare boundary condition phenomena. Under extreme kinetic conditions, this approach may underrepresent critical nonlinearities, indicating that alternative modeling strategies would be required for extrapolative or failure mode analysis.

Beyond preprocessing, the feature selection results highlight that RFE functions not merely as a dimensionality reduction technique, but as a scientific filtering mechanism that extracts the dominant structural relationships embedded in heterogeneous experimental data. The consistent retention of temperature, pressure, space velocity, copper content, and promoter related descriptors indicates that the STY response is governed by a limited subset of physically meaningful variables. At the same time, the systematic elimination of several frequently reported experimental descriptors suggests that such variables may be redundant when included without critical evaluation. This finding extends existing theory by demonstrating that not all experimentally accessible parameters contribute independent informational value in machine learning based catalytic modeling, particularly in systems characterized by strong parameter coupling (Barbieri et al., 2024).

Importantly, the improved performance of the reduced feature model supports the argument that excessive feature complexity can obscure, rather than enhance, scientific insight. In catalytic systems, over optimized models that rely on high dimensional input spaces may

achieve superficially high accuracy while embedding spurious correlations that mislead experimental decision making (Shaik et al., 2024). In contrast, models exhibiting slightly higher but stable prediction errors are often more valuable for experimental guidance, as they provide consistent and interpretable trends rather than fragile dataset specific fits. The observed improvement in test set performance following feature reduction reflects a more favorable balance between accuracy and generalization, which is essential for reliable reactor optimization (Suvarna et al., 2022).

Compared with prior machine learning investigations of carbon dioxide hydrogenation, the present study extends existing approaches by explicitly examining feature redundancy and interpretability rather than focusing solely on predictive accuracy. Earlier studies demonstrated the feasibility of machine learning for STY prediction using comprehensive descriptor sets (Suvarna et al., 2022). while broader assessments of catalytic carbon dioxide conversion modeling have primarily emphasized algorithm selection and performance optimization (Aklilu & Bounahmidi, 2024). Recent data driven reaction yield prediction studies similarly highlighted predictive capability without systematically evaluating dimensional redundancy or structural interpretability (Shi et al., 2024). In contrast, the integration of Recursive Feature Elimination in this study provides a structured mechanism to assess whether compact feature representations can preserve predictive fidelity while clarifying chemically dominant variables. Consistent with findings in other high dimensional domains, where feature elimination enhances robustness and interpretability by suppressing correlated or weakly informative descriptors (Barzani et al., 2024), the present results demonstrate that dimensional refinement strengthens both generalization and scientific coherence in catalytic modeling.

The outcomes of this analysis extend beyond predictive metrics and carry meaningful methodological and practical consequences. By isolating a compact subset of dominant variables, the effective dimensional scope of the experimental search space becomes more constrained and structurally focused. Such refinement enables researchers to prioritize critical operating parameters, including temperature, pressure, and space velocity, at the earliest stages of experimental planning, thereby reducing reliance on exhaustive parameter exploration. Concentrating investigative efforts on variables with the highest structural relevance to STY can decrease experimental costs, streamline catalyst screening, and enhance reactor optimization efficiency (Su et al., 2024).

Several limitations of the present study must be acknowledged. The dataset originates from a single experimental source and is restricted to copper-based catalysts, which limits the generalizability of the identified feature hierarchy to alternative catalytic systems such as zinc oxide zirconia or indium oxide-based formulations. This domain specificity highlights the need for future research integrating multiple experimental data sources and catalyst families to evaluate cross catalyst transferability and reduce dataset dependent bias. In addition, while RFE effectively identifies dominant variables, it does not explicitly quantify nonlinear feature interactions. Future work should therefore incorporate explainable machine learning techniques, such as SHAP values or partial dependence analysis, to map mathematical attributions onto mechanistic interpretations and further strengthen the linkage between data driven models and catalytic reaction theory (Bernal et al., 2025).

The findings presented here underscore the integrative role of feature selection in linking machine learning methodology with scientific reasoning in catalytic reaction engineering. Evidence that dimensional refinement strengthens both generalization and interpretability support the view that model simplification can enhance, rather than diminish, mechanistic transparency. Within this context, the proposed framework advances the development of predictive models that are not only robust but also structurally interpretable and experimentally actionable for sustainable chemical process optimization.

CONCLUSION

This study demonstrates that integrating RFE with the XGBoost algorithm provides a compact and scientifically coherent framework for predicting STY in carbon dioxide hydrogenation to methanol. The results show that feature reduction enhances model generalization and interpretability by isolating chemically and operationally dominant variables, thereby reinforcing the role of machine learning as a reliable scientific approach rather than a black box predictor. The strong alignment between selected features and established kinetic and catalytic principles confirms coherence between data driven inference and physicochemical understanding. By narrowing the variable space to structurally influential parameters, the modelling framework facilitates more focused experimental planning and more rational allocation of resources during catalyst evaluation and reactor development. Although the analysis is limited to a single data source and copper-based catalysts, these constraints define a clear direction for future work involving multi source datasets, broader catalyst systems, and explainable machine learning techniques to further strengthen the integration of data driven modelling with catalytic reaction theory.

REFERENCES

Aklilu, E. G., & Bounahmidi, T. (2024). Machine learning applications in catalytic hydrogenation of carbon dioxide to methanol: A comprehensive review. *International Journal of Hydrogen Energy*, *61*, 578–602. https://doi.org/10.1016/j.ijhydene.2024.02.309

Barbieri, M. C., Grisci, B. I., & Dorn, M. (2024). Analysis and comparison of feature selection methods towards performance and stability. *Expert Systems with Applications*, *249*, 123667. https://doi.org/10.1016/j.eswa.2024.123667

Barzani, A. R., Pahlavani, P., Ghorbanzadeh, O., Gholamnia, K., & Ghamisi, P. (2024). Evaluating the Impact of Recursive Feature Elimination on Machine Learning Models for Predicting Forest Fire-Prone Zones. *Fire*, *7*(12), 440. https://doi.org/10.3390/fire7120440

Benjamin, K. J. M., Katipalli, T., & Paquola, A. C. M. (2023). dRFEtools: Dynamic recursive feature elimination for omics. *Bioinformatics*, *39*(8), btad513. https://doi.org/10.1093/bioinformatics/btad513

Bernal, L., Rastelli, G., & Pinzi, L. (2025). Improving Machine Learning Classification Predictions through SHAP and Features Analysis Interpretation. *Journal of Chemical Information and Modeling*, *65*(21), 11716–11732. https://doi.org/10.1021/acs.jcim.5c02015

Chen, C., Liang, J., Sun, W., Yang, G., & Meng, X. (2025). An automatically recursive feature elimination method based on threshold decision in random forest classification. *Geo-Spatial Information Science*, *28*(4), 1494–1519. https://doi.org/10.1080/10095020.2024.2387457

Ding, X., Li, Y., & Chen, S. (2024). Maximum margin and global criterion based-recursive feature selection. *Neural Networks*, *169*, 597–606. https://doi.org/10.1016/j.neunet.2023.10.037

Khalil, M. T., Wu, X., Liu, S., Liu, Y., Ashraf, S., Shen, R., Zhang, H., Peng, Z., Jiang, J., & Li, B. (2025). Recent advancements in catalytic CO_2 conversion to methanol: Strategies, innovations, and future directions. *Green Chemistry*, *27*(30), 9016–9054. https://doi.org/10.1039/D5GC01906K

Lamens, A., & Bajorath, J. (2025). Contrastive explanations for machine learning predictions in chemistry. *Journal of Cheminformatics*, *17*(1), 143. https://doi.org/10.1186/s13321-

025-01100-6

Lee, Y., Cappellato, M., & Di Camillo, B. (2022). Machine learning–based feature selection to search stable microbial biomarkers: Application to inflammatory bowel disease. *GigaScience*, *12*, giad083. https://doi.org/10.1093/gigascience/giad083

Lv, B., Gong, H., Dong, B., Wang, Z., Guo, H., Wang, J., & Wu, J. (2025). An Explainable XGBoost Model for International Roughness Index Prediction and Key Factor Identification. *Applied Sciences*, *15*(4), 1893. https://doi.org/10.3390/app15041893

Mallikharjuna Rao, K., Saikrishna, G., & Supriya, K. (2023). Data preprocessing techniques: Emergence and selection towards machine learning models - a practical review using HPA dataset. *Multimedia Tools and Applications*, *82*(24), 37177–37196. https://doi.org/10.1007/s11042-023-15087-5

Schwaller, P., Vaucher, A. C., Laplaza, R., Bunne, C., Krause, A., Corminboeuf, C., & Laino, T. (2022). Machine intelligence for chemical reaction space. *WIREs Computational Molecular Science*, *12*(5), e1604. https://doi.org/10.1002/wcms.1604

Shaik, N. B., Jongkittinarukorn, K., & Bingi, K. (2024). XGBoost based enhanced predictive model for handling missing input parameters: A case study on gas turbine. *Case Studies in Chemical and Environmental Engineering*, *10*, 100775. https://doi.org/10.1016/j.cscee.2024.100775

Shao, X., Wang, H., Zhu, X., Xiong, F., Mu, T., & Zhang, Y. (2023). EFFECT: Explainable framework for meta-learning in automatic classification algorithm selection. *Information Sciences*, *622*, 211–234. https://doi.org/10.1016/j.ins.2022.11.144

Shi, R., Yu, G., Huo, X., & Yang, Y. (2024). Prediction of chemical reaction yields with large-scale multi-view pre-training. *Journal of Cheminformatics*, *16*(1), 22. https://doi.org/10.1186/s13321-024-00815-2

Su, Y., Wang, X., Ye, Y., Xie, Y., Xu, Y., Jiang, Y., & Wang, C. (2024). Automation and machine learning augmented by large language models in a catalysis study. *Chemical Science*, *15*(31), 12200–12233. https://doi.org/10.1039/D3SC07012C

Suvarna, M., Araújo, T. P., & Pérez-Ramírez, J. (2022). A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO2 hydrogenation. *Applied Catalysis B: Environmental*, *315*, 121530. https://doi.org/10.1016/j.apcatb.2022.121530

Yao, S., Kronenburg, A., Shamooni, A., Stein, O. T., & Zhang, W. (2022). Gradient boosted decision trees for combustion chemistry integration. *Applications in Energy and Combustion Science*, *11*, 100077. https://doi.org/10.1016/j.jaecs.2022.100077

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, *1*(2), 107–116. https://doi.org/10.1016/j.eehl.2022.06.001