Engineering Chemistry
ISSN: 2813-6535, Vol. 13, pp 1-6
doi:10.4028/p-IBRw9O

Submitted: 2024-11-25
Accepted: 2025-10-10
Online: 2026-02-18

Effect of Feature Selection on Data-Driven Prediction of Catalyst Performance: A Case Study on Methanol Formation from Thermocatalytic CO_2 Hydrogenation on Cu-Based Catalysts

Novianto Nur Hidayat[1,3,a*], Usman Sudibyo[1,b], Achmad Wahid Kurniawan[1,c] Fariz Hasim Arvianto[3,d], Muhammad Naufal[2,3,e*], Wahyu Aji Eko Prabowo[1,4,f], Harun Al Azies[1,3,g], Muhamad Akrom[1,h*]

[1]Research Center for Quantum Computing and Material Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

[2]Research Center for Intelligent Distributed Surveillance and Security, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

[3]Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

[4]Distance Learning Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

[a]novianto.hidayat@dsn.dinus.ac.id, [b]usman.sudibyo@dsn.dinus.ac.id, [c]wahid@dsn.dinus.ac.id, [d]fariz.arvin@gmail.com, [e]m.naufal@dsn.dinus.ac.id, [f]prabowo@dsn.dinus.ac.id, [g]harun.alazies@dsn.dinus.ac.id, [h]m.akrom@dsn.dinus.ac.id

Keyword: methanol conversion, CO_2 hydrogenation, Cu, machine learning, feature selection

Abstract. CO_2 conversion to methanol via thermocatalytic hydrogenation is one of the viable alternatives to address climate change problem while producing a valuable industrial product. However, this comes with a challenge, i.e., predicting the performance of catalytic systems. In this work, we present a data-driven study to predict the performance of Cu-based catalyst based on a compiled dataset consisting of 15 features obtained from experiment data. Furthermore, we implement feature selection techniques such as univariate, RFE, and XGBoost to investigate how the performance of the prediction model changes with varied number of features. The results show that features selected by RFE method yields the best performance with 7 number of features, capable of even outperforms the baseline model in terms of accuracy and feasibilty. This suggests that feature selection technique is relevant in terms of constructing a machine learning model for predicting methanol production via CO_2 thermocatalytic hydrogenation.

Introduction

Conversion of CO_2 gas into low-carbon fuels such as methanol (CH_3OH) is one potential solution to reduce the concentration of this gas in the atmosphere while also decreasing reliance on fossil fuels. Consequently, this process is a crucial aspect of the transition to clean energy and green economy [1,2]. One effective option for optimizing this conversion process is through the use of catalyst. Catalysts can direct and accelerate the rate of chemical reactions, therefore enhancing the production yield of CH_3OH [3,4]. Traditionally, research on the role of catalysts in this conversion has been conducted using experimental methods such as impregnation and coprecipitation [3] as well as computational approaches (density functional theory and molecular dynamics [4]). However, evaluating the performance of a catalyst, whether experimentally or computationally, often involves significant costs, time, and resources, especially when multiple catalyst candidates are considered. Although recent studies have discovered a variety of catalysts [5,6] relevant to CH_3OH production, challenges persist in identifying and optimizing the most suitable catalysts for specific reactions [7-9]. Traditional experimental methods for catalyst discovery often require long periods of time, high costs, and significant resource investment [7-13]. In this regards, machine learning (ML) offers a new option to accelerate catalyst design by analyzing large and complex datasets, thereby uncovering

patterns and relationships that may not be apparent through experimental approaches. Indeed, previous study by Suvarna et al. [2] reports how the utilization of ML techniques is capable of predicting the performance of CH_3OH production catalyst. These predictive capabilities can thus significantly reduce the time, cost, and resources needed for experimental validation.

Despite the advancements in ML-based catalyst design, one of the interesting challenges is feature selection [14]. In the context of ML, feature selection refers to the process of identifying the most relevant input variables (features) that contribute to the predictive accuracy of the model. For CO_2 conversion to CH_3OH, a variety of features can influence the CH_3OH yield [2], including catalyst composition, reaction conditions (temperature, pressure), support materials, surface area, and electronic properties. The sheer number of potential features introduces complexity, and not all features carry equal importance for predicting catalytic performance. Using irrelevant or redundant features can lead to overfitting, reduced model interpretability, and increased computational costs.

Feature selection techniques are thus vital for improving model performance and making the predictions more interpretable, which is crucial for scientific insights and practical applications. Effective feature selection can help identify the key parameters that truly drive the CO_2 conversion efficiency, facilitating the design of better catalysts and optimizing reaction conditions. In the context of CO_2 conversion to methanol, feature selection helps in pinpointing the most influential factors among a wide range of catalyst characteristics and reaction variables, enabling the development of a more efficient and tailored catalytic system.

In this study, we implement feature selection techniques to evaluate and select the most important features for predicting the efficiency of CO_2 conversion to methanol. We will employ several approaches such as filtering, wrapper, and embedded approach. By comparing the selected features across different techniques, we aim to uncover the relation of feature selection to the performance of the ML for predicting CO_2-to-CH_3OH conversion.

Methodology

Dataset

This study constructs a dataset of Cu-based catalysts CO_2 hydrogenation from the available literature [2]. The original dataset is then pre-processed by selecting data that correspond to the Cu-based catalyst only. We then perform normalization and remove any outliers from the constructed data. Lastly, we remove some features that consist of missing value and/or NaN. The final dataset consists of 15 features (Table 1)which describes the 3 aspects of the experiment environment. First, the properties of the catalyst are represented by 7 features (X1-X7), e.g. Cu content, type of supports, its molecular weights, the type of promoter, and the promoter content. Secondly, The synthesis conditions are represented by calcination temperature (X8) and calcination duration (X9). Lastly, the reaction conditions are represented by the surface area of the catalyst (X10), the ratio of H_2/CO_2 feed (X11), Gasly Hour Space Velocity (X12), the amount of the catalyst used in the experiment (X13), pressure (X14), and reaction temperature (X15). The performance of the catalyst, which is the target of this dataset, is the CH3OH space-time yield.

Table 1. The feature index and its corresponding feature

Index	Corresponding Feature	Index	Corresponding Feature	Index	Corresponding Feature
X1	Amount of Cu	X6	Type of promoter	X11	H2/CO2
X2	Type of support 1	X7	Promoter 1 loading	X12	GHSV
X3	MW of support 1	X8	Calcination temperature	X13	Catalyst amount
X4	Type of support 2	X9	Calcination duration	X14	Pressure
X5	MW of support 2	X10	SBET	X15	Temperature

Machine learning model

The model that is discussed in this study employs ML technique based on extreme gradient boost (XGB) method. The optimum parameters used for XGB is taken from the previous study [2]. We implement univariate, RFE, and the embedded XGB technique to represent each the filtering, wrapper, and embedded approaches for the feature selection. The selected feature amounts are established at 25%, 50%, and 75% of the initial features, corresponding to 3, 7, and 11 features, respectively. The workflow of this model is illustrated in **Figure 1**.

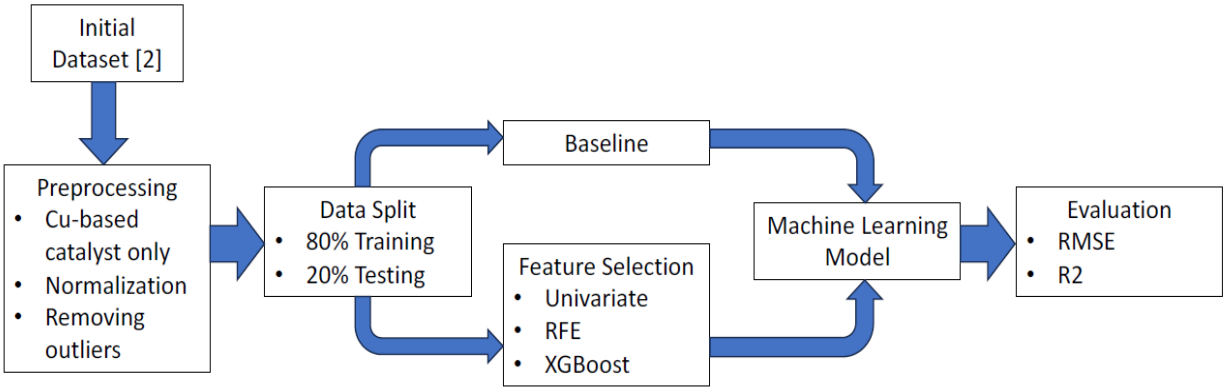

Fig. 1. The workflow that is used in this study

Model performance is assessed by utilizing reliable metrics such as the coefficient of determination (R^2) and root mean square error (RMSE). The feasibility of the model is indicated by the value of R^2 close to 1. RMSE is used to calculate the difference between actual and predicted values. As the RMSE value decreases, the prediction error also decreases. A reduced statistical error indicates a better predictive model; thus, the metric is used to evaluate the accuracy of the model.

Result and Discussion

We begin the discussion with evaluating the performance of the prediction model without any feature selection implemented so as to establish a baseline for our study. As shown in Table 2, the performance of the prediction model is already accurate as-is, with the RMSE value close to 0. The feasibility of the model is also considerably good, with R^2 value very close to 1.

Table 2. The baseline for the prediction model

Baseline	RMSE	R2
Without feature selection	0.108	0.931

Table 3. The performance of each model after treated with feature selection method

Feature selection technique	~25% of the initial features (3 features)		~50% of the initial features (7 features)		~75% of the initial features (11 features)	
	RMSE	R2	RMSE	R2	RMSE	R2
Univariate	0.198	0.766	0.139	0.883	0.120	0.914
RFE	0.199	0.763	0.090	0.952	0.111	0.926
XGBoost	0.196	0.770	0.105	0.934	0.130	0.898

Table 3 shows the performance of each model after treated with feature selection method. We observe that the performance of each of the model worsens when we reduce the number of features to 3. In this case, the model that utilizes features selected via XGBoost method prevails only slightly over the other two methods. After the number of features is increased to 7, it is apparent that the general performance of each of the model is improved. The model with XGBoost feature selection method shows a remarkable improvement that is roughly similar with our baseline. However, the model with features selected via RFE method especially yields the best performance as it even outperforms the baseline in terms of accuracy and feasibility. Interestingly, the performance of models treated with RFE and XGBoost seems to be slightly reduced as more features are added. In this case, the performance of the model with RFE feature selection is more or less on par with our baseline.

Additionally, we can see how the model treated with univariate performs better as more features are included. On the contrary, it is interesting to see how the performance of the model treated with each RFE and XGBoost is peaked when using 7 features. We are currently investigating further how the performance of the model changes with more varied number of features. Finally, this result suggests that utilizing feature selection can indeed give us a better performing model.

Table 4. List of the features selected for each method. The features that are not appeared in the 3 selection methods are listed inside the parentheses.

Method	Features Selected
25% of features (3 features) case	
Univariate	X8, X12, (X14)
RFE	X8, X12, (X13)
XGBoost	X8, (X14), X12
50% of features (7 features) case	
Univariate	X1, (X6), X8, X12, X13, X14, X15
RFE	X1, X8, (X9), X12, X13, X14, X15
XGBoost	X8, X14, X12, X13, (X9), X1, X15
75% of features (11 features) case	
Univariate	X1, X3, (X5), X6, (X7), X8, (X10), X12, X13, X14, X15
RFE	X1, X3, (X5), X6, X8, (X9), (X10), X12, X13, X14, X15
XGBoost	X8, X14, X12, X13, (X9), X1, X15, (X7), X3, (X4), X6

We then examine the selected features by each method to check whether it holds a physical meaning. The features selected by each method are shown in Table 4. In case of 3 selected features, we can see that the most common feature included by all methods are X8 and X12 which correspond to calcination temperature and the gas hourly space velocity, respectively. Previous study reports that gas hourly space velocity is indeed the most important feature to determine the space-time yield of CH_3OH [2]. In case of 7 selected features, we could see that our best performing model (RFE) utilizes features that are exactly the same as the ones selected by XGBoost method. Additionally, RFE also captures 3 most important features that affect CH_3OH yield according to [2], e.g X12 (gas hourly space velocity), X14 (reaction pressure), and X1 (metal (Cu) content). It should be noted that in our case, there are some selected features that are deemed to be the least important by the previous study [2] such as X8 (calcination temperature) and X9 (calcination duration).

Conclusion

In this work, we present a data-driven study to develop a model to predict the methanol yield via CO_2 thermocatalytic hydrogenation. We implement features selection model to investigate how the performance of the prediction model is affected by the number of features. The results show that feature selection is capable of affecting the accuracy and feasibility of the prediction model. The prediction model that utilizes features selected via RFE method shows the best performance and even outperforms the baseline with as little as 7 features. Another important finding is that RFE could capture 3 most important features that affect CH_3OH yield while suggesting new features. Further study to check how the performance of the prediction model varies with respect to the number of features is still ongoing.

Acknowledgements

This work is generously supported by Universitas Dian Nuswantoro through the Hibah Penelitian Dasar Pemula scheme (contract number: 076/A.38-04/UDN-09/VII/2024). N. N. H. especially expresses his deepest gratitude to Hideaki Kasai (Osaka University), Mohammad Kemal Agusta (Institut Teknologi Bandung), and Supriadi Rustad (Universitas Dian Nuswantoro) for the scientific discussions and encouragements.

References

[1] A. González-Garay et al., Plant-to-planet analysis of CO2-based methanol processes, Energy Environ. Sci. 12 (2019) 3425–3436, https://doi.org/10.1039/C9EE01673B.

[2] Manu Suvarna, et al., A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO2 hydrogenation, Appl. Catal. B: Environ. 315 (2022) 121530, https://doi.org/10.1016/j.apcatb.2022.121530.

[3] Andrés García-Trenco, Agustín Martínez, A simple and efficient approach to confine Cu/ZnO methanol synthesis catalysts in the ordered mesoporous SBA-15 silica, Catal. Today. Vol. 315 (2013) page 152-161, https://doi.org/10.1016/j.cattod.2013.03.005.

[4] Shyam Kattel et al., Active sites for CO_2 hydrogenation to methanol on Cu/ZnO catalysts. Science **355** (2017) 1296-1299, https://doi.org/10.1126/science.aal3573.

[5] X. Jiang, et al., Recent advances in carbon dioxide hydrogenation to methanol via heterogeneous catalysis, Chem. Rev. 120 (2020) 7984–8034, https://doi.org/10.1021/acs.chemrev.9b00723.

[6] D. Wu, et al., Understanding and application of strong metalsupport interactions in conversion of CO2 to methanol: a review, Energy Fuels 35 (2021) 19012–19023, https://doi.org/10.1021/acs.energyfuels.1c02440.

[7] X. Tang, et al., Effect of modifiers on the performance of Cu-ZnO-based catalysts for low-temperature methanol synthesis, J. Fuel Chem. Technol. 42 (2014) 704–709, https://doi.org/10.1016/S1872-5813(14)60031-1.

[8] A. Bansode, et al., Impact of K and Ba promoters on CO2 hydrogenation over Cu/Al2O3 catalysts at high pressure, Catal. Sci. Technol. 3 (2013) 767–778, https://doi.org/10.1039/C2CY20604H.

[9] A. Bansode, A. Urakawa, Towards full one-pass conversion of carbon dioxide to methanol and methanol-derived products, J. Catal. 309 (2014) 66–70, https://doi.org/10.1016/j.jcat.2013.09.005.

[10] T. Zou, et al., ZnO-promoted inverse ZrO2-Cu catalysts for CO2-based methanol synthesis under mild conditions, ACS Sustain, Chem. Eng. 10 (2021) 81–90, https://doi.org/10.1021/acssuschemeng.1c04751.

[11] M.S. Frei et al., Nanostructure of nickel-promoted indium oxide catalysts drives selectivity in CO2 hydrogenation, Nat. Commun. 12 (2021) 1960, https://doi.org/10.1038/s41467-021-22224-x.

[12] Z. Han, et al., Atomically dispersed Ptn+ species as highly active sites in Pt/In2O3 catalysts for methanol synthesis from CO2 hydrogenation, J. Catal. 394 (2021) 236–244, https://doi.org/10.1016/j.jcat.2020.06.018.

[13] B. Hu, et al., Hydrogen spillover enabled active Cu sites for methanol synthesis from CO2 hydrogenation over Pd doped CuZn catalysts, J. Catal. 359 (2018) 17–26, https://doi.org/10.1016/j.jcat.2017.12.029.

[14] J. Barrera-García et al., Feature Selection Problem and Metaheuristics: A Systematic Literature Review about Its Formulation, Evaluation and Applications. *Biomimetics* **2024**, *9*, 9. https://doi.org/10.3390/biomimetics9010009