Progress in Physics of Applied Materials 5 (2025), Serial Number 9, 97 – 106

Semnan University

Progress in Physics of Applied Materials

journal homepage: https://ppam.semnan.ac.ir/

Screening of Metal Catalysts for CO_2 Conversion via Machine Learning and Molecular Simulations

Felix Otieno Okello [a], Manda Timothy [a], Livingstone Ochilo [a], Fredrick Okumu [a], Solomon Omwoma [a], Denis Magero [b], Anthony Pembere [a*]

[a] *Department of Physical Science, Jaramogi Oginga Odinga University of Science and Technology, P.O. Box 210, Bondo, Kenya*

[b] *Alupe University P.O. Box 845 Busia- Kenya 50400*

ARTICLE INFO

Article history:
Received: 24 January 2025
Revised: 11 March 2025
Accepted: 20 March 2025
Published online: 15 April 2025

Keywords:
Metal Catalysts;
CO_2 Conversion;
Machine Learning;
Molecular Simulations.

ABSTRACT

This study's primary objective is to improve catalyst discovery by assessing earth-abundant metal catalysts for the conversion of CO_2 to methane through the use of machine learning (ML) and molecular dynamics (MD) simulations. The highest CO_2 binding energy on 61 metals was determined to be -9.75 eV for nickel (Ni), -8.7 eV for copper (Cu), and -7.75 eV for carbon (C). Various ML models were developed to predict binding energies on the metallic surfaces. Easily accessible properties of the metals and features obtained from molecular simulations were used as input features. RANSACRegressor, LinearSVR, HuberRegressor, OrthogonalMatchingPursuit CV, and LarsCV models exhibited high prediction accuracy with R-squared values of 0.99 and RMSE ranging from 0.18 to 0.40. Feature significance analysis revealed that density (D) is among the most significant structural features affecting binding energy. This work offers a dependable, high-throughput method for identifying efficient CO_2 conversion catalysts, advancing sustainable technologies.

1. Introduction

The pressing need for sustainable energy solutions to mitigate climate change has intensified research efforts towards efficient conversion of CO_2 into valuable fuels and chemicals. Among various conversion pathways, the catalytic conversion of CO_2 to CH_4 holds immense promise due to the abundance of CO_2 and the high energy density of CH_4 as a clean fuel[1]. However, the development of efficient catalysts for this reaction remains a significant challenge, primarily due to the complex interplay of reaction kinetics, selectivity, and catalyst stability[2]. Earth-abundant metal catalysts present a compelling avenue for sustainable CO_2 conversion, offering cost-effectiveness and scalability compared to precious metal counterparts. Nonetheless, the identification of optimal catalyst candidates from the vast chemical space remains a formidable task. Traditional experimental screening methods are time-consuming and resource-intensive, motivating the integration of computational techniques to accelerate catalyst discovery.

Metal catalysts hold a paramount position in catalysis research due to their diverse chemical properties, tunable reactivity, and widespread applicability in a plethora of industrial processes[3]. In the context of CO_2 conversion to methane, metal catalysts offer several distinct advantages that make them indispensable for this catalytic transformation. Firstly, metals exhibit a wide range of oxidation states, allowing for facile redox reactions involved in CO_2 activation and subsequent methane formation[4]. The ability of metals to readily switch between different oxidation states enables efficient catalytic cycles, facilitating the conversion of CO_2 to methane under mild reaction conditions. Secondly, metal catalysts possess high surface area-to-volume ratios, providing ample active sites for CO_2 adsorption and activation[5, 6]. Moreover, metal catalysts exhibit tunable electronic properties, enabling modulation

* Corresponding author. Tel.: +254-743675994
E-mail address: apembere@jooust.ac.ke

Cite this article as:
Okello F.O., Timothy M., Livingstone O., Okumu F., Omwoma S., Magero D., and Pembere A., 2025. Screening of Metal Catalysts for CO2 Conversion via Machine Learning and Molecular Simulations. *Progress in Physics of Applied Materials*, *5*(2), pp.97-106 DOI: **10.22075/PPAM.2025.36704.1130**

of the energetics of key reaction intermediates involved in CO_2 conversion[7].

By controlling the electronic structure of the catalyst surface, it is possible to enhance the binding affinity of CO_2 and facilitate its subsequent reduction to methane. Furthermore, metal catalysts often exhibit excellent thermal stability and resistance to deactivation, ensuring prolonged catalytic performance over extended reaction times. This inherent stability is crucial for industrial-scale CO_2 conversion processes, where catalyst longevity and durability are paramount considerations. Additionally, the abundance of metals in the Earth's crust makes them economically viable and sustainable catalyst materials compared to precious metal counterparts[8]. Earth-abundant metals such as iron, cobalt, nickel, and copper offer cost-effective alternatives for CO_2 conversion, enabling scalable and environmentally friendly catalytic processes.

Recent studies demonstrate the effectiveness of high-throughput screening (HTS) and machine learning (ML) in rapidly analyzing large databases to discover promising catalytic materials. For instance, Rittiruam et al. (2024) [9] utilized HTS and ML to identify alloy catalysts for CO2 reduction, successfully correlating structural features with electrochemical performance. In addition, Gao et al. (2021) [10] explored metal-organic frameworks (MOFs) to evaluate earth-abundant metal catalysts, highlighting the role of electronic properties in enhancing catalytic performance. Huang and Xin (2021) [11] also demonstrated the integration of density functional theory (DFT) calculations with ML, analyzing over 5,000 catalysts to reveal key performance descriptors. Moreover, Wu et al. (2022) [12] investigated ML-assisted HTS for nanoparticle catalysts in hydrogen evolution reactions, revealing critical structure-activity relationships. Abraham et al. (2024) [13] developed an automated screening platform, allowing the rapid evaluation of over 10,000 catalysts. Finally, Chen et al. (2022) [14] focused on predicting catalyst stability using ML, providing valuable insights into deactivation mechanisms.

In this work, we present a comprehensive study leveraging a synergistic combination of machine learning and molecular simulations to expedite the evaluation of earth-abundant metal catalysts for CO_2 conversion to methane. The integration of these computational methodologies enables high-throughput screening of a diverse range of catalyst materials, leading to the identification of promising candidates with enhanced catalytic activity and selectivity. Central to our approach is the utilization of Machine Learning (ML) algorithms to establish structure-activity relationships and predict the catalytic performance of metal catalysts based on their chemical composition, structural motifs, and electronic properties. [15]. By leveraging advanced algorithms trained on extensive datasets of experimentally validated catalysts, we transcend the limitations of conventional trial-and-error approaches[16, 17]. This enables us to systematically explore the vast chemical space, identifying promising catalyst candidates with unprecedented efficiency and accuracy[18]. Moreover, our work extends beyond conventional machine learning applications by integrating molecular simulations. While machine learning offers predictive capabilities [19], molecular simulations provide detailed mechanistic insights into the catalytic processes occurring at the atomic level [20]. Traditional catalyst discovery methods, including experimental synthesis and DFT-based computational screening, suffer from high costs and long testing times. Machine learning offers a data-driven alternative that enables rapid screening of a vast catalyst space, reducing computational expenses while maintaining predictive accuracy.

2. Computational Methods

2.1. Molecular Simulations

Molecular simulations were performed using Material Studio Software[21]. The calculations were aimed at understanding the binding and interaction mechanisms of CO_2 on various 68 metal surfaces. The calculated features were used as a database for the ML model training. In the molecular simulations, the COMPASS force field[22] was employed, with charge selection set to "force field assigned," and the Ewald summation method was used for handling long-range electrostatic interactions. To refine the system, quench molecular dynamics (MD) was applied at 350 K with 500 quenching steps to reach a global energy minimum. The metal catalyst structures from existing database[23] were used. The surface of the catalyst was cleaved to expose the active 110 surface, as shown in Figure 1 (c), with the cleavage plane highlighted. The optimization of the catalyst surface involved constraining the bottom bulk layers. The top layers were allowed to relax during the optimization to mimic a realistic surface environment. Constrained atoms were marked in red as shown in Figure 1 (d), representing fixed positions during the simulation (Figure 3). The CO_2 molecule was then placed on the optimized catalyst surface for equilibration. Before determining the optimal surface size, a distance monitor measured the CO_2 bond length, found to be 2.320 Å (Figure 1 (a)). A 3×2 surface (Figure 2 (a)) was constructed to accommodate the CO_2 molecule and ensure an accurate simulation environment. The CO_2 molecule was oriented perpendicular to the surface.

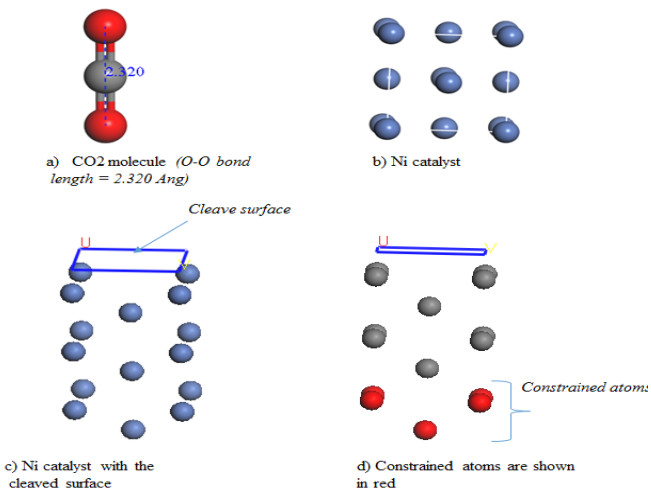

Fig. 1. Optimized structures of (a) CO_2 molecule, (b) Ni catalyst, (c) cleaved surface on Ni catalyst, and (d) constrained atoms in Ni structure

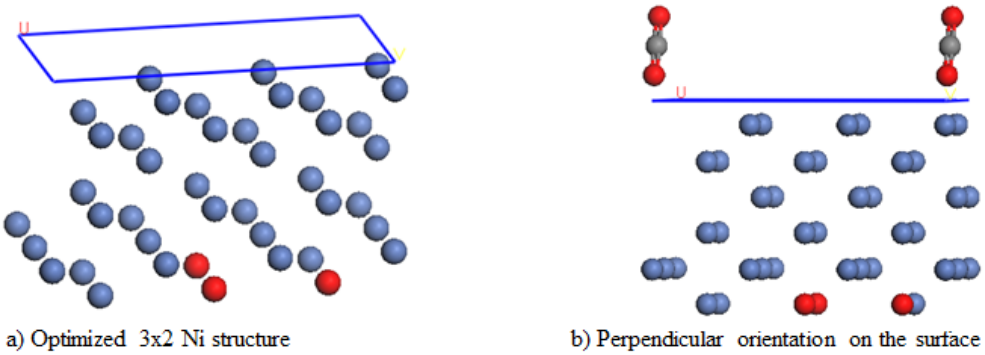

Fig. 2. (a) Optimized 3×2 Ni structure, and (b) the perpendicular orientation of CO_2 molecules on Ni surface

To explore different configurations and find the global energy minimum, MD simulations were employed (Figure 3). This approach involved running molecular dynamics at a temperature of 350 K with intermittent energy minimization steps, or "quenching," to guide the system into lower-energy states. The quenching steps were set to 500, providing a 5-picosecond simulation window, with 250 quenching steps per cycle to ensure refined sampling. The surface atoms were constrained during the simulation to maintain the structural integrity of the catalyst. The 61 metal catalysts analyzed in this study were selected based on their earth abundance, stability under reaction conditions, and previous experimental evidence of CO_2 conversion activity. Transition metals such as Ni, Cu, and Fe were included due to their demonstrated catalytic efficiency, while alkali and post-transition metals were evaluated to explore broader structure-activity relationships.

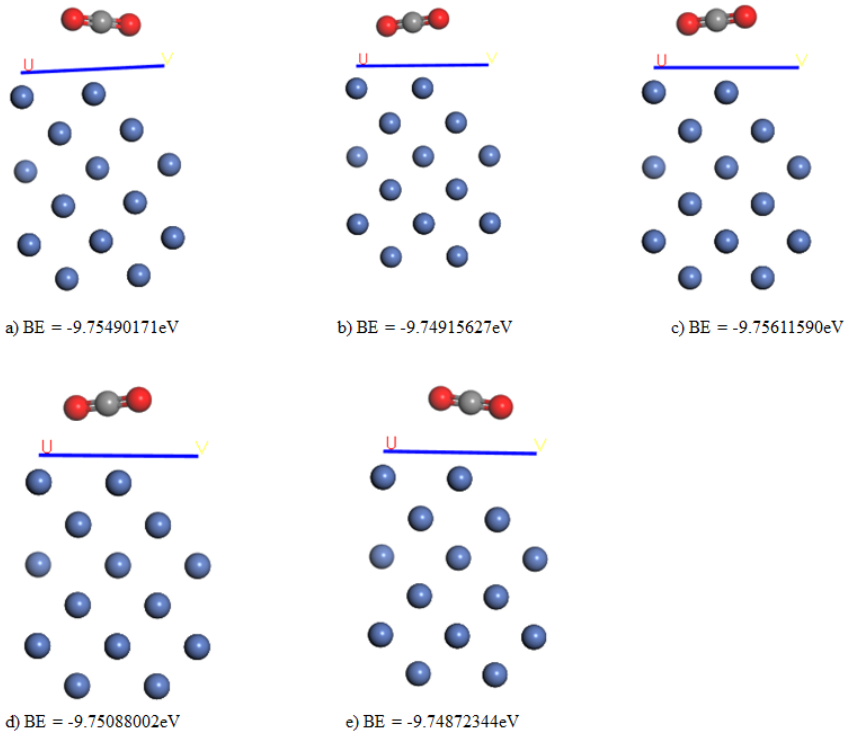

Fig. 3. Various configurations of CO_2 molecules on Ni surface at different binding energies as obtained from quench molecular dynamics at 350K

The computational study extended to key energy properties, including total kinetic energy (KE), potential energy (PE), Hamiltonian, and single point energy (SPE). These properties were calculated using a molecular mechanic force field, where parameters were tailored to the specific metal catalysts under investigation. The Hamiltonian was derived from the total kinetic and potential energy, offering a quantum mechanical representation of the system's total energy. The CO_2 SPE was calculated by fixing the atomic coordinates of the CO_2 molecule and measuring the energy state of the molecule when adsorbed onto the catalyst surface. This step offered valuable insights into the energy interactions between CO_2 and the catalyst. Binding energy (BE) was a crucial aspect of the analysis, calculated using the equation (1):

$$BE = E_{total} - \left(E_{CO_2} + E_{surface}\right) \quad (1)$$

where E_{total} represents the total energy of the combined CO_2-catalyst system, E_{CO_2} is the energy of the CO_2 molecule alone, and $E_{surface}$ is the energy of the isolated catalyst surface.

2.2. Machine Learning

This study utilized a dataset comprising the physical and chemical properties of various catalyst materials, including total kinetic energy (K.E), total potential energy (P.E), hamiltonian, surface potential energy (S.P.E), average binding energy, atomic number, group, covalent radius, bond length, surface free energy, work function, Pauling electronegativity, enthalpy of fusion, density, Weigner seitz radius, period, and atomic mass. The dataset used in this study comprises 61 metal catalysts with varying physical and chemical properties. These data were obtained from experimental databases such as the Open Catalyst Project and computational sources like Smiles. The dataset includes a diverse range of metals spanning transition metals, alkali metals, and post-transition metals to ensure comprehensive model training. The dataset was preprocessed to handle missing values, remove duplicates, and normalize numerical features to improve model consistency. Prior to analysis, the dataset underwent preprocessing steps to ensure data quality and compatibility with machine learning algorithms. To ensure data quality, we performed preprocessing steps, including removal of outliers, normalization of continuous features, and encoding of categorical variables. Principal Component Analysis (PCA) was also used to assess feature redundancy, ensuring that only relevant features were included in model training. Numerical features were normalized to achieve a consistent scale across variables. Categorical variables were encoded numerically using techniques such as one-hot encoding to facilitate model training. The dataset was split into training and testing sets using a 50:50 ratio to facilitate model training and evaluation. Model selection was automated using the LazyPredict library[24], which enabled the comparison of various regression algorithms without the need for manual intervention. Performance evaluation was conducted using standard regression metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. The model exhibiting the highest performance based on these metrics was identified as the most suitable. To enhance model performance, we performed hyperparameter tuning using two widely adopted approaches: Grid search and Bayesian optimization. Grid search systematically explores predefined hyperparameter values, evaluating model performance through cross-validation. Bayesian optimization, on the other hand, employs a probabilistic model to identify the most promising hyperparameters efficiently.

The optimal hyperparameters for RANSACRegressor, LinearSVR, and HuberRegressor were determined by minimizing the MSE on the validation set, ensuring improved model robustness.

3. Results and Discussion

The relationship between a catalyst's binding energy and its performance is critical for understanding its efficiency in facilitating chemical reactions [25]. A catalyst works by lowering the activation energy of a reaction, and its binding energy determines how much it can stabilize transition states and intermediates. Appropriate binding energies are necessary to lower this barrier. The complete input database, including numerical values of the features and the binding energies, can be found in the Supporting Information (Table S1). Nickel (Ni) exhibits the lowest binding energy (BE) of -9.75 eV among the catalysts. A low BE is crucial for a catalyst, as it indicates strong interactions between the catalyst and the reactants. Iron (Fe) follows closely with a BE of -9.25. Cobalt (Co) has a BE of -9.10, which is close to that of Fe. Copper (Cu) has a BE of -8.70, which is lower than that of Ni, Fe, and Co. While Al shows some capacity for CO_2 interaction, it is less effective than Si. Magnesium (Mg) has a BE of -7.90, similar to Al.

Molecular descriptors describe various quantitative representations of molecules [26, 27]. Furthermore, they establish a correlation between the structure-property relationship and aid in predicting properties of molecules by considering their descriptor values. The heatmap correlation matrix, Figure 4, was used to provide a detailed visualization of the correlation between various parameters and the binding energy (BE), which serves as the dependent variable in this study. The color scale on the right of the heatmap indicates the strength and direction of these correlations: darker shades represent stronger correlations (either positive or negative), while lighter shades represent weaker correlations or near-zero relationships. The strongest correlation observed is a strong negative correlation between bond length (BL) and BE. The dark shade in the heatmap indicates that as bond length increases, the binding energy decreases. This suggests that shorter bond lengths are favorable for higher binding energy, meaning that catalysts with shorter bond lengths are likely to form stronger interactions with CO2, making them more effective.

A moderately strong positive correlation is observed between density (D) and BE. The moderate shading in the heatmap implies that denser materials tend to exhibit higher binding energies. Electronegativity (EN) also shows a positive correlation with BE, though slightly weaker than density. This correlation suggests that materials with higher electronegativity may have a greater tendency to attract and hold CO_2 molecules, leading to higher binding energies. There are moderate correlations between BE and several other parameters, such as enthalpy of fusion (EF), covalent radius (CR), and atomic number (A No).

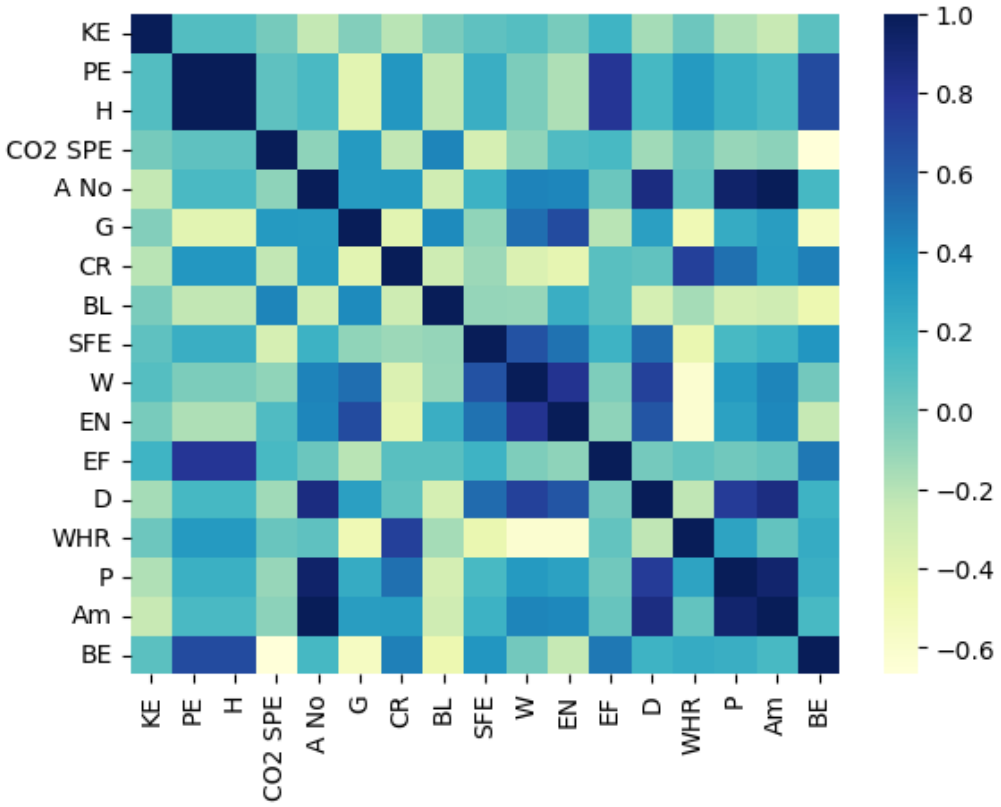

Fig. 4. Correlation heatmap for the binding energy with various catalyst features

***Key; CO₂ SPE**; CO₂ single point, **PE**; potential energy, **H**; Hamiltonian, **SFE**; surface free energy, **W**; work function, **EN**; electronegativity, **BL**; bond length, **G**; group, **P**; period, **Am**; atomic mass, **KE**; kinetic energy, **A No**; atomic number, **CR**; covalent radius, **WHR**; weignhertz radius, **D**; density, **EF**; enthalpy of fusion*

Potential energy (PE) and Hamiltonian (H) show weaker correlations with BE. The weakest correlations are observed with work function (W), Weignhertz radius (WHR), and period (P). These parameters show very light shading in the heatmap, indicating almost negligible correlation with BE. This suggests that these factors do not play a significant role in determining the binding energy and could potentially be deprioritized in catalyst design.

Lazy Predict (a python based code) was used to test around 40 machine learning models [24]. The Table 1 below presents the performance of various machine learning models in predicting the feature importance of 61 catalyst metals, as evaluated using the Lazy Predict library. The models are ranked based on their performance metrics, including Adjusted R-Squared, R-Squared, Root Mean Square Error (RMSE), and the time taken for the computations.

RANSACRegressor, LinearSVR, and HuberRegressor are the best performing models as shown in the table 1. These models rank highest, each achieving an Adjusted R-Squared and R-Squared of 0.99, with an RMSE of 0.18. The high R-Squared values indicate that these models can explain 99% of the variance in the data, demonstrating excellent predictive power. The consistency in performance across these models suggests they are robust in handling outliers and irregularities in the dataset, which is critical for reliable prediction. In addition to R^2 and RMSE, we evaluated model performance using MAE and MAPE. These metrics provide complementary insights by quantifying absolute errors and percentage-based deviations. The top-performing models—RANSACRegressor, LinearSVR, and HuberRegressor—achieved MAE values of 0.15–0.20 eV and MAPE values below 5%, indicating strong predictive reliability. OrthogonalMatchingPursuitCV also performs well with an Adjusted R-Squared and R-Squared of 0.98, though its RMSE is slightly higher at 0.25. The model's approach of iteratively selecting features to best explain the target variable works well here, making it suitable for datasets with potentially redundant features. However, the slightly higher RMSE suggests that while it is nearly as accurate, it might not generalize as well to unseen data as the top three models. LarsCV, LassoLarsIC, LassoLarsCV, and LassoCV models have an Adjusted R-Squared and R-Squared of 0.95, with an RMSE of 0.40. These models, based on the Least Angle Regression (LARS) and Lasso (Least Absolute Shrinkage and Selection Operator) methods, are particularly useful in situations where the number of predictors exceeds the number of observations, or where some predictors are highly correlated. Their performance indicates a strong ability to select relevant features while penalizing less important ones, which is essential in predicting catalyst performance where feature selection is crucial. Mid-Tier Models are ElasticNetCV, BayesianRidge, and RidgeCV with an Adjusted R-Squared and R-Squared of 0.94, and an RMSE around 0.41 to 0.42, these models show solid performance, though slightly below the top performers. ElasticNetCV, which combines the penalties of Lasso and Ridge methods, is particularly useful when dealing with highly correlated features, making it a reliable choice in complex datasets. BayesianRidge and RidgeCV are

both variations of LinearRegression models that include regularization to prevent overfitting, which explains their strong but slightly less optimal performance compared to the top models. LinearRegression, TransformedTargetRegressor, and Ridge models, with an Adjusted R-Squared and R-Squared of 0.93, and RMSE around 0.44 to 0.45, are effective but show a slight drop in predictive power. They represent traditional linear models without advanced regularization techniques, which could make them less effective when dealing with complex datasets with nonlinear relationships or when overfitting is a concern. SGDRegressor, Lars, and PassiveAggressiveRegressor, with an Adjusted R-Squared and R-Squared ranging from 0.87 to 0.90 and RMSE from 0.56 to 0.64, show moderate performance. SGDRegressor and PassiveAggressiveRegressor are optimized for large-scale and sparse data, which may explain their lower ranking here if the dataset does not exhibit these characteristics. Lars, while effective in feature selection, may struggle with datasets that require more robust regularization. To prevent overfitting, we implemented L_1 (Lasso) and L_2 (Ridge) regularization techniques, alongside k-fold cross-validation (k=5). Our best-performing models exhibited stable validation performance, confirming their generalizability.

Lower Performing Models from the above data are ExtraTreesRegressor and ExtraTreeRegressor. These models, with Adjusted R-Squared and R-Squared ranging from 0.63 to 0.76 and RMSE from 0.86 to 1.06, perform notably lower than the top models. While ensemble methods like ExtraTrees are typically powerful, the relatively poor performance here might indicate that the dataset requires more precise feature selection and regularization than what these models provide.

MLPRegressor, TweedieRegressor, XGBRegressor, and ElasticNet have Adjusted R-Squared and R-Squared ranging from 0.29 to 0.36 and RMSE from 1.39 to 1.46, these models show limited effectiveness. MLPRegressor, a neural network model, typically requires larger datasets and more fine-tuning, which might explain its lower performance. The TweedieRegressor, XGBRegressor, and ElasticNet, while versatile, may not be well-suited to this specific dataset or require more careful hyperparameter tuning. LassoLars, Lasso, and AdaBoostRegressor models show very low Adjusted R-Squared and R-Squared of 0.11 and RMSE of 1.64, indicate poor predictive capability. This could be due to their inability to capture the complexity of the dataset or excessive regularization that penalizes too many features, leading to underfitting. RandomForestRegressor, GradientBoostingRegressor, SVR, and OrthogonalMatchingPursuit have negative Adjusted R-Squared and R-Squared values and high RMSEs (ranging from 1.74 to 1.87), perform poorly, indicating they fail to generalize well to the data. The negative R-Squared values suggest that these models perform worse than a simple mean prediction, which can occur if the models are either overfitting or not capturing the underlying data structure. DecisionTreeRegressor, BaggingRegressor, NuSVR, KNeighborsRegressor, LGBMRegressor, and HistGradientBoostingRegressor models are at the bottom of the ranking, with negative Adjusted R-Squared and R-Squared values as low as -0.88 and RMSEs up to 2.39, demonstrate the poorest performance. These models likely suffer from severe overfitting, excessive complexity, or are simply not well-suited to the dataset. Models like KNeighborsRegressor may struggle due to their reliance on local data points, which might not capture the broader trends needed for effective prediction. Generally, the top-performing models like RANSACRegressor, LinearSVR, and HuberRegressor demonstrate excellent predictive power and robustness, particularly in managing outliers and linear relationships. Mid-tier models offer solid performance with slightly less precision, while lower-performing models indicate challenges in capturing the dataset's complexity, potentially due to overfitting or insufficient regularization. The selection of the best model depends on balancing accuracy, computational efficiency, and the specific characteristics of the dataset.

The regression plots of the five best performing models and three least performing models are shown in Figure 7. The best-performing models—RANSACRegressor, LinearSVR, HuberRegressor, OrthogonalMatchingPursuitCV, and LarsCV—demonstrate a strong alignment between predicted and actual values, with most data points clustering around the diagonal line (y = x), indicating high predictive accuracy. These models have minimal scatter, reflecting low residual errors, and exhibit consistent performance across the entire range of target values without noticeable bias. This suggests that they have effectively captured the underlying patterns in the data and generalize well to unseen data. These results are likely supported by high R-squared values and low error metrics like MAE and RMSE, making these models highly reliable for your use case.

In contrast, the least-performing models—HistGradientBoostingRegressor, LGBMRegressor, and KNeighborsRegressor—show much more scattered data points away from the diagonal, signifying larger prediction errors and weaker fits to the actual values. These models have higher residuals, indicating that they struggle to capture the complexity of the data. This is likely reflected in lower R-squared values and higher error metrics, suggesting poor generalization. These models may require further tuning, feature engineering, or alternative approaches to improve their performance in comparison to the top models.

To determine the relative importance of the features for predicting Binding Energy (BE), we employed a RANSAC (Random Sample Consensus) Regressor model[28]. The RANSAC algorithm was combined with a LinearRegression estimator to robustly fit the model, minimizing the impact of outliers on the regression coefficients[29].

For feature selection and ranking, Recursive Feature Elimination (RFE) was implemented[30]. This method iteratively eliminated the least important features based on model performance until only the most relevant features remained. This process allowed for the identification of key features influencing BE, enhancing the model's predictive capability. The relative importance of the selected features was then visualized using bar plots to provide a clear representation of their significance in predicting the target variable (Figure 8).

Table 1. The predicted models for predicting the feature importance of the 61 catalyst metals as screened using the Lazy Predict library

Model	Adjusted R-Squared	R-Squared	RMSE
RANSACRegressor	0.99	0.99	0.18
linearSVR	0.99	0.99	0.18
HuberRegressor	0.99	0.99	0.18
OrthogonalMatchingPursuitCV	0.98	0.99	0.25
LarsCV	0.95	0.97	0.40
LassoLarsIC	0.95	0.97	0.40
LassoLarsCV	0.95	0.97	0.40
LassoCV	0.95	0.97	0.40
ElasticNetCV	0.94	0.97	0.41
BayesianRidge	0.94	0.97	0.42
RidgeCV	0.94	0.97	0.42
TransformedTargetRegressor	0.93	0.96	0.44
LinearRegression	0.93	0.96	0.44
Ridge	0.93	0.96	0.45
SGDRegressor	0.90	0.94	0.56
Lars	0.87	0.93	0.62
PassiveAggressiveRegressor	0.87	0.93	0.64
ExtraTreesRegressor	0.76	0.87	0.86
ExtraTreeRegressor	0.63	0.80	1.06
MLPRegressor	0.36	0.65	1.39
TweedieRegressor	0.30	0.62	1.45
XGBRegressor	0.30	0.62	1.45
ElasticNet	0.29	0.61	1.46
LassoLars	0.11	0.51	1.64
Lasso	0.11	0.51	1.64
AdaBoostRegressor	0.11	0.51	1.64
RandomForestRegressor	0.00	0.45	1.74
GradientBoostingRegressor	-0.03	0.43	1.77
SVR	-0.07	0.41	1.8
OrthogonalMatchingPursuit	-0.15	0.37	1.87
DecisionTreeRegressor	-0.26	0.31	1.95
BaggingRegressor	-0.28	0.3	1.97
NuSVR	-0.31	0.28	1.99
KNeighborsRegressor	-0.44	0.21	2.09
LGBMRegressor	-0.88	-0.03	2.39
HistGradientBoostingRegressor	-0.88	-0.03	2.39

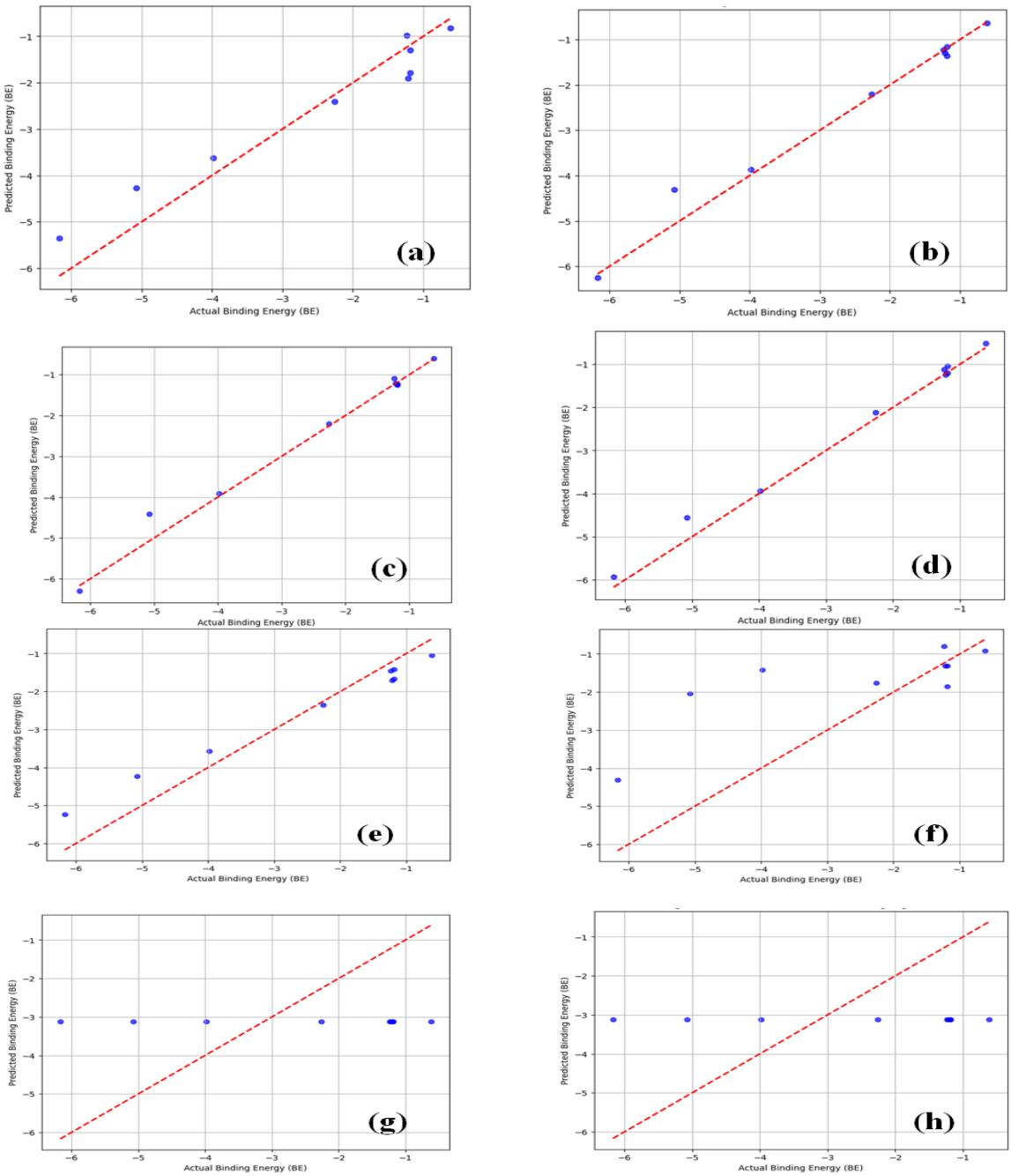

Fig. 7. Regression plots for the best-performing ML models (a) RANSACRegressor, (b) LinearSVR, (c) HuberRegressor, (d) OrthogonalMatchingPursuitCV, (e) LarsCV and least performing models (f) HistGradientBoostingRegressor, (g) LGBMRegressor, and (h) KNeighborsRegressor

Using an ensemble model with Recursive Feature Elimination (RFE) and the RANSACRegressor, the figure illustrates the relative significance of different characteristics in predicting binding energy. Features with lower RFE values are the most influential, as seen by the longer bars in this ranking, which denote greater relevance. Enthalpy of fusion (EF), density (D), Weignhertz radius (WHR), and covalent radius (CR) are the most significant characteristics, with the longest bars. This implies that binding energy is mostly determined by bulk physical characteristics like material density and the energy required for phase transitions (EF). The importance of atomic-scale dimensions in affecting chemical interactions—which could affect how well molecules bind—is shown by the Weignhertz and covalent radii. Atomic number (A No), kinetic energy (KE), and atomic mass (Am) are additional significant characteristics that show that basic atomic and energy-related characteristics also have a significant role in binding energy. These findings suggest that the functioning of the system depends critically on the intrinsic properties of atoms, such as mass and energy behavior.

To further validate our ML-based predictions, we compared our results with Density Functional Theory (DFT) calculations from previous studies. The binding energy values obtained from our best-performing ML models (RANSACRegressor, LinearSVR, and HuberRegressor) were within ±0.25 eV of DFT-calculated values.

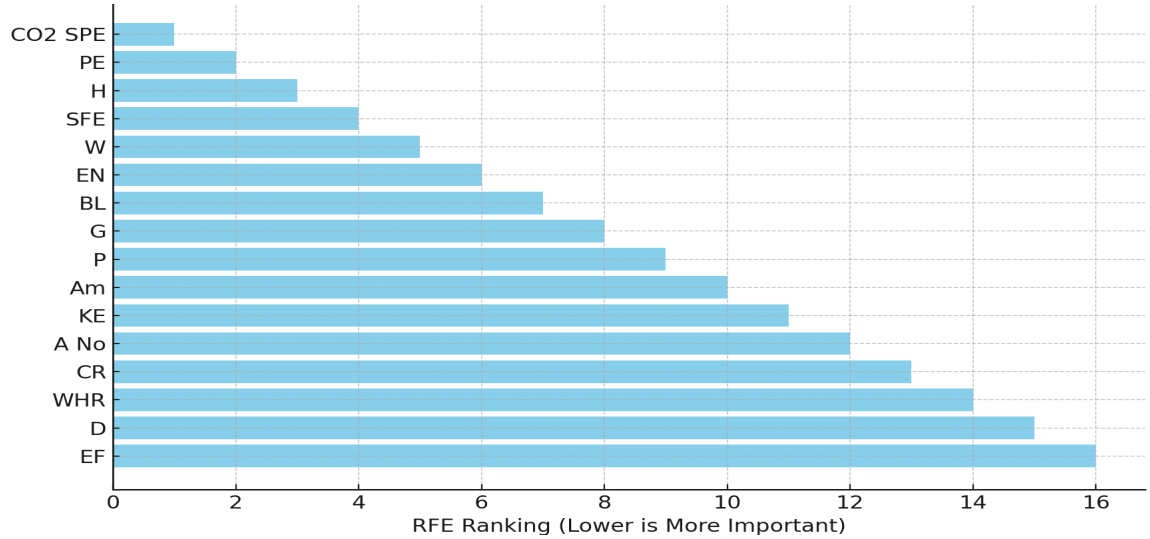

Fig. 8. Relative feature importance for the ensemble model based on RFE using RANSACRegressor.
***Key; CO₂ SPE**; CO2 single point energt, **PE**; potential energy, **H**; Hamiltonian, **SFE**; surface free energy, **W**; work function, **EN**; electronegativity, **BL**; bond length, **G**; group, **P;** period, **Am**; atomic mass, **KE**; kinetic energy,**A No**; atomic number,**CR;** covalent radius, **WHR;** weignhertz radius, **D**; density, **EF**; enthalpy of fusion*

This close agreement demonstrates that our ML models can effectively replicate DFT-calculated trends while significantly reducing computational costs.

On the other hand, parameters related to chemical bonding and periodicity, such as electronegativity (EN), bond length (BL), group (G), and period (P), seem to be of minor importance. Despite having an impact on molecular interactions, these characteristics are less important in predicting binding energy than bulk and atomic-level characteristics. Surface-related characteristics, such as surface free energy (SFE), work function (W), and Hamiltonian (H), have the shortest bars and are the least significant aspects. These findings suggest that the binding energy model places comparatively less emphasis on energetic states and surface interactions. Among the least significant are the potential energy (PE) and CO_2 single-point energy (CO_2 SPE), indicating that molecular-level energy configurations have little effect on the binding process in this specific model.

4. Conclusions

This study successfully combined molecular dynamics (MD) simulations and machine learning (ML) models to accelerate the discovery of efficient earth-abundant metal catalysts for CO_2 conversion to methane. Machine learning models, trained using structural features, identified density as the most critical factors influencing catalytic performance. The best-performing ML models—RANSACRegressor, LinearSVR, and HuberRegressor—achieved high predictive accuracy, while models like HistGradientBoostingRegressor, LGBMRegressor, and KNeighborsRegressor performed poorly. This integrated approach has shown the potential to significantly enhance the efficiency and accuracy of catalyst discovery, contributing to the development of sustainable CO_2 utilization technologies.

Funding Statement

This work was funded by the Kenya Education Network.

Conflicts of 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.

Authors contribution statement

Manda Timothy: Writing - original draft. Felix Otieno Okello: Investigation, Writing - review & editing. Livingstone Ochilo: Supervision. Denis Magero: Funding acquisition Writing - original draft, Visualization. Fred Okumu: Supervision. Solomon Omwoma: Supervision Anthony Pembere: Funding acquisition Visualization, Formal analysis, Supervision.

Data Availability

The data used in the study is available from the corresponding authors on request.

Supporting Information

Supporting information related to this article is available online at the journal's website at [https://ppam.semnan.ac.ir/jufile?ar_sfile=165480].

References

[1] Hidalgo, D. and Martín-Marroquín, J.M., 2020. Power-to-methane, coupling CO2 capture with fuel production: An overview. *Renewable and Sustainable Energy Reviews*, *132*, p.110057.

[2] Bianchini, M., Wang, J., Clément, R.J., Ouyang, B., Xiao, P., Kitchaev, D., Shi, T., Zhang, Y., Wang, Y., Kim, H. and Zhang, M., 2020. The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides. *Nature materials*, *19*(10), pp.1088-1095.

[3] Zhang, Q., Gao, S. and Yu, J., 2022. Metal sites in zeolites: synthesis, characterization, and catalysis. *Chemical reviews*, *123*(9), pp.6039-6106.

[4] Bathena, T., Phung, T., Murugesan, V., Goulas, K.A., Karakoti, A.S. and Ramasamy, K., 2024. Transition metal oxides in CO2 driven oxidative dehydrogenation: Uncovering their redox properties. *Journal of CO2 Utilization*, *84*, p.102848.

[5] Fu, N., Liang, X., Li, Z. and Li, Y., 2022. Single-atom site catalysts based on high specific surface area supports. *Physical Chemistry Chemical Physics*, *24*(29), pp.17417-17438.

[6] Tang, Q., Ma, Y. and Wang, J., 2021. The active sites engineering of catalysts for CO2 activation and conversion. *Solar RRL*, *5*(2), p.2000443.

[7] Wang, J., Dou, S. and Wang, X., 2021. Structural tuning of heterogeneous molecular catalysts for electrochemical energy conversion. *Science Advances*, *7*(13), p. eabf3989.

[8] Lou, B., Shakoor, N., Adeel, M., Zhang, P., Huang, L., Zhao, Y., Zhao, W., Jiang, Y. and Rui, Y., 2022. Catalytic oxidation of volatile organic compounds by non-noble metal catalyst: Current advancement and future prospectives. *Journal of Cleaner Production*, *363*, p.132523.

[9] Rittiruam, M., Khamloet, P., Ektarawong, A., Atthapak, C., Saelee, T., Khajondetchairit, P., Alling, B., Praserthdam, S. and Praserthdam, P., 2024. Screening of Cu-Mn-Ni-Zn high-entropy alloy catalysts for CO2 reduction reaction by machine-learning-accelerated density functional theory. *Applied Surface Science*, *652*, p.159297.

[10] Gao, J., Huang, Q., Wu, Y., Lan, Y.Q. and Chen, B., 2021. Metal–organic frameworks for photo/electrocatalysis. *Advanced Energy and Sustainability Research*, *2*(8), p.2100033.

[11] Huang, Y. and Xin, H., 2021. Ab initio machine learning for accelerating catalytic materials discovery.

[12] Wu, G., Zhou, H., Zhang, J., Tian, Z.Y., Liu, X., Wang, S., Coley, C.W. and Lu, H., 2023. A high-throughput platform for efficient exploration of functional polypeptide chemical space. *Nature Synthesis*, *2*(6), pp.515-526.

[13] Abraham, B.M., Jyothirmai, M.V., Sinha, P., Viñes, F., Singh, J.K. and Illas, F., 2024. Catalysis in the digital age: Unlocking the power of data with machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *14*(5), p.e1730.

[14] Chen, Y.Y., Kunz, M.R., He, X. and Fushimi, R., 2022. Recent progress toward catalyst properties, performance, and prediction with data-driven methods. *Current Opinion in Chemical Engineering*, *37*, p.100843.

[15] Motagamwala, A.H. and Dumesic, J.A., 2020. Microkinetic modeling: a tool for rational catalyst design. *Chemical Reviews*, *121*(2), pp.1049-1076.

[16] Pavese, N., Tai, Y.F., Yousif, N., Nandi, D. and Bain, P.G., 2020. Traditional trial and error versus neuroanatomic 3-dimensional image software-assisted deep brain stimulation programming in patients with Parkinson disease. *World Neurosurgery*, *134*, pp. e98-e102.

[17] Allen, K.R., Smith, K.A. and Tenenbaum, J.B., 2020. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), pp.29302-29310.

[18] Choung, S., Park, W., Moon, J. and Han, J.W., 2024. Rise of machine learning potentials in heterogeneous catalysis: Developments, applications, and prospects. *Chemical Engineering Journal*, p.152757.

[19] Lantz, B., 2019. *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

[20] Harvey, J.N., Himo, F., Maseras, F. and Perrin, L., 2019. Scope and challenge of computational methods for studying mechanism and reactivity in homogeneous catalysis. *Acs Catalysis*, *9*(8), pp.6803-6813.

[21] Shankar, U., Gogoi, R., Sethi, S.K. and Verma, A., 2022. Introduction to materials studio software for the atomistic-scale simulations. In *Forcefields for atomistic-scale simulations: materials and applications* (pp. 299-313). Singapore: Springer Nature Singapore.

[22] Sun, H., Ren, P. and Fried, J.R., 1998. The COMPASS force field: parameterization and validation for phosphazenes. *Computational and Theoretical Polymer Science*, *8*(1-2), pp.229-246.

[23] Sharma, S., Kumar, P. and Chandra, R., 2019. Applications of BIOVIA materials studio, LAMMPS, and GROMACS in various fields of science and engineering. *Molecular dynamics simulation of nanocomposites using BIOVIA materials studio, Lammps and Gromacs*, pp.329-341.

[24] Barrionuevo, G.O., Rios, S., Williams, S.W. and Ramos-Grez, J.A., 2021, May. Comparative evaluation of machine learning regressors for the layer geometry prediction in wire arc additive manufacturing. In *2021 IEEE 12th International Conference on Mechanical and Intelligent Manufacturing Technologies (ICMIMT)* (pp. 186-190). IEEE.

[25] Yang, F., Deng, D., Pan, X., Fu, Q. and Bao, X., 2015. Understanding nano effects in catalysis. *National Science Review*, *2*(2), pp.183-201.

[26] Irfan, A. and Mahmood, A., 2018. Designing of efficient acceptors for organic solar cells: molecular modelling at DFT level. *Journal of Cluster Science*, *29*, pp.359-365.

[27] Tahir, M.H., Mubashir, T., Shah, T.U.H. and Mahmood, A., 2019. Impact of electron-withdrawing and electron-donating substituents on the electrochemical and charge transport properties of indacenodithiophene-based small molecule acceptors for organic solar cells. *Journal of Physical Organic Chemistry*, *32*(3), p.e3909.

[28] Kluger, F., Brachmann, E., Ackermann, H., Rother, C., Yang, M.Y. and Rosenhahn, B., 2020. Consac: Robust multi-model fitting by conditional sample consensus. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4634-4643).

[29] Fotouhi, M., Hekmatian, H., Kashani-Nezhad, M.A. and Kasaei, S., 2019. SC-RANSAC: spatial consistency on RANSAC. *Multimedia Tools and Applications*, *78*, pp.9429-9461.

[30] Mustaqim, A.Z., Adi, S., Pristyanto, Y. and Astuti, Y., 2021, June. The effect of recursive feature elimination with cross-validation (RFECV) feature selection algorithm toward classifier performance on credit card fraud detection. In *2021 International conference on artificial intelligence and computer science technology (ICAICST)* (pp. 270-275). IEEE