Research Article

Prediction of Corrosion Inhibition Efficiency Based on Machine Learning for Pyrimidine Compounds: A Comparative Study of Linear and Non-linear Algorithms

Wise Herowati[1,2], Wahyu Aji Eko Prabowo[1,3]*, Muhamad Akrom[1,2], Totok Sutojo[1,2], Noor Ageng Setiyanto[1,2], Achmad Wahid Kurniawan[1,2], Novianto Nur Hidayat[1,2], Supriadi Rustad[1,4]

[1]Research Center for Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia
[3]Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia
[3]Distance Learning Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia
[4]DoctoralStudy Program in Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

Corresponding Author: Wahyu Aji Eko Prabowo; email: prabowo@dsn.dinus.ac.id

Published 7 March 2024

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the JICOMS Conference Committee.

Abstract.

The corrosion of materials poses a significant challenge in various industries, leading to substantial economic impacts. In this context, pyrimidine compounds emerge as promising, non-toxic, cost-effective, and versatile corrosion inhibitors. However, conventional methods for identifying such inhibitors are typically time-consuming, expensive, and labor-intensive. Addressing this challenge, our study leverages machine learning (ML) to predict pyrimidine compounds corrosion inhibition efficiency (CIE). Using a quantitative structure-property relationship (QSPR) model, we compared 14 linear and 12 non-linear ML algorithms to identify the most accurate predictor of CIE. The bagging regressor model demonstrated superior performance, achieving a root mean square error (RMSE) of 5.38, a mean square error (MSE) of 28.93, a mean absolute error (MAE) of 4.23, and a mean absolute percentage error (MAPE) of 0.05 in predicting the CIE values for pyrimidine compounds. This research marks a significant advancement in corrosion science, offering a novel and efficient ML-based approach as an alternative to traditional experimental methods. It shows that machine learning can quickly and accurately determine how well organic chemical inhibitors like pyrimidine stop material corrosion. This method gives the industry a new perspective and a workable solution to a problem that has existed for a long time.

Keywords: machine learning, corrosion inhibition, pyrimidine, QSPR model, predictive analysis

How to cite this article: Wise Herowati, Wahyu Aji Eko Prabowo*, Muhamad Akrom, Totok Sutojo, Noor Ageng Setiyanto, Achmad Wahid Kurniawan, Novianto Nur Hidayat, Supriadi Rustad, (2024), "Prediction of Corrosion Inhibition Efficiency Based on Machine Learning for Pyrimidine Compounds: A Comparative Study of Linear and Non-linear Algorithms" in *Jakarta International Conference on Multidisciplinary Studies towards Creative Industries*, KnE Engineering, pages 68–77. DOI 10.18502/keg.v6i1.15350

1. Introduction

The economic consequences of corrosion are significant, with annual costs for corrosion control estimated at approximately US$ 2.5 trillion [5]. Utilizing corrosion inhibitor technologies, which can decelerate corrosion rates in metals such as steel, iron, and aluminum, has decreased these expenses by as much as 35% [6]. Although Density Functional Theory (DFT) has the potential to assess corrosion inhibitors, conventional experimental techniques for corrosion treatment continue to be expensive and time-consuming [7-9].

The integration of various algorithmic models and artificial intelligence has been made easier by recent advancements in computing technology. This has led to improved machine learning processes, including classification, clustering, and model development. Machine learning (ML) techniques that utilize quantitative structure-activity relationships (QSAR) or quantitative structure-property relationships (QSPR) have shown promise in the field of corrosion inhibitors, both in terms of efficiency and effectiveness [8-11]. Prior studies have examined different algorithmic models, including Partial Least Squares (PLS), Multiple Linear Regression (MLR), Random Forest (RF), Autoregressive with Exogenous Inputs (ARX), Support Vector Machine (SVM), and similar models [12-15]. The Bagging Regressor, which is a non-linear model algorithm, has yielded satisfactory predictive results [16, 17]. The objective of this study is to assess and compare the effectiveness of linear and non-linear algorithm models in a machine learning-based quantitative structure-property relationship (QSPR) approach. The main objective is to forecast the inhibition efficiency (IE %) of corrosion inhibitors.

The primary objective of this research is to create a QSPR/QSAR model that explains the correlation between the chemical compositions of inhibitors and their ability to prevent corrosion. This method can assess untried substances and aid in creating new inhibitors with specific characteristics. Nevertheless, the existing literature needs to highlight the complete potential of machine learning models in this particular context. This study aims to address these discrepancies by conducting a thorough and evaluative examination of existing methodologies and their constraints. Our objective is to provide a fresh viewpoint and contribute to the field by showcasing the effectiveness of integrating linear and non-linear algorithms in predicting the effectiveness of corrosion inhibitors. This approach is characterized by its innovation and ability to address the limitations of prior research. It offers a more efficient and cost-effective approach to corrosion management. The corrosion of materials presents a significant obstacle in diverse industries, resulting in substantial economic consequences. Pyrimidine compounds are emerging as corrosion inhibitors that show promise due to their non-toxicity, cost-effectiveness, and versatility [12]. Nevertheless, conventional approaches to detect these inhibitors are time-consuming, costly, and require significant manual effort.

To tackle this challenge, our study utilizes ML to predict pyrimidine compounds' corrosion inhibition efficiency (CIE). Using a QSPR model, we evaluated 14 linear and 12 non-linear machine learning algorithms to determine the most precise predictor of CIE. This research represents notable progress in corrosion science, providing a new and effective machine learning-based technique as a substitute for conventional experimental methods. Furthermore, it can be possible to evaluate new compounds that have not yet been synthesized or tested. This technique can be used in designing new corrosion inhibitors with some desirable traits.

2. Material and Methods

2.1. Materials

In this study, we utilized data from Alamri et al. [12], comprising 54 pyrimidine structures characterized by 14 quantum chemistry descriptors, as outlined in Table 1. These descriptors include parameters like Energy of HOMO and LUMO, hardness, electron sharing fraction, dipole moment, ionization potential, softness, electron affinity, absolute electronegativity, electrophilicity, partition coefficient logarithm, molecular mass, and molecular volume. Identifying these descriptors is the foundational stage in developing our machine learning model, which aims to evaluate the efficacy of pyrimidine as an anti-corrosion agent. Our model's target variable for prediction is the Inhibition Efficiency (IE).

To predict IE, we divided our algorithm models into linear and non-linear categories [18] as presented in Tables 2 and 3. Linear models, such as linear regression, Bayesian Ridge, and SGD regression, best suit scenarios where the relationship between variables is proportional and linear. These models are relatively straightforward and offer easy interpretability. They are ideal for more superficial relationships where the output changes constantly as the input changes.

On the other hand, non-linear models, including the Gradient Boosting Regressor, Adaboost Regressor, and XGB Regressor, are designed to handle more complex and flexible relationships. They can identify and learn patterns that linear models may not adequately capture. Non-linear models are beneficial in scenarios where the relationship

TABLE 1: Quantum Chemistry Descriptors.

No.	Descriptors	Parameter
1	E_{HOMO} (eV)	Energy of HOMO
2	E_{LUMO} (eV)	Energy of LUMO
3	E_{L-H} (eV)	Energy gap
4	μ (D)	The dipole moment
5	IP (eV)	Ionization Potential
6	EA (eV)	The electron affinity
7	χ (eV)	The absolute electronegativity
8	η (eV)	The hardness
9	σ (eV^{-1})	The softness
10	ΔN	The fraction of electrons shared
11	ω (eV)	Electrophilicity
12	Log P	The logarithm of the partition coefficient
13	M (g.mol–1)	The molecular mass
14	V_m (cm^3/mol)	The molecular volume

TABLE 2: Linear Models.

No.	Algorithm Models
1	Linear Regression
2	ARD Regression
3	Bayesian Ridge
4	Elastic Net
5	Gamma Regressor
6	Huber Regressor
7	Orthogonal Matching Pursuit
8	Passive Aggressive Regressor
9	Poisson Regressor
10	Ransac Regressor
11	Ridge
12	SGD Regressor
13	Theilsen Regressor
14	Tweedie Regressor

between variables is more intricate and does not follow a straightforward proportional increase or decrease.

The next step involves applying and comparing these models to determine their effectiveness in predicting the IE of pyrimidine compounds. We will train each model using the identified quantum chemistry descriptors as input variables and IE as the target variable. The performance of each model will be evaluated based on standard metrics such as accuracy, mean square error, and others to identify the most effective

TABLE 3: Non-linear Models.

No.	Algorithm Models
1	Adaboost Regressor
2	Bagging Regressor
3	Gradient Boosting Regressor
4	Random Forest Regressor
5	PLS Regression
6	Decision Tree Regressor
7	Extra Tree Regressor
8	Dummy Regressor
9	Gaussian Process Regressor
10	Kernel Ridge
11	KNeighbors Regressor
12	XGB Regressor

model for predicting corrosion inhibition efficiency. This step is crucial, as it will provide insights into which models are best suited for predicting the effectiveness of pyrimidine compounds as corrosion inhibitors and, hence, contribute to the optimization of corrosion inhibition strategies in various industrial applications.

2.2. Methods

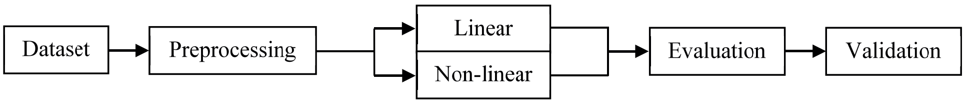

Figure 1: Flowchart of the Machine Learning Model Development Process.

After the preprocessing phase, algorithmic models were applied to the pre-processed data. These models included both linear models (listed in Table 2) and non-linear models (seen in Table 3). In order to reduce the impact of certain features on the models, we applied the MinMaxScaler normalization technique [19]. In order to mitigate potential biases and variance problems, we employed the K-Fold cross-validation technique, thereby bolstering the reliability of our model evaluation. The model's performance was evaluated using a set of metrics including MSE, RMSE, MAE, MAPE, and the R-Square [2-9], which that metrics are quantify the model's predictions accuracy. A lower value for these metrics indicates a higher level of precision in the model [20, 21]. The R-Square value, which ranges from zero to one, measures the degree to which the model accurately fits the observed data. A value approaching one indicates a model that precisely captures the variability in the dataset.

DOI 10.18502/keg.v6i1.15350

MSE = $\frac{1}{n}\left(\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2\right)$ (1)

RMSE = $\sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2\right)}$ (2)

MAE = $\frac{\sum_{i=1}^{n}}{y_i{}'-y_i n}$ (3)

MAPE = $\frac{1}{n}\sum_{1}^{n}\frac{\left(y_i-y_i{}'\right)}{y_i}$ (4)

R-Square = $\frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2\right]\left[n\sum y^2-\left(\sum y\right)^2\right]}}$ (5)

Where *n* represents the number of observations or samples, *x* and *y* are variables representing the independent and dependent data features, respectively. Y_i is the observation value, $\overset{b}{Y}_i$ is the predictive value.

The following steps include training the models using the prepared dataset, followed by testing and validation to ensure the reliability of the predictions. The final model selection will be based on comparative analysis, weighing the performance metrics to determine the most effective algorithm for predicting the IE of pyrimidine compounds in corrosion inhibition.

3. Result and Discussion

Linear models revealed that the gamma regressor yielded the lowest MSE at 33.91 and an RMSE of 5.82, alongside MAE and MAPE values of 4.47 and 0.05, respectively (Table 4). The bagging regressor algorithm demonstrated the best performance in non-linear models, corresponding to MSE and RMSE values of 28.93 and 5.38 and MAE and MAPE of 4.23 and 0.05 (Table 5). Compared to previous studies, such as the one by Alamri et al. [12], which reported an MSE of 64.64 for the partial least square regression (PLS) algorithm, our results indicate a substantial improvement in model performance. The findings highlight the bagging regressor model as a notably effective tool in predicting corrosion inhibition, marking progress in the field, and underscoring the potential of machine learning for enhancing corrosion control strategies.

The bagging regressor model yielded the most accurate results, both linear and non-linear, compared to the other tested algorithms. As illustrated in Figure 2, the training and testing data prediction line closely agrees with the actual data points, as evidenced by the R-Square value of 0.12.

4. Conclusion

Table 4: Linear Model Analysis Results.

Model	MSE	RMSE	MAE	MAPE	R-Square
Linear Regression	53.52	7.31	5.29	0.06	-0.62
ARD Regression	34.03	5.83	4.69	0.05	-0.03
Bayesian Ridge	33.98	5.83	4.68	0.05	-0.03
Elastic net	34.14	5.84	4.44	0.05	-0.03
Gamma Regressor	33.91	5.82	4.47	0.05	-0.03
Huber Regressor	54.91	7.41	5.45	0.06	-0.66
Orthogonal Matching Pursuit	34.14	5.84	4.71	0.05	-0.03
Passive Aggressive Regressor	46.06	6.79	5.41	0.06	-0.4
Poisson Regressor	34.12	5.84	4.73	0.05	-0.03
Ransac Regressor	65.49	8.09	5.86	0.07	-0.98
Ridge	34.03	5.83	4.69	0.05	-0.03
SGD Regressor	41.81	6.47	5.19	0.06	-0.27
Theilsen Regressor	50.83	7.13	5.43	0.06	-0.54
Tweedie Regressor	33.94	5.83	4.47	0.05	-0.03

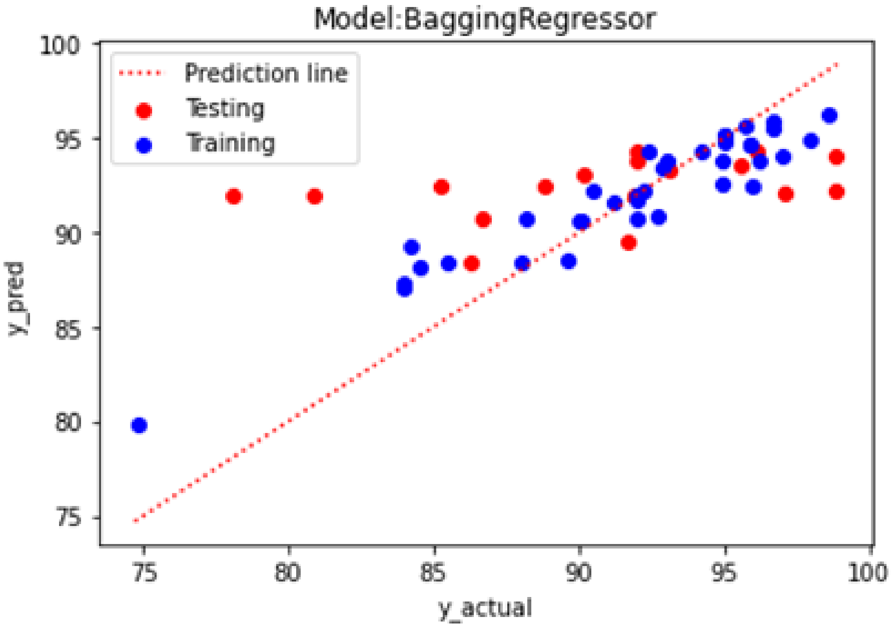

Figure 2: Bagging Regressor Model Performance.

Table 5: Nonlinear Model Analysis Results.

Model	MSE	RMSE	MAE	MAPE	R-Square
Adaboost Regressor	37.30	6.11	4.90	0.06	-0.13
Bagging Regressor	28.93	5.38	4.23	0.05	0.12
Gradient Boosting Regressor	33.37	5.78	4.73	0.05	-0.01
Random Forest Regressor	30.56	5.53	4.19	0.05	0.07
PLS Regression	35.66	5.97	4.77	0.05	-0.08
Decision Tree Regressor	48.22	6.94	5.85	0.07	-0.46
Extra Tree Regressor	46.93	6.85	5.56	0.06	-0.42
Dummy Regressor	34.21	5.85	4.45	0.05	-0.04
Gaussian Process Regressor	33.73	5.81	4.69	0.05	-0.02
Kernel Ridge	40.93	6.39	5.29	0.06	-0.24
KNeighbors Regressor	34.30	5.86	5.00	0.06	-0.03
XGB Regressor	34.67	5.89	4.61	0.05	-0.05

Acknowledgements

WAEP is grateful for the financial support provided by Universitas Dian Nuswantoro through the Internal Research Grand Project 2023.

References

[1] Wanli Wu FL, Chen R, Yang Z, He Z, Zhou Y. Corrosion resistance of 45 carbon steel enhanced by laser graphene-based coating, Diamond and Related Materials. Diamond Related Materials. 2021;116:108370.

[2] Bupesh Raja PG, Palanikumar K, Rohith Renish R, Ganesh Babu AN. Jashwanth Varma, “Corrosion resistance of corten steel – A review,” *Mater. Today Proc.*, vol. 46 Part 9, p. Pages 3572-3577, 2021, https://doi.org/10.1016/j.matpr.2021.01.334.

[3] Ichchou I, Larabi L, Rouabhi H, Harek Y, Fellah A; I. I. and L. L. and H. R. and Y. H. and A. Fellah. Electrochemical evaluation and DFT calculations of aromatic sulfonohydrazides as corrosion inhibitors for XC38 carbon steel in acidic media. J Mol Struct. 2019;1198:126898.

[4] Gutiérrez E, Rodríguez JA, Cruz-Borbolla J, Alvarado-Rodríguez JG, Thangarasu P; E. G. and J. A. R. and J. C.-B. and J. G. A.-R. and P. Thangarasu. Development of a predictive model for corrosion inhibition of carbon steel by imidazole and benzimidazole derivatives. Corros Sci. 2016;108:23–35.

[5] Koch G. 1-Cost of Corrosion. Trends in Oil and Gas Corrosion Research and Technologies. 2017;108:3–30.

[6] Marzorati S, Verotta L, Trasatti SP. Green corrosion inhibitors from natural sources and biomass wastes. Molecules. 2018 Dec;24(1):48.

[7] Quraishi MA, Chauhan DS, Saji VS; M. A. Q. and D. S. C. and V. S. Saji. Heterocyclic biomolecules as green corrosion inhibitors. J Mol Liq. 2021;341:117265.

[8] Sutojo T, Rustad S, Akrom M, Syukur A, Shidik GF, Dipojono HK. "A machine learning approach for corrosion small datasets," *npj Mater*. NPJ Mater Degrad. 2023;7(1):18.

[9] Budi S, Akrom M, Trisnapradika GA, Sutojo T, Aji W, Prabowo E. Optimization of Polynomial Functions on the NuSVR Algorithm Based on Machine Learning: Case Studies on Regression Datasets. Sci. J. Informatics. 2023;10(2):151–8.

[10] Akrom M, Rustad S, Saputro AG, Dipojono HK; M. A. and S. R. and A. G. S. and H. K. Dipojono. Data-driven investigation to model the corrosion inhibition efficiency of Pyrimidine-Pyrazole hybrid corrosion inhibitors. Comput Theor Chem. 2023;1229:114307.

[11] M. A. and S. R. and H. {Kresno Dipojono}, "Machine learning investigation to predict corrosion inhibition capacity of new amino acid compounds as corrosion inhibitors," *Results Chem.*, p. 101126, 2023, https://doi.org/10.1016/j.rechem.2023.101126.

[12] Alamri AH, Alhazmi N. Development of data driven machine learning models for the prediction and design of pyrimidine corrosion inhibitors. J Saudi Chem Soc. 2022;26(6):101536.

[13] Quadri TW, Olasunkanmi LO, Fayemi OE, Lgaz H, Dagdag O, Sherif EM, et al. Predicting protection capacities of pyrimidine-based corrosion inhibitors for mild steel/HCl interface using linear and nonlinear QSPR models. J Mol Model. 2022 Aug;28(9):254.

[14] Zhao H, Zhang X, Ji L, Hu H, Li Q. Quantitative structure-activity relationship model for amino acids as corrosion inhibitors based on the support vector machine and molecular design. Corros Sci. 2014;83:261–71.

[15] Du L, Zhao H, Hu H, Zhang X, Ji L, Li H, et al. Quantum chemical and molecular dynamics studies of imidazoline derivatives as corrosion inhibitor and quantitative structure-activity relationship (QSAR) analysis using the support vector machine (SVM) method. J Theor Comput Chem. 2014;13(2):1450012.

[16] Khan PW, Park SJ, Lee SJ, Byun YC. Electric Kickboard Demand Prediction in Spatiotemporal Dimension Using Clustering-Aided Bagging Regressor. J Adv Transp. 2022;2022:1–15.

[17] Nasir Amin M, Iftikhar B, Khan K, Faisal Javed M, Mohammad AbuArab A, Faisal Rehman M; M. Nasir A. and B. I. and K. K. and M. Faisal J. and A. Mohammad A. and M. Faisal Rehman. Prediction model for rice husk ash concrete using AI approach: boosting and bagging algorithms. Structures. 2023;50:745–57.

[18] Jacob Creutzig P. Wojtaszczyk, “Linear vs. nonlinear algorithms for linear problems,”. J Complexity. 2004;20(6):807–20.

[19] Pedregosa F, et al. “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011, [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[20] Al Mamun A, Sohel M, Mohammad N, Haque Sunny MS, Dipta DR, Hossain E. A Comprehensive Review of the Load Forecasting Techniques Using Single and Hybrid Predictive Models. IEEE Access. 2020;8:134911–39.

[21] Hu M, Zhao Y, Khushi M. Zexin and Zhao, Yiqi and Khushi, “A Survey of Forex and Stock Price Prediction Using Deep Learning,”. Appl Syst Innov. 2021;4(1):1.