Journal of Multiscale Materials Informatics

Vol. 1, No. 2, August 2024
https://publikasi.dinus.ac.id/index.php/jimat/ ISSN 3047-5724
DOI: 10.62411/jimat.v1i2.11021

XGBoost Performance In Predicting Corrosion Inhibition Efficiency of Benzimidazole Compounds

Diah Rahayu Ningtias[1], Muhamad Akrom[2]*

[1]Electromedical Engineering, STIKES Semarang, Semarang 50222, Indonesia
[2]Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

Article Info
Received : June 28, 2024
Revised : July 05, 2024
Accepted : July 06, 2024

Keywords:
XGBoost
Corrosion inhibition
Benzimidazole

ABSTRACT

In this study, we compare the performance of the XGBoost model with a Support Vector Machine (SVM) model from the literature in predicting a given task. Performance metrics such as the coefficient of determination (R2), root mean squared error (RMSE), and mean absolute error (MAE) were utilized to evaluate and compare the models. The XGBoost model achieved an R^2 of 0.99, an RMSE of 2.54, and an MAE of 1.96, significantly outperforming the SVM model, which recorded an R^2 of 0.96 and an RMSE of 6.79. The scatter plot for the XGBoost model further illustrated its superior performance, showing a tight clustering of points around the ideal line (y = x), indicating high accuracy and low prediction errors. These findings suggest that the XGBoost model is highly effective for the given prediction task, likely due to its ability to capture complex patterns and interactions within the data.

Corresponding Author:
email: m.akrom@dsn.dinus.ac.id

1. INTRODUCTION

Since material corrosion causes significant losses in various areas, including economics, the environment, society, industry, security, and safety, it is a major problem for both the industrial and academic worlds. [1], [2], [3]. Using inhibitor technology to manage corrosion is one of the easiest, most efficient, and least expensive approaches [4], [5], [6]. The capacity of inhibitor chemicals to create an adsorbed or protective layer on the metal surface, which can prevent mass transfer and charge transfer and shield the metal from corrosive environments, determines how successful they are [7], [8], [9]. Extensive expenses, time, and resources are needed for experimental research assessing different inhibitor chemical possibilities [10], [11], [12].

The quantitative structure-property relationship (QSPR) model based on the machine learning (ML) approach can be used further in investigating different candidate inhibitor compounds because electronic properties and chemical reactivity can be quantified against the chemical structure of compounds [13], [14], [15]. Density functional theory (DFT) calculations of quantum chemical descriptors (QCD) are an important component in creating accurate and dependable QSPR models. Typically, feature selection is done to extract pertinent quantum chemical descriptors that are then utilized in the QSPR model's construction [16], [17], [18], [19], [20]. Numerous reports on QSPR modeling of several quantum chemical descriptors are available. Furthermore, to attain performance efficacy and efficiency, inhibitor synthesis performance may be optimized through machine learning techniques before experimental analysis.

To assess inhibitor performance, a variety of machine learning (ML) algorithms have been combined and widely used, including genetic algorithms (GA), multiple linear regressions (MLR), partial least squares (PLS), ordinary least squares regressions (OLS), artificial neural networks (ANN), adaptive neural fuzzy inference systems (ANFIS), and autoregressive with exogenous inputs (ARX). Support vector machine (SVM), an ML algorithm, was utilized by Lu Li et al. [21] to investigate the potential of benzimidazole as a corrosion inhibitor. The results show that the SVM model has a coefficient of determination (R2) of 0.96 and an average root mean square error (RMSE) of 6.79. In this work, we implemented an XGBoost model to assess how well benzimidazole compounds suppress corrosion.

2. METHODS

2.1. Dataset

In this study, the dataset used was extracted from published literature and consisted of 20 benzimidazole sensitivities with 12 features and 1 target [21]. The features that are used are the senyawa molecular structure of benzimidazole; these include the energy of the highest occupied molecular orbital (E-HOMO), the energy of the lowest unoccupied molecular orbital (E-LUMO), polarizability (α), total natural charges (Qtotal), molar volume (Vi), the adiabatic ionization potential (Ia), the adiabatic electron affinity (Aa), electrophilicity (ωa), the fraction electron shared (ΔN), indexes of aromaticity in the benzene (ΛNICS(1)B) and indexes of aromaticity in the imidazole (ΛNICS(1)I) that are variable independent. The dependent variable of interest is the corrosion inhibition efficiency (CIE).

2.2. ML Modeling

During the preprocessing stage, the data normalization procedure is carried out using the MinMax scaling technique to reduce the data's sensitivity to the particular feature. The k-fold technique is used as a cross-validation (CV) model to account for bias and variance in data by training the model in a stepwise manner until the lowest statistical significance is reached [22], [23]. We use the k = 10 nilai, with 1 fold used as the test set and the remaining 9 folds as the training set (train). The k-fold nilai is sensitive to the data used; nonetheless, k = 5 or k = 10 are commonly used [24], [25]. The regression metrics root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R2) are used to assess the performance of the prediction model. The best models have values for MSE, RMSE, and MAE regularly, along with R2 being less than 1.

3. RESULTS and DISCUSSIONS

Table 1 provides a comparative analysis of the performance measures between the XGBoost model developed in this work and an SVM model from the literature.

Table 1. Model prediction performance

Model	R2	RMSE	MAE	Ref.
XGBoost	0.99	2.54	1.96	This work
SVM	0.96	6.79	-	[20]

R2 metric indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. RMSE is the square root of the average of squared differences between predicted and observed values. It provides a measure of the magnitude of the prediction error. MAE represents the average of the absolute differences between predicted and observed values, providing a measure of the average prediction error [26], [27], [28]. The XGBoost model explains 99% (R2 = 0.99) of the variance in the data, indicating an excellent fit. The relatively low RMSE (2.54%) value suggests that the XGBoost model has a small prediction error, reflecting high accuracy. MAE = 1.96 value shows that the average prediction error is 1.96 units, further supporting the model's high accuracy. The SVM model explains 96% (R2 = 0.96) of the variance, which is still very good but slightly lower than the XGBoost model.

The higher RMSE (6.96) value indicates that the SVM model has a larger prediction error compared to the XGBoost model. The XGBoost model shows a marginally higher R^2 value, suggesting that it has a slightly better ability to explain the variance in the dataset compared to the SVM model. The XGBoost model has a significantly lower RMSE, indicating much smaller prediction errors and thus higher precision in its predictions compared to the SVM model. While the SVM's MAE is not available for direct comparison, the provided RMSE values already suggest that the XGBoost model is superior in terms of prediction accuracy. The XGBoost model outperforms the SVM model from the literature based on the provided metrics. The higher R^2 and lower RMSE values indicate that XGBoost provides more accurate and reliable predictions. The absence of the MAE value for the SVM model slightly limits the depth of comparison, but the substantial difference in RMSE is a strong indicator of XGBoost's superior performance.

XGBoost, being an ensemble learning method based on decision trees, can capture complex patterns and interactions in the data more effectively than SVMs, which rely on a single decision boundary.

XGBoost has several hyperparameters that can be fine-tuned to optimize performance, potentially giving it an edge over SVM if both models are not equally well-tuned. XGBoost can handle non-linear relationships better due to its tree-based structure, whereas SVM may require kernel tricks to achieve similar performance.

Scatter plots are essential tools for visualizing the performance of predictive models. They allow us to examine the relationship between the predicted and actual values, providing insights into the model's accuracy and potential biases. In a perfect model, all points would lie on the line (y = x), indicating that the predicted values are equal to the actual values. Data points should be closely clustered around the line (y = x). The closer the points are to this line, the more accurate the predictions are. A uniform spread of points around the (y = x) line indicates that the model performs consistently across the range of actual values. Points that deviate significantly from the line indicate potential outliers or areas where the model performs poorly. If points systematically deviate above or below the (y = x) line, it suggests a bias in the model. For example, if points consistently fall below the line, the model tends to underpredict. Variability in the spread of points, such as a fan shape where spread increases with actual values, indicates heteroscedasticity. This means prediction errors vary with the magnitude of the actual values [29], [30].

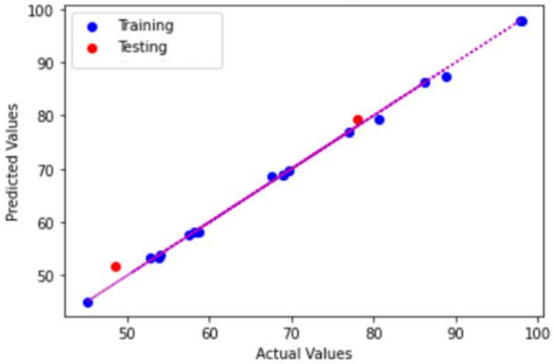

Figure 1. Scatter plot of data points from XGBoost performance

From Figure 1, the scatter plot of the XGBoost model performance shows points tightly clustered around the (y = x) line, reflecting the high proportion of variance explained by the model. The points have a minimal deviation from the (y = x) line, indicating low prediction errors. The spread of points is relatively uniform across the range of actual values, suggesting consistent model performance without significant biases or heteroscedasticity. The visual representation in the scatter plot validates the high R², low RMSE, and MAE values. It confirms that the numerical metrics accurately reflect the model's strong predictive performance.

4. CONCLUSION

The XGBoost model demonstrates superior performance compared to the SVM model from the literature in terms of R² and RMSE. This suggests that XGBoost is more effective for the given prediction task in this work, likely due to its ability to handle complex patterns and interactions within the data. The scatter plot for the XGBoost model shows a tight clustering of points around the ideal line (y = x), indicating high accuracy and low prediction errors. These results confirm that XGBoost not only provides a better fit but also maintains consistent prediction accuracy, making it a highly suitable choice for the given prediction task.

REFERENCES

[1] V.C. Anadebe, V.I. Chukwuike, S. Ramanathan, and R.C. Barik, Cerium-based metal organic framework (Ce-MOF) as corrosion inhibitor for API 5L X65 steel in CO2- saturated brine solution: XPS, DFT/MD-simulation, and machine learning model prediction, *Process Safety and Environmental Protection*, **168**, 499–512 (2022), https://doi.org/10.1016/J.PSEP.2022.10.016.

[2] M. Akrom, Investigation of natural extracts as green corrosion inhibitors in steel using density functional theory, *Jurnal Teori dan Aplikasi Fisika*, **10**(1), 89-102 (2022), https://doi.org/10.23960%2Fjtaf.v10i1.2927.

[3] T.L. Yusuf, T.W. Quadri, G.F. Tolufashe, L.O. Olasunkanmi, E.E. Ebenso, and W.E. Van Zyl, Synthesis and structures of divalent Co, Ni, Zn and Cd complexes of mixed dichalcogen and dipnictogen ligands with corrosion inhibition properties: Experimental and computational studies, *RSC Adv*, **10**(69), 41967–41982 (2020), https://doi.org/10.1039/d0ra07770d.

[4] H. Kumar and V. Yadav, Highly efficient and eco-friendly acid corrosion inhibitor for mild steel: Experimental and theoretical study, *J Mol Liq*, **335**, (2021), https://doi.org/10.1016/j.molliq.2021.116220.

[5] M. Akrom, DFT Investigation of Syzygium Aromaticum and Nicotiana Tabacum Extracts as Corrosion Inhibitor, *Science Tech: Jurnal Ilmu Pengetahuan dan Teknologi*, **8**(1), 42-48 (2022), https://doi.org/10.30738/st.vol8.no1.a11775.

[6] C. Verma, M.A. Quraishi, and E.E. Ebenso, Quinoline and its derivatives as corrosion inhibitors: A review, *Surfaces and Interfaces*, **21**, 100634 (2020), https://doi.org/10.1016/J.SURFIN.2020.100634.

[7] M. Akrom and T. Sutojo, Investigasi Model Machine Learning Berbasis QSPR pada Inhibitor Korosi Pirimidin Investigation of QSPR-Based Machine Learning Models in Pyrimidine Corrosion Inhibitors, *Eksergi*, **20**(2), 107-111 (2023), https://doi.org/10.31315/e.v20i2.9864.

[8] F.E. Abeng and V.C. Anadebe, Combined electrochemical, DFT/MD-simulation and hybrid machine learning based on ANN-ANFIS models for prediction of doxorubicin drug as corrosion inhibitor for mild steel in 0.5 M H2SO4 solution, *Comput Theor Chem*, **1229**, 114334 (2023), https://doi.org/10.1016/J.COMPTC.2023.114334.

[9] M. Akrom, S. Rustad, and H.K. Dipojono, A machine learning approach to predict the efficiency of corrosion inhibition by natural product-based organic inhibitors, *Phys Scr*, **99**,(3), 036006 (2024), https://doi.org/10.1088/1402-4896/ad28a9.

[10] T.W. Quadri, L.O. Olasunkanmi, O.E. Fayemi, H. Lgaz, O. Dagdag, E.M. Sherif, A.A. Alrashdi, E.D. Akpan, H. Lee, and E.E. Ebenso, Computational insights into quinoxaline-based corrosion inhibitors of steel in HCl: Quantum chemical analysis and QSPR-ANN studies, *Arabian Journal of Chemistry*, **15**(7), 103870 (2022), https://doi.org/10.1016/J.ARABJC.2022.103870.

[11] R.L. Camacho-Mendoza, L. Feria, L.Á. Zárate-Hernández, J.G. Alvarado-Rodríguez, and J. Cruz-Borbolla, New QSPR model for prediction of corrosion inhibition using conceptual density functional theory, *J Mol Model*, **28**(8), (2022), https://doi.org/10.1007/s00894-022-05240-6.

[12] M. Boudalia, R.M. Fernández-Domene, L. Guo, S. Echihi, M.E. Belghiti, A. Zarrouk, A. Bellaouchou, A. Guenbour, and J. García-Antón, Experimental and Theoretical Tests on the Corrosion Protection of Mild Steel in Hydrochloric Acid Environment by the Use of Pyrazole Derivative, *Materials*, **16**(2), (2023), https://doi.org/10.3390/ma16020678.

[13] M. Akrom, S. Rustad, and H.K. Dipojono, Machine learning investigation to predict corrosion inhibition capacity of new amino acid compounds as corrosion inhibitors, *Results Chem*, **6**, 101126 (2023), https://doi.org/10.1016/J.RECHEM.2023.101126.

[14] L.B. Coelho, D. Zhang, Y.V. Ingelgem, D. Steckelmacher, A. Nowé, and H. Terryn, Reviewing machine learning of corrosion prediction in a data-oriented perspective, *npj Materials Degradation*, **6**(1), (2022), https://doi.org/10.1038/s41529-022-00218-4.

[15] T.W. Quadri, L.O. Olasunkanmi, O.E. Fayemi, E.D. Akpan, H. Lee, H. Lgaz, C. Verma, L. Guo, S. Kaya, and E.E. Ebenso, Multilayer perceptron neural network-based QSAR models for the assessment and prediction of corrosion inhibition performances of ionic liquids, *Comput Mater Sci*, **214**, (2022), https://doi.org/10.1016/j.commatsci.2022.111753.

[16] M. Akrom, S. Rustad, A.G. Saputro, and H.K. Dipojono, Data-driven investigation to model the corrosion inhibition efficiency of Pyrimidine-Pyrazole hybrid corrosion inhibitors, *Comput Theor Chem*, **1229**, 114307 (2023), https://doi.org/10.1016/J.COMPTC.2023.114307.

[17] M. Akrom, S. Rustad, and H.K. Dipojono, Prediction of Anti-Corrosion performance of new triazole derivatives via Machine learning, *Comp and Theoretical Chem*, **1236**, 114599 (2024), https://doi.org/10.1016/j.comptc.2024.114599.

[18] T.W. Quadri, L.O. Olasunkanmi, E.D. Akpan, O.E. Fayemi, H. Lee, H. Lgaz, C. Verma, L. Guo, S. Kaya, and E.E. Ebenso, Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors, *Mater Today Commun*, **30**, 103163 (2022), https://doi.org/10.1016/J.MTCOMM.2022.103163.

[19] M. Akrom, S. Rustad, and H.K. Dipojono, SMILES-based machine learning enables the prediction of corrosion inhibition capacity, *MRS Comm*, (2024), https://doi.org/10.1557/s43579-024-00551-6.

[20] M. Akrom, S. Rustad, and H.K. Dipojono, Variational quantum circuit-based quantum machine learning approach for predicting corrosion inhibition efficiency of pyridine-quinoline compounds, *Mater Today Quantum*, (2024), https://doi.org/10.1016/j.mtquan.2024.100007.

[21] L. Li et al., The discussion of descriptors for the QSAR model and molecular dynamics simulation of benzimidazole derivatives as corrosion inhibitors," *Corrosion Science*, vol. 99, p. 76–88, Oct 2015, https://doi.org/10.1016/j.corsci.2015.06.003.

[22] C. Beltran-Perez, A.A.A. Serrano, G. Solís-Rosas, A. Martínez-Jiménez, R. Orozco-Cruz, A. Espinoza-Vázquez, and A. Miralrio, A General Use QSAR-ARX Model to Predict the Corrosion Inhibition Efficiency of Drugs in Terms of Quantum Mechanical Descriptors and Experimental Comparison for Lidocaine, *Int J Mol Sci*, **23**(9), (2022), https://doi.org/10.3390/ijms23095086.

[23] Y.G. Skrypnik, T.F. Doroshenko, and S.Y. Skrypnik, ON THE INFLUENCE OF THE NATURE OF SUBSTITUENTS ON THE INHIBITING ACTIVITY OF META-AND PARA-SUBSTITUTED PYRIDINES, *Materials Science*, **31**, 324-330 (1996), https://doi.org/10.1007/BF00558554.

[24] T.V. Doroshenko, S.N. Lyashchuk, and Y.G. Skrypnik, The HSAB Principle in the Description of the Inhibitive Effectiveness of Heterocyclic N-Bases, *Protection of Metals*, **36**, 244-247 (2000).

[25] M. Akrom, T. Sutojo, A. Pertiwi, S. Rustad, and H.K. Dipojono, Investigation of Best QSPR-Based Machine Learning Model to Predict Corrosion Inhibition Performance of Pyridine-Quinoline Compounds, *J Phys Conf Ser*, **2673** (1), 012014 (2023), https://doi.org/10.1088/1742-6596/2673/1/012014.

[26] S. Budi, M. Akrom, H. Al Azies, U. Sudibyo, T. Sutojo, G.A. Trisnapradika, A.N. Safitri, A. Pertiwi, and S. Rustad, Implementation of Polynomial Functions to Improve the Accuracy of Machine Learning Models in Predicting the Corrosion Inhibition Efficiency of Pyridine-Quinoline Compounds as Corrosion Inhibitors, *KnE Engineering*, 78-87 (2024), https://doi.org/10.18502/keg.v6i1.15351.

[27] M. Akrom, A.G. Saputro, A.L. Maulana, A. Ramelan, A. Nuruddin, S. Rustad, and H.K. Dipojono, "DFT and microkinetic investigation of oxygen reduction reaction on corrosion inhibition mechanism of iron surface by Syzygium Aromaticum extract, *Appl Surf Sci*, **615**, 156319 (2023), https://doi.org/10.1016/j.apsusc.2022.156319.

[28] W. Herowati, W.A.E. Prabowo, M. Akrom, T. Sutojo, N.A. Setiyanto, A.W. Kurniawan, N.N. Hidayat, and S. Rustad, Prediction of Corrosion Inhibition Efficiency Based on Machine Learning for Pyrimidine Compounds: A Comparative Study of Linear and Non-linear Algorithms, *KnE Engineering*, 68-77 (2024), https://doi.org/10.18502/keg.v6i1.15350.

[29] M. Akrom, S. Rustad, A.G. Saputro, A. Ramelan, F. Fathurrahman, and H.K. Dipojono, A combination of machine learning model and density functional theory method to predict corrosion inhibition performance of new diazine derivative compounds, *Mater Today Commun*, **35**, 106402 (2023), https://doi.org/10.1016/J.MTCOMM.2023.106402.

[30] M. Akrom, S. Rustad, and H.K. Dipojono, Development of quantum machine learning to evaluate the corrosion inhibition capability of pyrimidine compounds, *Mater Today Comm*, **39**, 108758 (2024), https://doi.org/10.1016/j.mtcomm.2024.108758.