Investigation of Best QSPR-Based Machine Learning Model to Predict Corrosion Inhibition Performance of Pyridine-Quinoline Compounds

Investigation of Best QSPR-Based Machine Learning Model to Predict Corrosion Inhibition Performance of Pyridine-Quinoline Compounds

Muhamad Akrom[1*], Totok Sutojo[1], Ayu Pertiwi[1], Supriadi Rustad[1], Hermawan Kresno Dipojono[2]

[1]Universitas Dian Nuswantoro, Semarang 50131, Indonesia
[2]Institut Teknologi Bandung, Bandung 40132, Indonesia

*m.akrom@dsn.dinus.ac.id

Abstract. Corrosion is a major concern for the industrial and academic sectors because it causes significant losses in many fields. Currently, there is a great deal of interest in the topic of material damage control using organic chemicals. Pyridine and quinoline are potential corrosion inhibitors because they are non-toxic, inexpensive, and efficient in various corrosive conditions. Experimental studies in searching for candidate corrosion inhibitor candidates require a lot of time, cost, and labor intensive. Using a machine learning (ML) strategy based on a quantitative structure-property relationship (QSPR) model, we evaluate gradient boosting regressor (GBR), support vector regression (SVR), and k-nearest neighbor (KNN) algorithms as predictive models to investigate corrosion inhibition efficiency (CIE) of pyridine-quinoline compounds in this study. We found that the GBR model, when compared with the SVR and KNN models as well as models from the literature for the pyridine-quinoline compound dataset, has the best predictive performance based on the metric coefficient of determination (R^2) and root mean square error (RMSE). Overall, our study provides a new perspective on how the ML model can estimate the effectiveness of corrosion prevention on iron surfaces by organic inhibitor compounds.

1. Introduction

Because it causes significant losses in various fields including the economy, environment, society, industry, security, and safety, corrosion is a serious concern for industry and academia [1]–[3]. One of the easiest, most efficient, and inexpensive corrosion management methods is to use inhibitor technology [4], [5]. The efficiency of corrosion inhibitor chemicals depends on their capacity to create a barrier on the metal surface that can stop charge and mass transfer, protecting the metal from corrosive conditions [6], [7]. Experiments to assess the many potential candidate compounds as corrosion inhibitors required a significant investment of time, money, and resources.

The QSPR-based ML method can be used to study various potential inhibitor compounds because the electronic properties and chemical reactivity can be assessed to the chemical structure of these compounds [8]–[10]. The calculation of the density functional theory (DFT) method that generates quantum chemical descriptors (QCDs) is very important in creating an accurate and reliable QSPR model. Significant QCDs are obtained by feature selection for use in the construction of the QSPR model [11], [12]. There has been much interest in QSPR modeling of different QCDs using linear and non-linear regression. To increase performance effectiveness and efficiency, the application of the ML algorithm can also optimize the performance of inhibitor synthesis before experimental analysis.

Many ML techniques are used in the development of the QSPR model to assess the effectiveness of inhibitors. To assess the effectiveness of pyridine-quinoline compounds as corrosion inhibitors, Mendoza et al. [13] use a linear multilinear regression (MLR) model and produce a predictive model performance of $R^2 = 0.93$. Ser et al. [14] used a nonlinear genetic algorithm-artificial neural network (GA-ANN) model to estimate the effectiveness of the same compound as a corrosion inhibitor and the resulting model predicted the performance of %RMSE = 8.80. We hypothesize that there should be an identical model that can be applied to the two datasets from the studies mentioned above because the compounds used in the two studies are the same. Developing an accurate ML model is one of the main challenges. To assess the effectiveness of the ability of pyridine-quinoline organic compounds as corrosion inhibitors, therefore, we propose a QSPR-based ML model in this study and compare its performance with the gradient boosting regressor (GBR) model, support vector regression (SVR), and k-nearest neighbors (KNN).

2. Method

2.1. Dataset, feature, and target

We used the pyridine-quinoline compound dataset from the literature by Mendoza et al. [13] and Ser et al. [14] which has previously been published. To assist in the design of corrosion inhibition, the QSPR model was constructed using several *QCDs* such as HOMO-LUMO, electron affinity, electronegativity, global hardness, global softness, electrophilicity, a fraction of electrons transferred, total energy, gap energy, dipole moment, and ionization potential. as a feature. CIE is used as a target. In general, *QCDs* can be calculated using the DFT approach and Koopman's theory. CIE is generally obtained from the results of experimental studies. Corrosion inhibition is highly dependent on the chemical reactivity of inhibitor molecules, which are characterized by various quantum chemical descriptors [15], [16].

The ability of an inhibitor molecule to donate electrons can be described by the HOMO orbital, while its ability to accept electrons can be described by the LUMO orbital. The HOMO-LUMO orbital conditions allow the analysis of electron transfer based on the energy values of the electrons. Electron donor acceptors from metal surfaces are made possible by molecules. The ability of molecules to bind to metal surfaces is indicated by the gap energy, which is the energy difference between LUMO and HOMO. The ionization potential and electron affinity of molecules also affect their reactivity. The ability of the inhibitor molecule to attract electrons and achieve electron equilibrium is referred to as electronegativity. Global hardness indicates the capacity of a molecule to accept a charge, while global softness indicates a molecule's resistance to charge transfer [13], [14], [17]–[19]. The ability of a molecule to interact with a metal surface dipole is described by the molecular dipole moment. This is due to the increased surface area of the molecule with the metal surface, which increases its capacity to block corrosive substances. The capacity of a molecule to absorb electrons is known as electrophilicity. The transfer of electrons from the inhibitor molecule to the atoms on the metal surface when the molecule and the metal surface come into contact. Electron transfer is due to the lower electronegativity of the molecules compared to the metal surface. Electrons will move from low electronegativity to high electronegativity until the chemical potential is balanced. The abilities of molecules to adsorb on a metal surface are directly related to the total energy. In general, the interaction of molecules with metal surfaces results in corrosion inhibition. On metal surfaces, corrosion inhibitors can work by chemical or physical adsorption [20]–[22].

2.2. ML Model

We used three algorithms, namely GBR, SVR, and K-NN to create a prediction model for CIE. The tendency of the three models to explain the potential relationship between features (QCD) and targets (CIE) was evaluated. A 70:30 ratio between training and testing is used in the dataset. To prevent problems with the sensitivity of certain features in the prediction results, data normalization is carried out in the preprocessing step [23]. The model is validated using the k-fold cross-validation technique, which uses one set of data for validation and the remaining set for model training. Model performance

is evaluated using R^2 and RMSE metrics. All parameters and other variable default settings are from sci-kit learn version 0.23.2.

3. Result and Discussion

Table 1 displays the model's predictive performance metrics. Figures 1 and 2 show how the data points are distributed. Figure 3 shows the relationship between the predicted value and the actual (experimental) value.

Table 1. Model prediction performances.

Dataset	Current work			Other work		
	Model	R^2	RMSE	Model	R^2	RMSE
Pyridine-Quinoline_M (PKM)	GBR	0.999	0.001	MLR [13]	0.930	na
	KNN	0.644	6.359			
	SVR	0.042	27.981			
Pyridine-Quinoline_S (PKS)	GBR	0.999	0.041	GA-ANN [14]	na	8.800
	KNN	0.469	8.428			
	SVR	0.065	31.087			

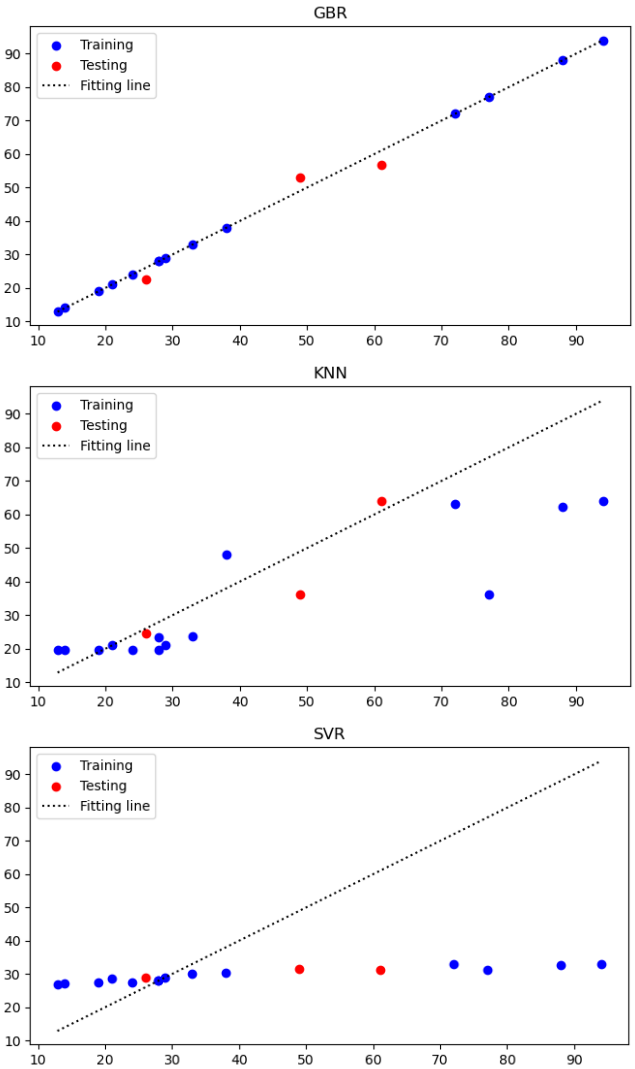


Figure 1. Scatter plot of data point prediction for the PKM dataset.

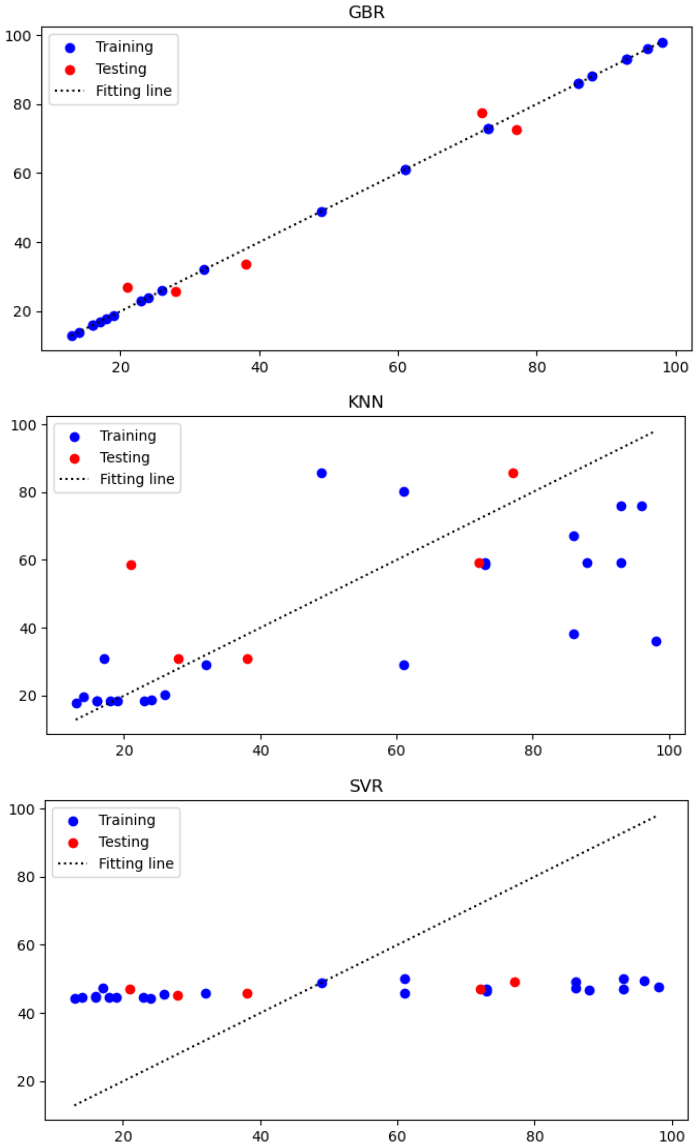
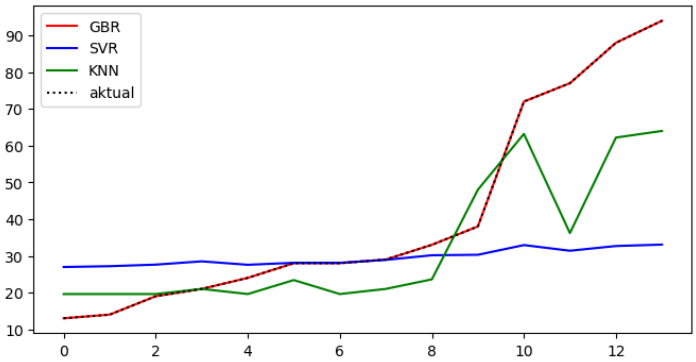
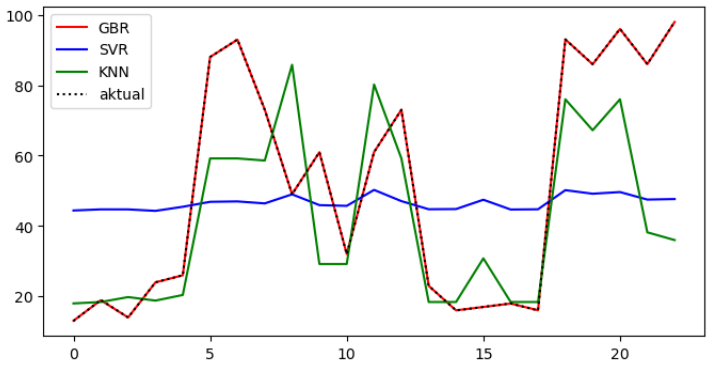


Figure 2. Scatter plot of data point prediction for the PKS dataset.



(a)



(b)

Figure 3. A plot of CIE value between prediction and actual (experimental) for (a) PKM and (b) PKS dataset.

According to our research findings shown in Table 1, the GBR model has the highest R^2 value and lowest RMSE for both datasets (PKM and PKS) when compared to the KNN and SVR models. The ideal model has an R^2 value close to 1 and the lowest RMSE. The data points predicted by the GBR model tend to be closer to the prediction line (fitting line) than the other two models, as shown in Figure 1 for the representation of the distribution of data points for the PKM dataset and the same representation in Figure 2 for the PKS dataset. This strongly confirms the findings based on those R^2 and RMSE metrics. In addition, it can be shown in Figure 3 that the predicted value of the CIE (target) from the GBR model (represented in red) shows a pattern that is most comparable to the pattern of the actual (experimental) CIE value. The GBR model outperforms the KNN and SVR models and shows the best predictive performance in both datasets because the resulting predicted values are almost identical to the actual values. These results indicate that there are identical specific models that can provide good and superior accuracy in the investigation of pyridine-quinoline compounds as corrosion inhibitors.

Table 1 further shows that our three proposed models (GBR, SVR, and KNN) offer competitive predictive performance with models from the literature. In the PKM dataset, compared to the model from the literature (MLR), the GBR model is superior, but the KNN and SVR models are the opposite, based on the R^2 value. In the PKS dataset, compared to models from the literature (GA-ANN), the GBR and KNN models are superior, but the SVR model is the opposite, based on RMSE values. The most important finding is that the GBR model outperforms the other two models (KNN and SVR) as well as models from the literature (MLR and GA-ANN) in both pyridine-quinoline datasets. These results confirm that there is a specific reliable model to investigate the corrosion inhibition performance of pyridine-quinoline compounds.

4. Conclusion

We propose a superior QSPR-based ML technique to evaluate the predictive performance of the GBR, SVR, and KNN models in predicting the inhibitory effectiveness of pyridine-quinoline compounds against iron corrosion. Compared to the other two models, the GBR model has the best predictive capacity based on R^2 and RMSE values. When compared to models from previous works in literature, the GBR model is also robust. Overall, our study provides a new perspective on how the ML model can predict corrosion prevention on iron surfaces. Our model can still be improved in future studies to improve prediction accuracy, for example by including polynomial functions and/or virtual samples.

Acknowledgments

All calculations were performed using the Computation Facility at the Research Center for Materials Informatics, Universitas Dian Nuswantoro, Semarang, Indonesia. SR acknowledges funding support by BRIN Indonesia through the “RIIM 2023” program. HKD acknowledges funding support by DIKTIRISTEK through the “PFR 2023”.

References

[1] Anadebe V C, Nnaji P C, Onukwuli O D, Okafor N A, Abeng F E, Chukwuike V I, Okoye C C, Udoh I I, Chidiebere M A, Guo L and Barik L C 2022 *J. Mol. Liq.* **349** 118482
[2] Quadri T W, Olasunkanmi L O, Akpan E D, Fayemi O E, Lee H, Lgaz H, Verma C, Guo L, Kaya S and Ebenso E E 2022 *Mater. Today Commun.* **30** 103163
[3] Akrom M and Sutojo T 2023 *Eksergi* **20** 107
[4] Kumar D, Jain V and Rai B 2022 *Corros. Sci*. **195** 109960
[5] Akrom M, Sudibyo U, Kurniawan A W, Setiyanto N A, Pertiwi A, Safitri A N, Hidayat N N, Azies H, Herowati W 2023 *JOMMIT* **7** 15
[6] Thakur A, Kaya S, Abousalem A S and Kumar A 2022 *Sustain. Chem. Pharm*. **29** 100785
[7] Haladu S A, Mu'azu N D, Ali S A, Elsharif A M, Odewunmi N A and Abd El-Lateef H M 2022 *J. Mol. Liq*. **350** 118533
[8] Belghiti M E, Echihi S, Dafali A, Karzazi Y, Bakasse M, Elalaoui-Elabdallaoui H, Olasunkanmi L O, Ebenso E E and Tabyaoui M 2019 *Appl. Surf. Sci.* **491** 707
[9] Beltran-Perez C, Serrano A A A, Solís-Rosas G, Martínez-Jiménez A, Orozco-Cruz R, Espinoza-Vázquez A and Miralrio A 2022 *Int. J. Mol. Sci*. **23** 5086
[10] Akrom M, Rustad S and Dipojono H K 2023 *Results in Chemistry* **6** 101126
[11] Quadri T W, Olasunkanmi L O, Fayemi O E, Akpan E D, Lee H, Lgaz H, Verma C, Guo L, Kaya S, Ebenso EE 2022 *Comp. Materials Sci.* **214** 111753
[12] Akrom M, Rustad S, Saputro A. G, Ramelan A, Fathurrahman F and Dipojono H K 2023 *Mater. Today Commun.* **35** 106402
[13] Camacho-Mendoza R L, Feria L, Zárate-Hernández L. Á, Alvarado-Rodríguez J. G and Cruz-Borbolla J 2022 *J. Mol. Model.* **28** 238
[14] Ser C T, Žuvela P and Wong M W 2020 *Appl. Surf. Sci.* **512** 145612
[15] Sutojo T, Rustad S, Akrom M, Syukur A, Shidik G F and Dipojono H K 2023 *Npj Mater. Degrad.* **7** 18
[16] Quadri T W, Olasunkanmi L O, Fayemi O E, Lgaz H, Dagdag O, Sherif E M, Alrashdi A A, Akpan E D, Lee H and Ebenso E E 2022 *Arabian Journal of Chemistry* **15** 103870
[17] Assiri H, Driouch M, Lazrak J, Bensouda Z, Elhaloui A, Sfaira M, Saffaj T and Taleb M 2020 *Heliyon* **6**
[18] Li L, Zhang X, Gong S, Zhao H, Bai Y, Li Q and Ji L 2015 *Corros. Sci.* **99** 76
[19] Akrom M, Rustad S, Saputro A G and Dipojono H K 2023 *Comput. Theor. Chem.* **1229** 114307
[20] Kozlica D K, Kokalj A and Milošev I 2021 *Corros. Sci.* **182** 109082
[21] Kokalj A 2022 *Corrosion Science* **196** 109939
[22] Akrom M, Saputro A G, Maulana A L, Ramelan A, Nuruddin A, Rustad S and Dipojono H K 2023 *Appl. Surf. Sci*. **615** 156319
[23] Budi S, Akrom M, Trisnapradika G A, Sutojo T and Prabowo W A E 2023 *Scientific Journal of Informatics* **10** 151