Development of a Machine Learning Model to Predict the Corrosion Inhibition Ability of Benzimidazole Compounds

Aprilyani Nur Safitri¹,², Gustina Alfa Trisnapradika¹,², Achmad Wahid Kurniawan¹,², Wahyu Aji Eko Prabowo¹,², Muhamad Akrom¹,²*

¹Research Center for Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia
²Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

Article Info

ABSTRACT

This study uses a quantitative structure-property relationship (QSPR)-based machine learning (ML) to examine the corrosion inhibition capabilities of benzimidazole compounds. The primary difficulty in ML development is creating a model with high precision so that the predictions are correct and pertinent to the material's actual attributes. We assess the comparison between the extra trees regressor (EXT) as an ensemble model and the decision tree regressor (DT) as a basic model. It was discovered that the EXT model had better predictive performance in predicting the corrosion inhibition performance of benzimidazole compounds based on the coefficient of determination (R²) and root mean square error (RMSE) metrics compared DT model. This method provides a fresh viewpoint on the capacity of ML models to forecast potent corrosion inhibitors.

Keywords:
- machine learning
- QSPR
- corrosion inhibition
- benzimidazole

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

This publication is licensed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. INTRODUCTION

Corrosion The process of material degradation or decay brought on by chemical reactions between metal and the environment, where a variety of corrosive chemicals exist, is known as corrosion [1], [2], [3]. Oxides, hydroxides, and metal salts are among the corrosion products that are created when oxygen in the air or other corrosive materials oxidize metals. This corrosion response can lower the material's service life, impair its quality and performance, and result in large financial losses [4], [5], [6]. The kind of metal involved, the corrosive environment (such as humidity, pH, temperature, concentration of corrosive chemicals), and other elements like mechanical stress or frictional wear are some of the variables that affect the pace of corrosion [7], [8], [9]. In addition, stress-induced corrosion, microorganism interaction (such as bacteria), and galvanic corrosion (contact between two distinct metals in the electrolyte) can all speed up the corrosion process [10], [11], [12]. Understanding corrosion mechanisms, creating corrosion control strategies, and assessing material performance in corrosive settings are all part of corrosion studies [7]. Many industries, including the oil and gas, chemical, automotive, and construction sectors, benefit greatly from controlling the corrosion process [13], [14], [15].

The chemical compound known as the benzimidazole compound (C₇H₆N₂) is made up of a heterocyclic ring including the main structures of imidazole (C₃H₃N₂) and benzene (C₆H₅). Benzimidazole compounds are employed in a variety of industries, such as materials chemistry, agrochemistry, and medicines. Research has revealed that their derivatives exhibit a wide range of molecular functions [16], [17], [18]. Chemical synthesis uses benzimidazole molecules as organic pigments, corrosion rate regulators, and catalysts. It takes a lot of money, time, and resources to research the application of benzimidazole as an experimental corrosion inhibitor [19], [20], [21].
Presently, it is possible to apply quantum mechanical methods in conjunction with technological advancements to expedite the design and search for novel materials. Machine learning (ML) techniques, such as grouping, classification, and the creation of predictive models from one of the topics, corrosion, are made possible by the study of artificial intelligence. Lately, the investigation of novel materials has made extensive use of ML techniques. This is because the quantitative structure-property relationship (QSPR) of a compound and its structure are related; so, an ML technique may be used to create the QSPR model and assess the effectiveness of corrosion inhibitor compounds [22], [23], [24].

In this work, we examined the ML model to predict benzimidazole compounds' corrosion inhibition efficiency (CIE). We compared the extra trees regressor (EXT) as an ensemble model and the decision tree regressor (DT) as a basic model. It is anticipated that the findings of this study may shed light on how to develop ML models for the creation of possible compounds that block corrosion, preventing corrosion damage to materials.

### 2. METHODS

#### 2.1. Dataset
A dataset is essential for carrying out research using ML. The dataset used in this work, which comprises 20 benzimidazole compounds with 12 characteristics and 1 target, was acquired from published literature [24]. HOMO, LUMO, polarizability (α), total charge (Q), volume (V), ionization potential (I), electron affinity (A), electrophilicity (ω), transferred electron fraction (∆N), benzene aromatic index (ΛNICS(1)B), and imidazole aromatic index (ΛNICS(1)I) are among the molecular properties of the benzimidazole compound that are used as input features. The dependent variable, CIE, is the target in the interim [25], [26], [27].

#### 2.2. ML Model
The objective of this study was to compare EXT as an ensemble model and the DT as a basic model to determine the optimal model for benzimidazole compound CIE prediction. Before applying cross-validation (CV), data preprocessing is done to remove noise from the data and normalize (scale) it to prevent data sensitivity to particular features. The k-fold strategy was selected as a CV model to reduce statistical error by repeatedly training the model until bias and variation in the data were eliminated [28], [29]. We utilize k = 10, which designates one fold as the test set and the remaining nine as the training set. The data being used determines the appropriate k-fold value, however, values of k = 5 or k = 10 are frequently employed [30], [31].

Regression metrics, such as root mean square error (RMSE) and coefficient of determination (R²), are used to assess the performance of the prediction model. The model with the highest R² and the lowest values of RMSE, MAE, and MSE is the best one [32], [33].

### 3. RESULT AND DISCUSSION
The performance of each model is measured by the R² and RMSE values as denoted in Table 1. From Table 1, EXT shows superior prediction performance compared to DT based on the evaluation metrics used (R² and RMSE). These results are also confirmed by the distribution of data points in Figure 1, where the distribution is closer to the prediction line for the EXT model than for the DT model.

<table>
<thead>
<tr>
<th>Model</th>
<th>R²</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Testing</td>
</tr>
<tr>
<td>EXT</td>
<td>0.97</td>
<td>0.68</td>
</tr>
<tr>
<td>DT</td>
<td>0.93</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Analysis of the important features in Figure 2 shows that the descriptors total charge (Q) and ionization potential (I) respectively appear as the most influential features in determining the prediction results of the EXT model. Apart from that, it can also be seen that other features also show a positive correlation with the CIE target, this shows that there is a good correlation [32], [33] thus making the EXT model able to predict more accurately.

4. CONCLUSION

Investigation of the ML model to predict the CIE value of benzimidazole compounds has been carried out by comparing the EXT and DT models. The EXT model was confirmed as a more accurate model than the DT model based on the $R^2$ and RMSE metrics. This research provides important insights into developing effective and efficient material exploration methods so that they can be taken into consideration by the industry in designing corrosion inhibitor materials.

REFERENCES


