Faktor Exacta

Vol. 17, No. 3, Sept 2024, pp. 251~261
eISSN: 2502-339X, pISSN: 1979-276X, DOI: 10.30998/faktorexacta.v17i3.24679

Comparative Analysis of Linear Regression, Decision Tree, and Gradient Boosting Models for Predicting Drug Corrosion Inhibition Efficiency Using QSAR Descriptors

Darnell Ignasius[1], Muhamad Akrom[2], Setyo Budi[2]
[1] Study Program in Information Systems, Faculty of Computer Science
Universitas Dian Nuswantoro, Semarang 50131, Indonesia
[2] Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science
Universitas Dian Nuswantoro, Semarang 50131, Indonesia

Article Info

Article history:

Received 22 Jul 2024
Revised 18 Sept 2024
Accepted 25 Sept 2024

Keywords:

Corrosion Inhibition Efficiency
QSAR
Machine Learning
Gradient Boosting
Computational Chemistry

ABSTRACT

Corrosion in industrial environments poses significant economic and safety challenges, necessitating the development of effective inhibitors. Organic compounds, particularly pharmaceuticals, have emerged as promising corrosion inhibitors due to their efficiency and environmental benefits. However, predicting these compounds' corrosion inhibition efficiency (CIE) remains complex and requires advanced computational methods. This study investigates the predictive capabilities of three machine learning (ML) models, namely linear regression, decision tree, and gradient boosting regression, using Quantitative Structure-Activity Relationship (QSAR) descriptors. A dataset containing 14 QSAR descriptors was compiled from experimental studies on various pharmaceutical-based inhibitors. The dataset was divided into training (90%) and testing (10%) subsets to evaluate model performance. The research follows the CRISP-DM methodology, a systematic framework that includes data preparation, model training, and evaluation. Key performance metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), were used to assess model accuracy. Among the models, Gradient Boosting Regression achieved the most promising results, with the lowest MSE (21.52) and the highest R^2 (0.21), reflecting its ability to capture non-linear relationships in the data. Despite the relatively modest R^2, this model demonstrates the potential for improving computational approaches to corrosion inhibition prediction. This study highlights the value of machine learning in optimizing the selection of corrosion inhibitors, potentially reducing the reliance on extensive laboratory testing and accelerating the discovery of efficient, eco-friendly solutions for industrial applications.

Corresponding Author:

Muhamad Akrom,
Research Center for Quantum Computing and Materials Informatics,
Faculty of Computer Science, Universitas Dian Nuswantoro,
Email: m.akrom@dsn.dinus.ac.id

1. INTRODUCTION (10 PT)

Corrosion is a pervasive issue that affects industries worldwide, leading to significant economic losses and safety concerns. From pipelines and machinery to critical infrastructure, the degradation of materials due to corrosion can result in costly maintenance downtime and even catastrophic failures. In response, corrosion inhibitor chemicals that reduce or prevent corrosion have been widely used as a protective measure, especially in industries such as oil and gas, manufacturing, and marine transportation [1], [2]. However, selecting the suitable corrosion inhibitor, particularly at the molecular level, remains a complex and resource-intensive

challenge. The use of corrosion inhibitors, especially those derived from organic compounds, has been a practical approach to address this issue [3]. These molecules, often possessing complex structures and varying electronic properties, can adsorb onto metal surfaces, creating a protective barrier that prevents corrosive agents from reacting with the metal. Corrosion Inhibition Efficiency (CIE) is a critical parameter in evaluating the performance of corrosion inhibitors. A crucial metric that determines the protective ability of a compound. The development of new, more effective corrosion inhibitors thus requires an accurate prediction of CIE to streamline the selection and testing of candidate molecules.

Traditionally, the evaluation of corrosion inhibitors has relied on experimental methods, which are both time-consuming and costly. Although accurate, these methods require extensive laboratory testing, which can slow down the identification of suitable inhibitors, especially when dealing with vast libraries of potential compounds. As a result, there is growing interest in leveraging computational approaches to predict the efficacy of corrosion inhibitors, mainly through machine learning (ML) models [6]. Machine learning, which can analyse complex patterns in data, offers a promising alternative by enabling rapid predictions based on molecular descriptors, significantly reducing the need for exhaustive experimental testing. This study focuses on applying machine learning techniques to predict the CIE of organic molecules using quantitative structure-activity relationship (QSAR) descriptors. Quantitative Structure-Activity Relationship (QSAR) has proven to be a powerful approach to predicting various molecular properties, including CIE, where molecules' chemical and physical properties are related to their biological activities [7]. In this case, QSAR descriptors derived from the quantum mechanical properties of molecules are used as input features for machine learning models, enabling the prediction of CIE without the need for extensive experimental data. This approach aligns with the broader trend toward data-driven scientific research methods, particularly materials science.

Despite the potential of machine learning in corrosion research, there are challenges in identifying the most appropriate models for this application. The relationship between molecular descriptors and corrosion inhibition properties is often nonlinear and influenced by various factors, making it difficult for simple models to capture the full complexity of the data. Therefore, this research compares the performance of three distinct machine learning models: linear regression (a linear model), decision tree regression (a nonlinear model), and gradient boosting regression (an ensemble model). By evaluating the strengths and limitations of each model, this study aims to determine which approach is best suited for predicting CIE based on QSAR descriptors. The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology guided the research process [4], ensuring a structured and repeatable approach to data preparation, model development, and evaluation. This methodology is particularly well suited for machine learning research, emphasizing a cyclical process of understanding the data, building models, and validating results.

2. METHOD

2.1 CRISP-DM Framework

This research employs the Cross Industry Standard Process for Data Mining (CRISP-DM), a widely adopted framework in data science in Figure 1, to systematically address the task of predicting CIE using ML models. CRISP DM ensures a structured approach, guiding the study from understanding the problem to building and evaluating models. The process is divided into six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment, described below.

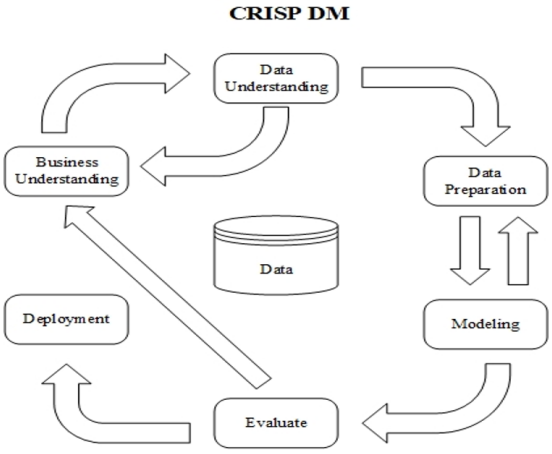

Figure 1. CRISP-DM Framework

2.2 Business Understanding

Corrosion is a significant challenge across multiple industries, including oil and gas, manufacturing, and marine transportation [5]. The deterioration of metals due to corrosion leads to costly repairs, replacements, and safety risks, causing economic losses that exceed billions of dollars annually. Failure to adequately manage corrosion can also lead to catastrophic consequences, such as pipeline leaks, equipment malfunctions, and structural collapses. For these reasons, developing and deploying corrosion inhibitors have become essential for maintaining industrial safety and longevity. Corrosion inhibitors are chemical compounds that reduce or prevent corrosion, primarily by forming a protective layer over the metal surface. Among the various inhibitors, organic compounds, particularly those derived from pharmaceuticals, have gained attention due to their eco-friendliness and effectiveness. These inhibitors work at the molecular level, adsorbing onto metal surfaces to create a barrier against corrosive agents like water, acids, or salts. However, selecting and designing the most effective inhibitors remains a resource-intensive task, as developing these compounds often requires extensive laboratory testing, which is both time-consuming and expensive.

CIE is a critical metric in determining the protective capacity of a compound. Predicting CIE accurately during the early stages of corrosion inhibitor development can significantly accelerate the identification of potential candidates, reducing the time and costs associated with traditional testing methods. This is where computational models, specifically machine learning, offer a transformative approach. ML models can analyse complex relationships between the molecular structure of a compound and its corrosion inhibition performance, offering the ability to predict CIE based on molecular descriptors. QSAR descriptors are well-established in fields such as drug design and computational chemistry. They are molecular properties, often derived from quantum chemical calculations, that describe how the structure of a molecule influences its behaviour. In corrosion science, these QSAR descriptors can be used as inputs to machine learning models to predict a molecule's CIE, effectively reducing reliance on physical experiments. Despite the apparent potential of machine learning for predicting CIE, one major challenge is identifying which model best captures the complex, often nonlinear relationships between molecular descriptors and corrosion inhibition properties. The proper model selection is critical because a poorly chosen model could yield inaccurate predictions, leading to suboptimal inhibitor selection.

Thus, this study aims to determine which ML model (linear regression, decision tree regression, or gradient boosting regression) is most suited for predicting corrosion inhibition efficiency based on QSAR descriptors. These models vary in their ability to handle linear and nonlinear relationships, and understanding their performance on this dataset will help develop more accurate, data-driven approaches to corrosion inhibitor design. An effective predictive model can accelerate the development process, guiding the selection of more efficient corrosion inhibitors with higher reliability and cost efficiency.

2.3 Data Understanding

The dataset used in this study consists of 14 molecular descriptors, including both QSAR and quantum chemical descriptors, which are key features influencing the inhibition efficiency of molecules. These descriptors capture various electronic and structural properties of molecules that are critical for their ability to prevent corrosion. The dataset was inspired by previous studies on drug molecules' corrosion inhibition efficiency and contains empirical and theoretical descriptors derived from quantum mechanical calculations.

The target variable for our prediction model is CIE, a crucial parameter in computational chemistry and drug design that affects the reactivity of molecules and their interaction with biological targets. The descriptors used in this study include Molecular Weight (g/mol), an essential molecular property, Acid Dissociation Constant (pKa), A measure of the strength of an acid, Octanol Water Partition Coefficient (Log P), Indicates the hydrophobicity of the molecule. Water Solubility (Log S): Represents the molecule's solubility in water. Polar Surface Area (Å^2): The surface area occupied by polar atoms in the molecule. Polarizability (Å^3): The molecule's ability to be polarized by an external electric field. The energy of HOMO (Highest Occupied Molecular Orbital, eV): Indicates the molecule's electron-donating ability. The energy of LUMO (Lowest Unoccupied Molecular Orbital, eV): Indicates the molecule's electron-accepting ability. Electron Affinity (eV): The energy changes when an electron is added to the molecule. Electronegativity (eV): A measure of the tendency of an atom to attract electrons. Hardness (eV): A property that reflects the resistance to electron transfer. Electrophilicity (eV): Indicates the ability of a molecule to accept electrons. Fraction of Electrons Shared (ΔN_Fe): A descriptor derived from quantum chemical calculations based on the Hard and Soft Acids and Bases (HSAB) theory [8], which is used to model the interaction between the inhibitor and the metal surface. These features are chosen for their ability to reflect the molecular interactions and electronic properties that influence CIE.

2.4 Data Preparation

Data preparation is vital in ensuring that the dataset is suitable for ML model development. This study performed two key steps during data preparation: handling missing values and feature scaling.

2.4.1 Handling Missing Value

The first step involved identifying and managing any missing or null values in the dataset. Since most machine learning algorithms cannot effectively handle missing data, rows with null values were removed. This approach ensures that the models are trained on complete, consistent data, reducing the risk of errors during the training process. Although this method can result in the loss of some data, it was considered necessary to maintain the quality of the model's input [9].

2.4.2 Feature Scaling

After addressing the missing values, the next step was standardizing the features. ML models, particularly those that rely on gradient-based optimization (such as Gradient Boosting), can be sensitive to the scale of the input features. If features have different ranges or units, the model may give disproportionate importance to certain variables, leading to suboptimal results. To avoid this issue, StandardScaler was applied to normalize the dataset. StandardScaler transforms each feature with a mean of 0 and a standard deviation of 1. This ensures all features are on a similar scale, allowing the models to process the input data more effectively. Feature scaling is essential in models where the magnitude of the data can influence learning, as it helps improve convergence rates and model performance [10].

$$z = (x - \mu) / \sigma \quad (1)$$

z	Standardized value
x	Original value
μ	Mean of the feature.
σ	The standard deviation of the feature

Table 1. Meanings of the symbols from StandardScaler

2.5 Data Splitting

Once the data was pre-processed, it was split into two subsets: 90% was used for training, and 10% was reserved for testing. This split ensures that the model is trained on sufficient data while maintaining an independent test set for evaluating the model's performance, with a ratio of 90/10 split balances that provide enough data for learning and preserving data for model evaluation [11].

2.6 Modelling

This study applied three machine learning models to predict pharmaceutical molecules' inhibition efficiency (CIE) based on 14 QSAR descriptors: linear regression, decision tree, and gradient boosting regression. Each model represents a different approach to capturing relationships between the molecular descriptors and the target variable (CIE), ranging from superficial linear relationships to complex nonlinear interactions [12].

2.6.1 Linear Regression

Linear regression is a statistical method that models the relationship between one or more independent variables (the QSAR descriptors) and a dependent variable (CIE) by fitting a linear equation to the observed data. The linear regression model assumes a direct, proportional relationship between the features and the target variable. Linear regression is arguably one of the most commonly used methods for statistical analysis [13].

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2)$$

Table 2 Meanings of the symbols from Linear Regression

$\hat{y}$	the predicted value of the target variable (CIE),
β_0	intercept (the predicted value when all input features are zero),
$\beta_1 + \beta_2 + \cdots + \beta_n$	coefficients (weights) for each feature $x_1 + x_2 + \cdots + x_n$
$x_1 + x_2 + \cdots + x_n$	input features (14 QSAR descriptors).

Linear regression serves as the baseline model. It captures only the linear relationships between the molecular descriptors and CIE. Its simplicity makes it easy to interpret but limits its ability to model complex, nonlinear relationships often present in chemical and physical data.

2.6.2 Decision Tree Regressor

A decision tree regressor is a nonlinear model that uses a tree-like structure to make predictions. It recursively splits the dataset into smaller subsets based on feature values, forming branches that represent decision rules. The model chooses splits that minimize prediction error at each node, and the leaves of the tree represent the final prediction values. Decision trees can capture complex nonlinear relationships in the data. However, they tend to overfit the training data, especially if the tree is too deep, meaning they perform well on the training data but poorly on unseen data [14]. The decision tree model works by following the rule:

$$\mathcal{Y} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \mathcal{Y}i \quad (3)$$

$\mathcal{Y}$	predicted value (CIE) at a given leaf node
N	number of data points in that node
$\mathcal{Y}i$	actual CIE values of the data points in that node

Table 3. Meanings of the symbols from the Decision Tree Regressor

2.6.3 Gradient Boosting Regressor

Gradient Boosting Regressor for its strength in combining multiple weak learners to form a robust model, with each new tree attempting to correct the errors made by the previous one. Gradient Boosting models are particularly effective at capturing complex, nonlinear relationships in the data, making them suitable for this study where the relationship between QSAR descriptors and CIE is non-linear [15]. The prediction in Gradient Boosting is calculated as follows:

$$\hat{y} = \sum_{m=1}^{M} \alpha_m h_m(x) \quad (4)$$

$\hat{y}$	final predicted value (CIE),
M	number of boosting rounds,
α_m	the weight assigned to the $weak$ learner (decision tree),
$h_m(x)$	prediction made by the m weak learner for the input features (x)

Table 4 Meanings of the symbols from Gradient Boosting Regressor

Using the gradient descent algorithm, each new tree is trained to minimize residual error from the previous trees. This residual error r_i for each data point is defined as:

$$r_i = y_i - \hat{y}i \quad (5)$$

y_i	actual target value (CIE),
$\hat{y}i$	predicted value from the model at the current iteration

Table 5 Meanings of the symbols from Residual Error

Gradient Boosting can progressively improve the accuracy of its predictions by iteratively adjusting for the residuals. This ensemble approach allows the model to handle more complex interactions between the molecular descriptors and CIE, leading to better overall performance [16].

2.7 Evaluation

Several evaluation metrics were used to assess the performance of the machine learning models: linear regression, Decision Tree Regression, and Gradient Boosting Regression. These metrics provide insight into how well each model predicts the Inhibition Efficiency (CIE) based on the QSAR descriptors. The evaluation was conducted on the test dataset (10%), which was not used during model training [17].

2.7.1 MSE

Mean Squared Error (MSE) is one of the most commonly used loss functions in regression problems. It measures the average squared difference between actual and predicted values. A lower MSE indicates better model performance, implying that the predictions are closer to the exact values [18].

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}i)^2 \quad (6)$$

y_i	actual value of CIE for the i - th data point
$\hat{y}i$	predicted value of CIE for the i - th data point
n	total number of data points

Table 6 Meanings of the symbols from MSE

Due to the squared term, MSE penalizes more significant errors than smaller ones, making it sensitive to outliers. For this reason, it's essential to interpret MSE in conjunction with other metrics.

2.7.2 MAE

Mean Absolute Error (MAE) measures the average absolute difference between actual and predicted values [19]. Unlike MSE, MAE does not square the differences, giving equal weight to all errors. This makes MAE more robust to outliers than MSE [20].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}i| \quad (7)$$

$\lvert y_i - \hat{y}i\rvert$	the absolute difference between the actual and predicted CIE for the i data point
n	total number of data points

Table 7 Meanings of the symbols from MAE

2.7.3 RMSE

Root Mean Squared Error (RMSE) is similar to MSE but takes the square root of the average squared differences [21]. This returns the error to the same unit as the target variable (CIE), making it easier to interpret. Like MSE, RMSE is sensitive to outliers and penalizes more significant errors more heavily [22].

$$RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}i)^2} \quad (8)$$

$\sqrt{\ }$	square root function
n	number of data points

Table 8 Meanings of the symbols from RMSE

Since RMSE is in the same units as CIE, it provides a clearer sense of the typical magnitude of the model's errors. However, because of its sensitivity to significant errors, RMSE is often used alongside MAE for a more balanced evaluation.

2.7.4 R^2

The Coefficient of Determination, commonly referred to as R^2, is a metric that explains how much of the variability in the target variable (CIE) can be explained by the model. It ranges from 0 to 1, and a value closer to 0 means that the model fails to capture the underlying patterns in the data [23], [24].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}i)^2} \quad (9)$$

y_i	actual value of CIE for the i - th data point
$\hat{y}i$	predicted value of CIE for the i - th data point
$\bar{y}i$	mean of the actual CIE values
n	Total number of data points.

Table 9 Meanings of the symbols from R2

R^2 value of 1 means that the model perfectly predicts the data, while an R^2 value of 0 means that the model performs no better than simply predicting the mean of the data.

2.8 Deployment

To make the predictive model accessible and user-friendly, the best-performing model, Gradient Boosting Regressor, was deployed using Streamlit, a popular framework for building web applications. The application allows users to input molecular descriptors (the 14 QSAR features) and receive real-time predictions for

Inhibition Efficiency (CIE). The features must be standardized before the user inputs are fed into the predictive model. During the model development process, StandardScaler was applied to ensure that all input features were on the same scale. Therefore, the same scaling transformation is applied to the user inputs during deployment to maintain consistency with the training phase [25]. Once the input features are standardized, they are passed to the Gradient Boosting Regressor model to predict the Inhibition Efficiency (CIE). The model processes the input and provides a real-time prediction

3. RESULT AND DISCUSSION (10 PT)

The performance of the three regression models, Linear Regression, Decision Tree Regressor, and Gradient Boosting Regressor, was assessed using multiple evaluation metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).

Model	MSE	MAE	RMSE	R2
Linear Regression	23.845986	4.325123	4.883235	0.126233
Decision Tree Regressor	25.081786	4.364286	5.008172	0.080951
Gradient Boosting Regressor	21.521361	3.817764	4.639112	0.211412

Table 10 Model Performance

Table 10 presents quantitative metrics MSE, MAE, RMSE, and R^2 that comprehensively evaluate each model's performance. As discussed, these metrics highlight the Gradient Boosting Regressor's superior ability to minimize average error (MSE and MAE) and significant deviations (RMSE). Lower values of MSE and MAE for the Gradient Boosting Regressor suggest that its overall accuracy is higher, with fewer significant errors compared to Linear Regression and Decision Tree Regressor. Notably, the Decision Tree Regressor has the highest MSE and MAE, indicating that it frequently produces more significant errors, further underscoring its overfitting problem. RMSE values reflect the Gradient Boosting Regressor's effectiveness at minimizing significant prediction errors. The higher RMSE values for Linear Regression and Decision Tree Regressor suggest that these models struggle with extreme values, producing more significant outliers in their predictions. This also reinforces the insights gained from the scatter plots and residual analysis, where both models exhibited higher variance in their predictions. R^2 values, which measure the proportion of variance in the target variable explained by the model, provide additional evidence of the Gradient Boosting Regressor's superior performance. An R^2 of 0.21 explains a more significant portion of the variance than Linear Regression (0.13) and Decision Tree Regressor (0.08). This low R^2 score across all models suggests that there are still unexplained patterns in the data, possibly due to nonlinear relationships or unaccounted factors. The low R^2 for the Decision Tree Regressor, in particular, highlights its tendency to overfit without providing meaningful generalization to new data.

Figure 2 represents how closely each model's predictions align with the actual values of the target variable. In a perfect model, the predicted values would fall directly on the ideal prediction line, showing no deviation from the actuals. For the Gradient Boosting Regressor, the scatter plot shows a tight clustering around the perfect line, particularly for most predicted values. This confirms that this model performs better in minimizing prediction errors. The Gradient Boosting model's iterative approach, combining weak learners and focusing on correcting residual errors, allows it to better model the data's underlying patterns. Its capability to handle non-linearity in the data helps it avoid the significant prediction errors seen in other models. In contrast, the scatter plots for the Linear Regression and Decision Tree Regressor models show wider spreads, with many data points deviating substantially from the ideal line, especially in the case of the Decision Tree Regressor. This further supports the conclusion that these models are struggling to generalize well. Linear Regression, while simple and interpretable, is limited by its assumption of linearity and inability to capture complex relationships in the data. The Decision Tree Regressor's scatter plot reveals significant deviation from actuals, indicating its overfitting to the training data and inability to handle the data's noise or complex interactions.

Figure 3 Representation The Gradient Boosting Regressor's R^2 score of 0.21, while modest, indicates its superiority over the Linear Regression (R^2 of 0.13) and Decision Tree Regressor (R^2 of 0.08). This performance advantage can be attributed to the Gradient Boosting model's ability to iteratively improve its predictions by correcting the residual errors of its weak learners, which contrasts with the simplistic, one-shot approach of Linear Regression and the local decision-making of Decision Trees. However, the relatively low R^2 scores across all models suggest that the data may be inherently noisy or complex and that a large portion of the variance is either unexplained or cannot be easily captured by the models. This indicates potential for further data exploration, feature engineering, or adopting more sophisticated models such as Random Forest or neural networks. The Decision Tree Regressor's poor R^2 score, combined with the evidence of overfitting, points to a fundamental weakness in the model's structure. While Decision Trees excel at capturing nonlinear interactions, they are prone to overfitting unless regularized or pruned. The poor generalization ability observed

 DOI: 10.30998/faktorexacta.v17i3.24679

here implies that the Decision Tree is overfitting to the training data, leading to poor performance on unseen data, further highlighting the importance of more flexible ensemble methods like Gradient Boosting in such tasks.

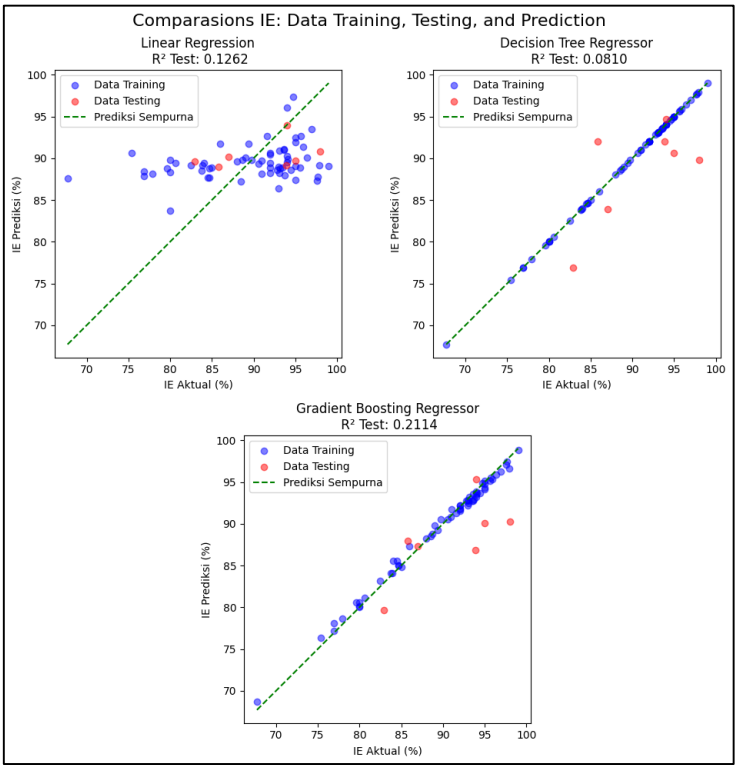

Figure 1 Comparison IE Data Training, Testing, and Prediction

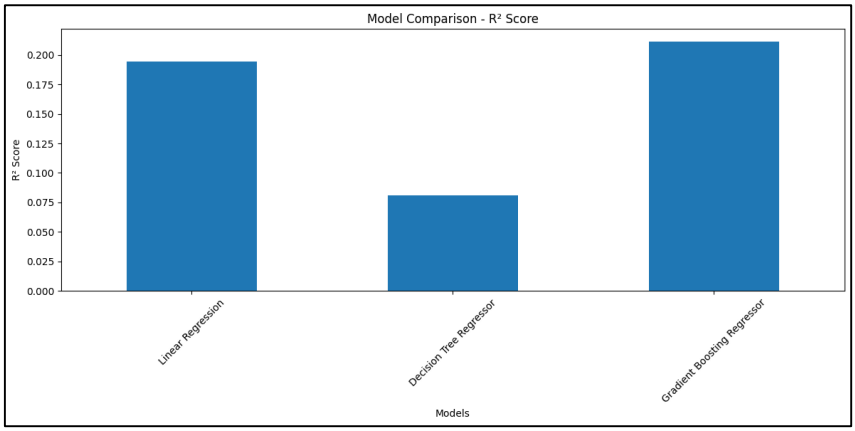

Figure 2 Model Accuracy Comparison

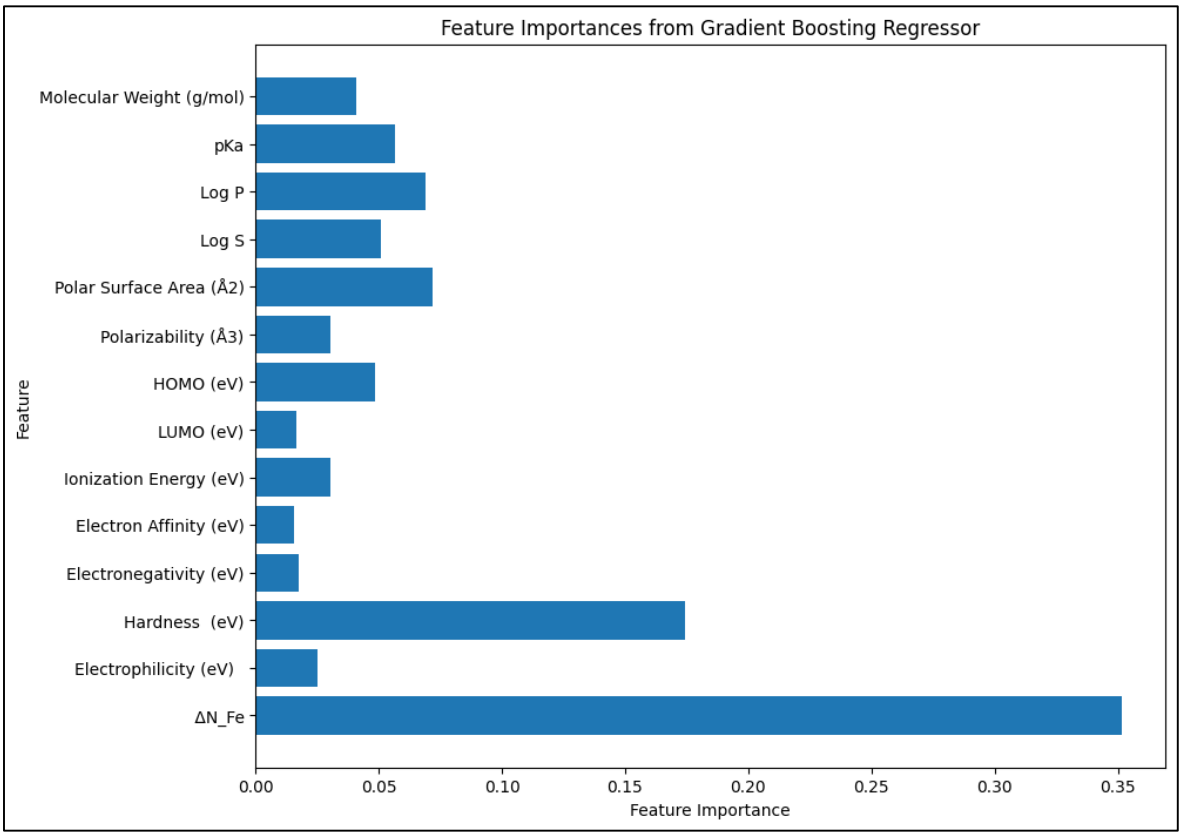

Figure 4 Feature importance

Figure 4 indicates a concentration of predictive power among a few key features. Features (ΔN_Fe) and Hardness (eV), as noted, dominate the feature importance rankings, while others contribute minimally to the model's predictive accuracy. This imbalance presents both opportunities and challenges for model optimization. From an analytical perspective, these highly ranked features are likely to capture core aspects of the target variable, whether directly through strong correlations or indirectly by serving as proxies for other complex interactions. These features might reflect significant environmental, behavioural, or system-level factors driving the predictions. A deeper understanding of the nature of these features is essential. If Feature (ΔN_Fe) represents a critical measurement like solar irradiance or temperature in a time series forecasting task, it would make sense that such a variable has significant predictive power. However, further analysis is needed to determine whether these dominant features represent actual critical drivers of the target variable or if they merely contain redundant information that correlates strongly with the target due to noise or overfitting. On the other hand, the negligible importance of certain features, such as Feature Electron Affinity (eV) and Feature Electronegativity (eV), may point to either a lack of variability in these features or an absence of helpful signals for prediction purposes. It is worth exploring if these features could be transformed, combined, or eliminated in future model iterations, potentially reducing model complexity and improving training efficiency. In some cases, though, minimal importance does not necessarily mean the feature has no value. It may provide a small but critical amount of information that could contribute to further refinement in susceptible domains.

4. CONCLUSION (10 PT)

This study successfully explored the application of ML models to predict the CIE of corrosion inhibitors using QSAR descriptors. The research aimed to streamline the corrosion inhibitor development process by employing computational models to reduce the need for labor-intensive laboratory testing. The results from the Gradient Boosting Regressor demonstrated a promising step toward achieving this objective, as the model effectively captured non-linear relationships inherent in chemical interactions. This suggests that machine learning can provide valuable preliminary insights into molecular characteristics that drive corrosion inhibition, offering a more efficient path for early-stage inhibitor selection. However, the relatively low R^2 values across all models indicate that the predictive power remains limited, pointing to unresolved complexities in the dataset. These findings highlight that while machine learning presents a viable approach, it currently falls short

of providing highly accurate predictions needed for practical application. The inability of the models to fully explain the variance in IE% suggests that critical factors influencing corrosion inhibition are either missing from the feature set or not sufficiently represented by the chosen models. This underscores the need for further refinements, such as incorporating additional molecular descriptors, exploring alternative computational techniques like Random Forests or deep learning, and improving data quality. These improvements could enhance the models' ability to capture the intricate chemical and physical interactions governing inhibition efficiency. This research contributes a foundational framework for using machine learning in corrosion science, moving toward a faster, cost-effective approach. However, realizing its full potential will require overcoming current limitations in prediction accuracy. Future work should focus on refining model performance to meet better the ultimate goal of reducing reliance on exhaustive experimental procedures in developing corrosion inhibitors.

REFERENCES

[1] M. Akrom, S. Rustad, and H. K. Dipojono, A machine learning approach to predict corrosion inhibition efficiency by natural product-based organic inhibitors, Phys. Scr., 2024, doi: 10.1088/1402-4896/ad28a9.

[2] C. Beltran-Perez et al., 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., vol. 23, no. 9, May 2022, doi: 10.3390/ijms23095086.

[3] A. Cherkasov, E. N. Muratov, D. Fourches, QSAR modeling: where have you been? Where are you going?, J. Med. Chem., 2014, doi: 10.1021/jm4004285.

[4] C. Shearer, The CRISP-DM model: the new blueprint for data mining, J. Data Warehous., 2000.

[5] M. Finšgar and J. Jackson, Application of corrosion inhibitors for steels in acidic media for the oil and gas industry: A review, Corros. Sci., 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010938X14002157

[6] M. Karelson, V. S. Lobanov, and A. R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, Chem. Rev., 1996, doi: 10.1021/cr950202r.

[7] D. C. Ghosh and R. Biswas, Theoretical calculation of absolute radii of atoms and ions. Part 1. The atomic radii, Int. J. Mol. Sci., 2002. [Online]. Available: https://www.mdpi.com/1422-0067/3/2/87

[8] R. T. V. Consonni, Handbook of Molecular Descriptors, vol. 11, John Wiley & Sons, New York, NY, 2009.

[9] Z. S. Priyambudi and Y. S. Nugroho, Which algorithm is better? An implementation of normalization to predict student performance, AIP Conf. Proc., 2024. [Online]. Available: https://pubs.aip.org/aip/acp/article-abstract/2926/1/020110/2999314

[10] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Stanford, 2009.

[11] H. Campbell, Equivalence testing for linear regression, Psychological Methods, *29*(1), 88–98 (2024), https://doi.org/10.1037/met0000596

[12] M. Bhaiyya, D. Panigrahi, P. Rewatkar, and H. Haick, Role of machine learning assisted biosensors in point-of-care testing for clinical decisions, ACS Sens., 2024, doi: 10.1021/acssensors.4c01582.

[13] C. Guestrin and T. Chen, XGBoost: A scalable tree boosting system, in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016.

[14] A. Jadon, A. Patil, and S. Jadon, A comprehensive survey of regression-based loss functions for time series forecasting, arXiv Preprint arXiv:2211.02989, 2022.

[15] G. Vasconcelos, M. B. Francisco, Prediction of surface roughness in duplex stainless steel face milling using artificial neural network, Int. J. Adv. Manuf. Technol., 2024, doi: 10.1007/s00170-024-13955-4.

[16] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, Comparative analysis of machine learning techniques for predicting air quality in smart cities, IEEE Access, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8746201/

[17] K. E. Lotterhos, Principles in experimental design for evaluating genomic forecasts, Methods Ecol. Evol., 2024, doi: 10.1111/2041-210X.14379.

[18] Y. Akkem, B. S. Kumar, and A. Varanasi, Streamlit application for advanced ensemble learning methods in crop recommendation systems–a review and implementation, Indian J. Sci. Technol., 2023. [Online]. Available: https://sciresol.s3.us-east-2.amazonaws.com/IJST/Articles/2023/Issue-48/IJST-2023-2850.pdf

[19] M. Akrom, S. Rustad, and H. K. Dipojono, SMILES-based machine learning enables the prediction of corrosion inhibition capacity, MRS Commun., vol. 14, pp. 379–387, 2024, doi: 10.1557/s43579-024-00551-6.

[20] M. Akrom, DFT investigation of Syzygium aromaticum and Nicotiana tabacum extracts as corrosion inhibitors, Sci. Tech. J. Ilmu Pengetahuan Teknol., vol. 8, no. 1, pp. 42-48, 2022.
[21] 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., vol. 1229, p. 114307, 2023.
[22] 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., vol. 6, p. 101126, 2023.
[23] M. Akrom and T. Sutojo, Investigasi model machine learning berbasis QSPR pada inhibitor korosi pirimidin, Eksergi, vol. 20, no. 1, 2023.
[24] 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 Eng., pp. 78-87, 2024.
[25] 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 Eng., pp. 68-77, 2024.