Comparison of Linear and Non-Linear Machine Learning Algorithms for Predicting the Effectiveness of Plant Extracts as Corrosion Inhibitors

Yudha Mulyana[1], Muhamad Akrom[2], Gustina Alfa Trisnapradika[3]
[1,2,3]Informatics Engineering Program, Dian Nuswantoro University, Semarang, Indonesia
[1]yudha.myn@gmail.com
[2]m.akrom@dsn.dinus.ac.id
[3]gustina.alfa@gmaail.com

Accepted 31 May 2024
Approved 18 June 2024

***Abstract*— This research aims to develop a Machine Learning (ML) model to accurately predict the corrosion inhibition potential of plant extracts. The ML model development involves data normalization, selecting both linear and non-linear ML algorithms, model training with k-fold cross-validation, and performance evaluation using regression metrics such as MSE, RMSE, MAE, and R^2. Various ML algorithms were compared, with the AdaBoost Regressor (ABR) model achieving the best performance, marked by an R^2 value of 0.993 and a low MSE of 0.002. These results highlight the potential of ML models in predicting effective corrosion inhibitors from plant extracts. Moreover, feature importance analysis reveals that two features, concentration (Conc) and LUMO, significantly influence the ABR model. This research contributes significantly to developing effective prediction methods in the corrosion control industry.**

Keywords— Machine learning; Corrosion; Plant extract; Adaboost regressor, Linear Regression.

I. Introduction

Corrosion is a process of material degradation or decay caused by a chemical reaction between the metal and the environment where many corrosive substances cause surrounding corrosion [1]. The corrosion process involves the oxidation of metals by oxygen in air or other corrosive substances, which results in corrosion products such as oxides, hydroxides, or metal salts [2]. These corrosion reactions can affect material quality and performance, reduce service life, and cause significant economic losses. Some of the factors that affect the corrosion rate include the type of metal involved, the nature of the corrosive environment such as humidity, pH, temperature, the concentration of corrosive substances, as well as other factors such as mechanical stress or friction [3]. The corrosion process can also be accelerated by the presence of galvanic (contact between two dissimilar metals in an electrolyte), interaction by microorganisms such as bacteria, or stress-induced corrosion [4]. The study of corrosion involves an in-depth understanding of corrosion mechanisms, development of corrosion control methods, and evaluation of material performance in corrosive environments [5]. Controlling corrosion processes is essential in various industries such as the oil and gas industry, chemical industry, automotive industry, and construction to ensure the sustainability and safety of the materials used [6]. Research in the field of corrosion inhibitors continues to grow, especially in the exploration of organic inhibitors. This is due to the negative properties possessed by inorganic inhibitors, such as toxicity, environmental unfriendliness, and high production costs [7]. One type of organic inhibitor that is increasingly in demand is natural plant extract-based corrosion inhibitors, which are often referred to as "green inhibitors" [8]. Green inhibitors are highly valued because they are environmentally friendly, easily degradable, renewable, do not pollute with toxins or pollutants, easy to produce, low cost, and have high anticorrosive efficiency [9]. Natural plant extracts contain natural compounds that play an important role as corrosion inhibitors [10]. The structure of compounds in natural plant extracts often contains heteroatom groups such as Oxygen (O), Nitrogen (N), Sulfur (S), Phosphorus (P), and also aromatic rings. This combination of structures is considered to be the most efficient as a corrosion inhibitor [7]. Experimentally studying the use of plant extracts as corrosion inhibitors is time-, cost-, and resource-intensive.

To address this gap, machine learning (ML) approaches based on quantitative structure-property relationship (QSPR) models have recently been used in the investigation and exploration of new anti-corrosion materials. Given the relationship between the structure and molecular properties of compounds, QSPR is an effective and reliable method [7]. Lu Li et al. [8] used an ML algorithm, namely a support vector machine (SVM), to investigate benzimidazole compounds as corrosion inhibitors. The results showed that the SVM model had a coefficient of determination (R^2) of 0.96 and a root mean square error (RMSE) of 6.79. In addition, Akrom et al. [9] compared several ML models on a dataset of pyrimidine compounds, and found that the gradient boosting regressor (GBR) model had the best accuracy with an R^2 value of 0.92 and an RMSE of 0.95, compared to the support vector regression (SVR) and k-nearest neighbor (KNN) models.

In this research, the focus is on the use of plant extracts as an alternative corrosion inhibitor. By utilizing the natural compounds contained in plant extracts, it is expected to develop a machine learning (ML) model that can predict corrosion inhibition efficiency with high accuracy [7]. The implementation of data normalization techniques during preprocessing and the use of k-fold cross-validation techniques in the development of ML models are expected to improve the accuracy of predicting corrosion inhibition efficiency [11]. This research makes an important contribution to the development of machine-learning models for designing potential corrosion inhibitor compounds sourced from plant extracts. It can serve as a reference for other researchers in developing more accurate machine learning models to predict corrosion inhibition efficiency by utilizing environmentally friendly natural resources [12].

II. METHOD

Figure 1 illustrates the process of developing the Machine Learning (ML) model in this study. The initial stage is the selection of a plant extract dataset, followed by data normalization to address scale differences and sensitivity to outliers. [9]. After that, ML algorithms, both linear and non-linear, are selected to model the relationship between input and output variables that can accurately predict the corrosion inhibition efficiency of plant extracts [1]. Furthermore, the models are trained using the k-fold cross-validation technique, which helps avoid overfitting and obtain a more generalized model [13]. Model performance evaluation is performed using regression metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R2) [14]. These stages not only help in the selection of the optimal ML models but also ensure that the model can produce accurate predictions regarding the corrosion inhibition efficiency of the plant extracts in this study [4].

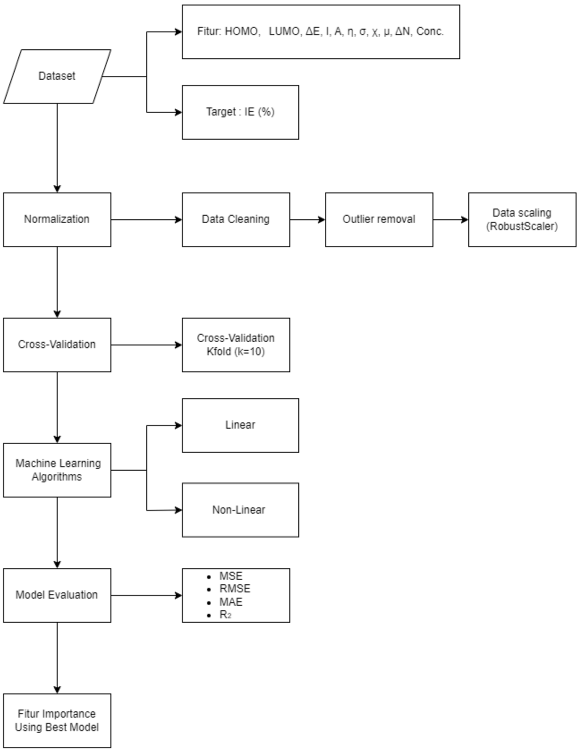
Fitur: HOMO, LUMO, ΔE, I, A, η, σ, χ, μ, ΔN, Conc.
Dataset
Target : IE (%)
Normalization
Data Cleaning
Outlier removal
Data scaling (RobustScaler)
Cross-Validation
Cross-Validation Kfold (k=10)
Linear
Machine Learning Algorithms
Non-Linear
Model Evaluation
MSE
RMSE
MAE
R_2
Fitur Importance Using Best Model

Fig 1. ML model development method

[1] Dataset

The dataset used in this study consists of plant extracts derived from published literature written by Akrom, et al [15]. This dataset includes specific plant extracts that have 12 features, namely HOMO, LUMO, ΔE, I, A, η, σ, χ, μ, ΔN, Conc., and IE (%). These features reflect the molecular properties and physicochemical characteristics of the plant extract [15]. Some of these features are related to the electronic properties, polarity, and corrosion inhibition efficiency (IE (%)) as dependent variables which are the main focus of the analysis and prediction in this study [14].

[2] Normalization

In the course of this research, the researcher adopted a very comprehensive approach to preprocessing the plant extract dataset used. A series of important steps were applied to ensure the quality and reliability of the processed data [16]. First of all, we performed data normalization using the RobustScaler technique. This was done to address the sensitivity to the presence of outliers or extreme data that could affect model performance. Data normalization is a crucial step in ensuring that the data used has a uniform scale, allowing machine learning models to provide consistent and accurate results [17]. In addition, researchers also remove outliers using the Interquartile Range (IQR) method [18]. Outlier removal is done to rid the dataset of unusual or unrepresentative data that may affect the interpretation of the analysis results. By calculating the IQR, researchers can identify and eliminate extreme data that may negatively affect model performance.

[3] Data Cleaning

In the data cleaning stage, researchers normalize the data using RobustScaler to address sensitivity to outliers or extreme data that can affect model performance [21],[17]. The RobustScaler technique was chosen because it can handle differences in scale in the data and maintain the consistency of the prediction results. In addition, researchers also applied the Interquartile Range (IQR) method to remove outliers [18]. Outlier removal is necessary to rid the dataset of unusual or unrepresentative data that may affect the analysis and prediction results [18], [1]. This data-cleaning process aims to prepare a clean and consistent dataset for further analysis [22].

[4] Scaling Data

After the data cleaning process is complete, the data scaling process is carried out to standardize the range of values on each feature in the dataset. The goal is for all features to have a uniform scale so that the model can provide consistent and accurate results [23]. In this study, researchers used data scaling techniques to improve the performance of Machine Learning models in predicting the corrosion inhibition efficiency of plant extracts. The combination of data normalization and data scaling helps prepare a dataset of optimal quality for use in developing accurate and relevant ML models [24].

[5] Cross-Validation

We applied the k-fold cross-validation method with k = 10 to divide the data into 10 equal subsets. In each iteration, the model is trained on 9 subsets and tested on the remaining one, rotating through all subsets [19]. This process helps identify the model with the smallest error rate, ensuring robustness and reliability. We chose k = 10 to maximize data use and minimize bias and variance [20].

[6] ML Algoritma

These experiments involve a comparison between linear algorithms, which tend to use linear relationships between input and output variables, and non-linear algorithms, which can handle more complex and non-linear relationships between input and output variables [25]. The linear algorithms evaluated, such as multilinear regressor (MLR), ridge, lasso, Elastic-Net (EN), Support Vector Regression (SVR), and Generalized Linear Model (GLM), focus on models that have a linear relationship between input and output variables, while non-linear algorithms, such as random forest (RF), k-nearest neighbors (KNN), nu-support vector regressor (NuSVR), decision tree regressor (DT), gradient boosting regressor (GBR), orthogonal matching pursuit (OMP), kernel ridge (KR), partial least square (PLS), adaboost regressor (ABR), and bagging regressor (BR), can address more complex and non-linear relationships between input and output variables [14], [1], [4]. Through this comparison, the experiment aims to find the most optimal algorithm for predicting the corrosion inhibition efficiency of plant extracts. The results show that the best model for predicting the corrosion inhibition efficiency is the AdaBoost Regressor (ABR). ABR is a boosting algorithm used to improve the accuracy of predictive models. The prediction of the AdaBoost Regressor is based on a combination of multiple weak learners. Each weak learner contributes to the final prediction through a weighted vote, where the weight reflects the accuracy of the weak learner [32]. The final prediction $H(x)$ is computed as shown in the equation.

$$H(x) = \sum_{m=1}^{M} \alpha_m h_m(x)$$

With $H(x)$ being the final prediction of the ensemble model, α_m the weight of the m-th weak learner, $h_m(x)$ the prediction of the m-th weak learner for the input x, and M the total number of weak learners in the ensemble. The weights α_m are calculated based on the accuracy of each weak learner. Where,

H(x) : Final prediction function of the AdaBoost Regressor model

α_m: The coefficient or weight of the m-th base model

$h_m(x)$: Prediction of the m-th base model (weak learner)

m : Iteration index for the base model

M : Total number of base models used in boosting

[7] Model Evaluation

To evaluate the performance of the prediction model, various regression metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R2) are used [26]. The main objective is to select a model that has the minimum possible MSE, RMSE, and MAE value while approaching the ideal R2 value which is close to 1 [27]. This is important because it shows that the models have a high level of accuracy in predicting the corrosion inhibition efficiency of plant extracts [9], [26]. These metrics provide comprehensive information on the model's ability to estimate the potential value of corrosion inhibitors, which is an important step in the development of effective and relevant prediction methods for industrial applications, especially in the context of corrosion control [11].

[8] Important Features

The analysis of important features in this study is a critical step that requires an in-depth understanding of the relationship between the molecular and physicochemical characteristics of plant extracts and their corrosion inhibition efficiency [1], [13]. This method not only helps in identifying the most influential features in the prediction but also provides a deeper insight into the corrosion inhibitor mechanism of action of the plant extracts. Thus, the results of feature importance analysis provide a strong foundation in the selection of the most relevant features to be used in the construction of accurate and effective Machine Learning models [18]. The information obtained from feature importance analysis also has important implications in the context of industrial applications and further research [28]. The discovery of the most significant features in indicating corrosion inhibition effectiveness can be used to direct the development of more efficient and environmentally friendly corrosion inhibitor materials [11]. Furthermore, a deeper understanding of the relationship between the molecular properties of plant extracts and their performance as corrosion inhibitors provides opportunities for further research in the optimization of inhibitor formulations that can be widely applied in the corrosion industry [4]. Thus, the analysis of feature importance is not only an important part of this research but also makes a real contribution to the development of science and technology in the field of corrosion control.

III. Result and Discussion

The first step in this research was to test the plant extract dataset using the linear algorithm available in the scikit-learn library using Python. The performance of each model was measured using R2 and RMSE values as evaluation metrics. Table 1 and Table 2 below show the performance results of the linear and non-linear models respectively. These tables serve as the basis for evaluating the performance and prediction accuracy of each model.

Based on the analysis results in Table 1, it can be concluded that among the linear models evaluated, the Linear Regression (MLR) model shows superior prediction performance compared to the Ridge Regression, Support Vector Regression (SVR)**,** and Generalized Linear Model (GLM) models. This is supported by the highest R-squared (R2) value obtained by MLR of 0.626, as well as the lowest Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) values (0.236, 0.486, and 0.330 respectively). Figures 2 to 5 are data distributions in linear models that show that the predicted points from MLR are closer to the predicted line than other linear models. This indicates that the MLR model is the best model of similar models.

TABLE 1. Linear Model Performance

Model	**Linear Model Evaluation**			
	MSE	***RMSE***	***MAE***	***R_2***
Ridge	0.265	0.515	0.363	0.580
MLR	0.236	0.486	0.330	0.626
SVR	0.283	0.532	0.343	0.552
GLM	0.246	0.476	0.331	0.624

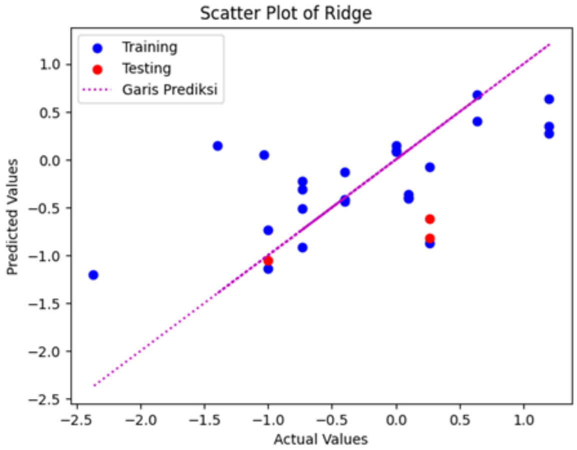

Fig. 2. Scatter Plot of Ridge

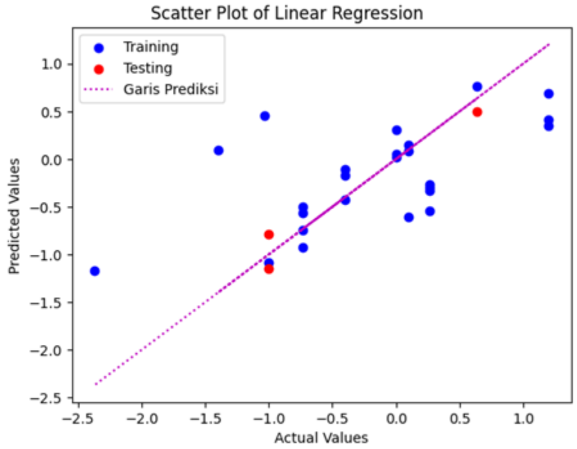

Fig. 3. Scatter Plot of MLR

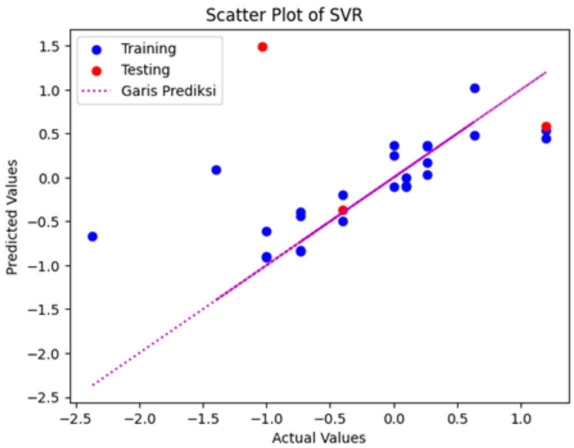

Fig. 4. Scatter Plot of SVR

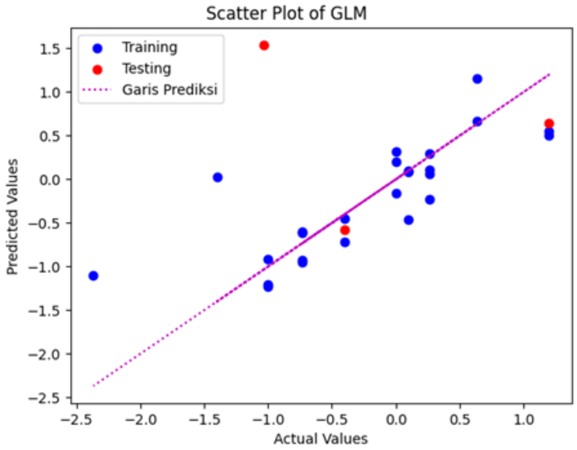

Fig. 5. Scatter Plot of GLM

For non-linear models, Table 2 presents model performances. The Adaboost Regressor (ABR) showed superior prediction performance when compared to the Random Forest (RF), Bagging Regressor (BR), and Gradient Boosting Regressor (GBR) based on the evaluation metrics used (R2 = 0.993, MSE = 0.002, RMSE = 0.053, and MAE = 0.034). The distribution of data points from the ABR model also showed a better fit to the prediction line. Thus, it can be concluded that overall, the Adaboost Regressor (ABR) model is the best choice in predicting the corrosion inhibitor potential values of plant extracts based on CIE values. Figures 6 to 9 are data distributions in non-linear models that show that the predicted points from ABR are closer to the predicted line than other linear models. This indicates that the ABR model is the best model of similar models.

TABLE 2. NON-LINEAR MODEL PERFORMANCE

Model	Non-Linear Model Evaluation			
	MSE	***RMSE***	***MAE***	***R_2***
RFR	0.030	0.173	0.114	0.934
GBR	0.041	0.203	0.129	0.934
ABR	0.002	0.053	0.034	0.993
BR	0.054	0.233	0.130	0.921

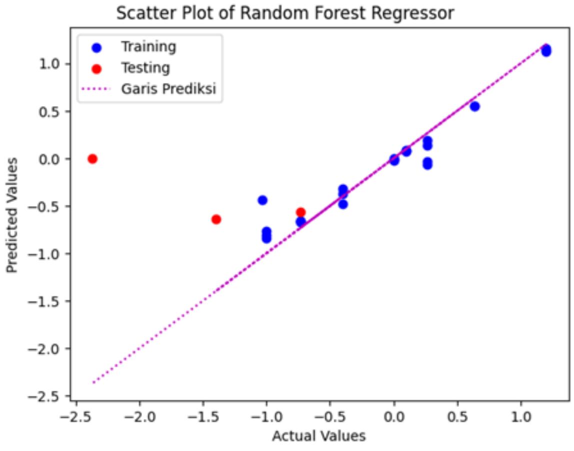

Fig. 6. Scatter Plot of RFR

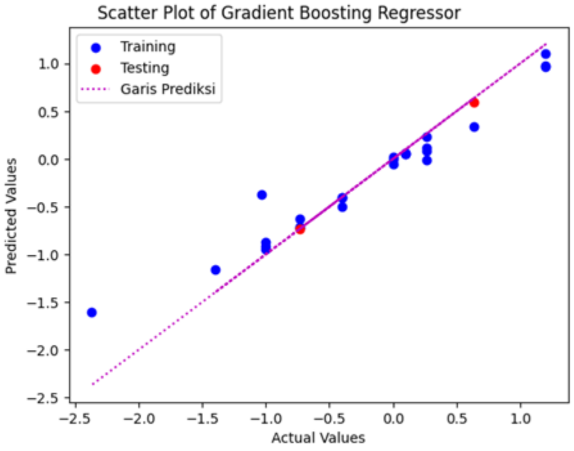

Fig. 7. Scatter Plot of GBR

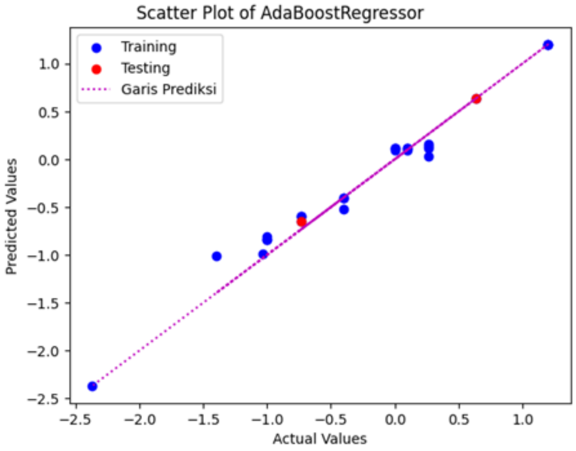

Fig. 8. Scatter Plot of ABR

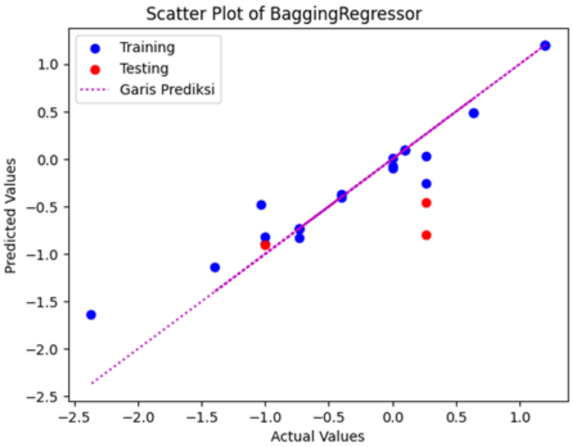

Fig. 9. Scatter Plot of BR

Figure 10 provides a clear picture of the performance comparison between Model Linear Regression (MLR) and AdaBoost Regressor (ABR) based on their R2 scores. From the results shown, it can be seen that the ABR model consistently achieves R2 scores close to 0.99 across all folds evaluated, demonstrating its ability to explain about 95% of the variability in the data consistently. On the other hand, the MLR model showed greater variation in R2 scores, with values ranging between 0.58 and 0.63. This analysis indicates that the ABR model has higher accuracy and better consistency in predicting plant extract efficiency than the MLR model. The ability of the ABR model to consistently approach R2 values close to 1 indicates that the model can describe the relationship between input and output variables with a high degree of accuracy. Therefore, based on this evaluation, it can be concluded that the ABR model is a better choice in this context to predict the corrosion inhibition efficiency of plant extracts. This is supported by the ABR model's ability to provide accurate and consistent predictions in explaining variations in the evaluation data used.

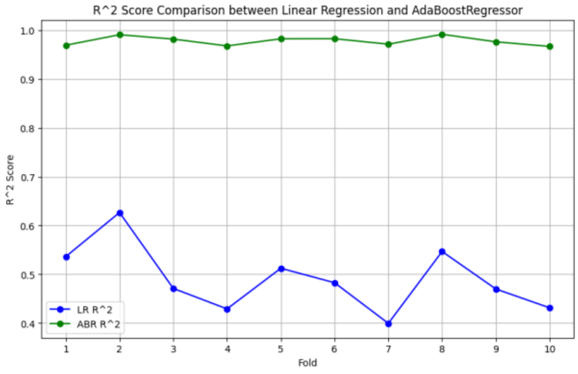

Fig. 10. Comparison of the best linear and best nonlinear algorithms for 10 experiments

From the residual analysis in Figure 11, it can be seen that most of the residual points from the ABR model are close to the zero line, indicating that the model has a low prediction error rate and can provide accurate estimates. In contrast, the residual error plot of the MLR model shows a larger spread and is not centered around the zero line, indicating that the MLR model has a higher error rate than the ABR model. Thus, based on this residual analysis, it can be concluded that the ABR model performs better in making predictions with a low error rate, compared to the MLR model. This corroborates the previous conclusion that the ABR model is a better choice for predicting the corrosion inhibition efficiency of plant extracts.

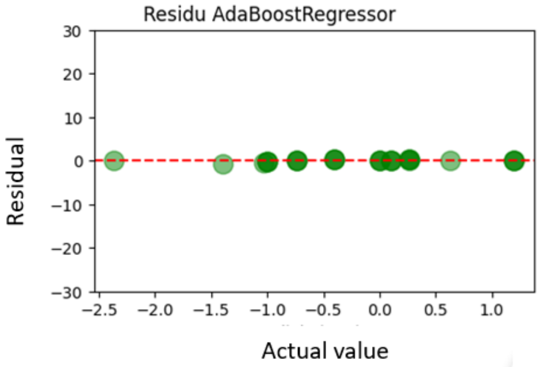

Fig. 11. Residual Error on ABR model

The selection of ABR as the superior model is also supported by feature importance analysis, which measures the extent to which a feature affects the algorithm's performance. Feature importance aims to reduce errors and eliminate noise in the dataset, thus providing more generalized and relevant results. In developing a predictive model, it is important to select input features that have a significant impact on the target variable. From the results of the critical feature analysis in Figure 12, it can be seen that two features have a significant impact on the ABR model, namely Conc and LUMO. This indicates that these features have a strong role in influencing the predictions of the model. The ABR model is strengthened by its ability to recognize complex patterns in the data, which may not be handled by the MLR model. Therefore, the nonlinear best model (ABR) is a better choice for predicting the corrosion inhibitor efficiency of plant extracts compared to the linear best model (MLR).

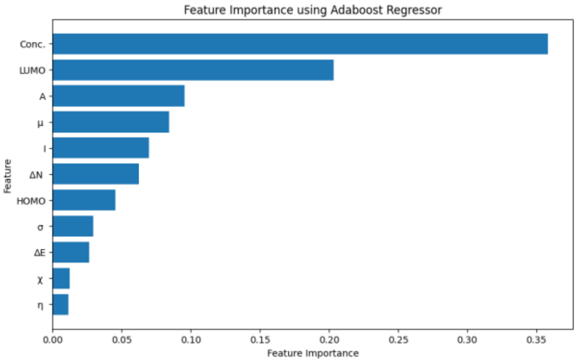

Fig. 12. Importance features of the Adaboost Regressor model

Based on Table 3 evaluation of linear (MLR) and non-linear (ABR) models, the ABR model shows better performance with lower MSE, RMSE, and MAE values and higher R^2 values. This indicates that the ABR model is more accurate and effective in predicting the data compared to the MLR model.

TABLE 3. EVALUATION OF THE BEST LINEAR AND NON-LINEAR MODELS

Evaluation	**Best Linear and Non-Linear Models**	
	MLR	***ABR***
MSE	0.236	0.002
RMSE	0.486	0.053
MAE	0.330	0.034
R2	0.626	0.992

IV. CONCLUSION

This study explores the best models for predicting the corrosion inhibitor potential of plant extracts using a Machine Learning (ML) approach by comparing the performance of linear and nonlinear models. The results showed that the nonlinear Adaboost Regressor (ABR) model was the most accurate, outperforming four linear models and three other nonlinear models. The ABR model achieved the highest R-squared (R^2) value of 0.992, indicating very high prediction accuracy. Additionally, the low Mean Squared Error (MSE) of 0.002, Root Mean Square Error (RMSE) of 0.053, and Mean Absolute Error (MAE) of 0.034 indicate minimal prediction error. These findings provide valuable insights for developing efficient material exploration methods in the corrosion control industry and can serve as a basis for designing more effective and environmentally friendly corrosion inhibitor materials.

ACKNOWLEDGMENT

The researchers would like to thank Allah SWT for His abundant grace and make it easy to complete this research. We also thank the Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Dian Nuswantoro University for the support and facilities provided during this research process. We would also like to express our deepest gratitude to our parents for their tireless prayers, support, and motivation during this research journey. The results of this research will be useful for the advancement of science and the welfare of mankind.

REFERENCE

[1] M. C. Adryan Putra Sumarjono, M. Akrom, and G. Alfa Trisnapradika, "Comparison of the Best Machine Learning Model to Predict Corrosion Inhibition Capability of Benzimidazole Compounds," 2023.

[2] S. Budi, M. Akrom, G.A. Trisnapradika, T. Sutojo, W.A.E. Prabowo, Optimization of Polynomial Functions on the NuSVR Algorithm Based on Machine Learning: Case Studies on Regression Datasets, *Scientific Journal of Informatics*, 10(2), (2023), https://doi.org/10.15294/sji.v10i2.43929.

[3] M. Akrom, S. Rustad, H.K. Dipojono, Development of Quantum Machine Learning to Evaluate the Corrosion Inhibition Capability of Pyrimidine Compounds, *Mater Today Commun*, 39, 108758 (2024), https://doi.org/10.1016/J.MTCOMM.2024.108758.

[4] N. V. Putranto, M. Akrom, and G. A. Trinapradika, "Implementation of Polynomial Function in Gradient Boosting Regressor Algorithm: Regression Study on Expired Medicines Dataset as Anticorrosion Material," J. Technol. and Manaj. Inform., vol. 9, no. 2, pp. 172-182, 2023, doi: 10.26905/jtmi.v9i2.11192.

[5] R. A. M. Napitupulu, J. Daely, R. Manurung, and C. S. P. Manurung, "Effect of Chrom Electroplating Time on Low Carbon Steel on Hardness, Corrosion Rate and Coating Thickness," vol. 1, no. V, pp. 76-85, 2022.

[6] K. For and M. Terjadinya, "Jurnal patria bahari," vol. 3, no. 1, 2023.

[7] M. Akrom et al., "QSPR-Based Artificial Intelligence in the Study of Corrosion Inhibitors," JoMMiT : Journal of Multi Media and IT , vol. 7, no. 1, p. 015–020, Jul 2023,doi: 10.46961/jommit.v7i1.721.

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

[9] M. Akrom and T. Sutojo, "Investigation of QSPR-Based Machine Learning Models in Pyrimidine Corrosion Inhibitors," Eksergi, vol. 20, no. 2, Art. no. 2, Jul 2023, doi:10.31315/e.v20i2.9864.

[10] P. Studies, P. Chemistry, and F. Science, "Potential of Guava Leaf Extract as an Alternative Iron Corrosion Inhibitor for Contextual Chemistry Learning Shofrina Surya Dewi*, Retno Aliyatul Fikroh, Fuadatul Mukoningah Introduction," vol. 6, no. 3, pp. 257-272, 2022.

[11] V. Frendyatha, M. Akrom, and G. Alfa, "A Study on the Corrosion Inhibition Efficiency of Quinoxaline Compounds Utilizing Machine Learning," vol. 21, no. 2, pp. 65-69, 2024.

[12] G. Against and L. Corrosion, "Effect of water hyacinth extract organic inhibitor addition on corrosion rate," vol. 2, no. 2, 2018.

[13] M. Akrom, S. Rustad, and H. Kresno, "Machine learning investigation to predict corrosion inhibition capacity of new amino acid compounds as corrosion inhibitors," Results in Chem., vol. 6, no. September, p. 101126, 2023, doi: 10.1016/j.rechem.2023.101126.

[14] F. M. Haikal, M. Akrom, and G. A. Trisnapradika, "Comparison of Multilinear Regression and Decision Tree

Regressor Algorithms in Predicting Pyridazine Corrosion Inhibition Efficiency," Edumatic J. Educ. Inform., vol. 7, no. 2, pp. 307-315, Dec. 2023, doi: 10.29408/edumatic.v7i2.22127.

[15] M. Akrom, Green Corrosion Inhibitors for Iron Alloys: A Comprehensive Review of Integrating Data-Driven Forecasting, Density Functional Theory Simulations, and Experimental Investigation, *Journal of Multiscale Materials Informatics*, vol. 1, pg. 22-37 (2024). https://doi.org/10.62411/jimat.v1i1.10495.

[16] A. C. Milano, "Classification of Rice Leaf Diseases Using the Efficientnet-B6 Deep Learning Model," J. Inform. and Tech. Applied Electrical, vol. 12, no. 1, 2024, doi: 10.23960/jitet.v12i1.3855.

[17] A. Khoirunnisa and N. G. Ramadhan, "Improving malaria prediction with ensemble learning and robust scaler: An integrated approach for enhanced accuracy," J. Infotel, vol. 15, no. 4, pp. 326-334, 2023, doi: 10.20895/infotel.v15i4.1056.

[18] L. Leng et al., "Machine learning predicting and engineering the yield, N content, and specific surface area of biochar derived from pyrolysis of biomass," Biochar, vol. 4, no. 1, 2022, doi: 10.1007/s42773-022-00183-w.

[19] antoni wibowo, "10 Fold-Cross Validation," vol. (accessed, 2023.

[20] D. Benkeser, M. Petersen, and M. Laan, "Improved Small-Sample Estimation of Nonlinear Cross-Validated Prediction Metrics," J. Am. Stat. Assoc., vol. 115, pp. 1-34, Oct. 2019, doi: 10.1080/01621459.2019.1668794.

[21] Tae KH, Roh Y, Oh YH, Kim H, Whang SE (2019) Data cleaning for accurate, fair, and robust models: a big data-AI integration approach. In: Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning, pp 1-4."

[22] F. Lian, M. Fu, and X. Ju, "An improvement of data cleaning method for grain big data processing using task merging," J Comput Commun, vol. 8, 2020, doi: 10.4236/jcc.2020.83001.

[23] A. Ambarwari, Q. Jafar Adrian, and Y. Herdiyeni, "Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 117–122, 2020, doi: 10.29207/resti.v4i1.1517.

[24] R. Subhi et al., "Implementation of Scaling Techniques in a Website-Based Server Balancing Management System," J. Comput. and Appl., vol. 09, no. 02, pp. 316- 326, 2021.

[25] D. Leni, "Selection of Optimal Machine Learning Algorithm for Prediction of Mechanical Properties of Aluminum," J. Energy Engines , Manufacturing, and Mater., vol. 7, no. 1, p. 35, 2023, doi: 10.30588/jeemm.v7i1.1490.

[26] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ. Comput. Sci., vol. 7, p. e623, 2021, doi: 10.7717/peerj-cs.623.

[27] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," no. 2, pp. 5481-5487, 2022.

[28] C. Verma and E. E. Ebenso, "Materials Advances inhibitors: design, performance and industrial," pp. 3806-3850, 2021, doi: 10.1039/d0ma00681e.

[29] 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, Nov. 2023, https://doi.org/10.1016/J.COMPTC.2023.114307.

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

[31] M. Akrom, S. Rustad, H. K. Dipojono, Prediction of Anti-Corrosion performance of new triazole derivatives via Machine learning, Comp. and Theoretical Chemistry, vol. 1236, pg. 114599, 2024.

[32] A. S. ElDen, M. A. Moustafa, H. M. Harb, and A. H. Emara,"Adaboost ensemble with simple genetic algorithm for student prediction model," AIRCC's International Journal of Computer Science and Information Technology, vol. 5, no. 2, pp. 73–85, 2013