Perbandingan Model Machine Learning Terbaik untuk Memprediksi Kemampuan Penghambatan Korosi oleh Senyawa Benzimidazole

Comparison of the Best Machine Learning Model to Predict Corrosion Inhibition Capability of Benzimidazole Compounds

Cornellius Adryan Putra Sumarjono[1], Muhamad Akrom[2], Gustina Alfa Trisnapradika[3]
[1,2,3]Prodi Teknik Informatika, Universitas Dian Nuswantoro
E-mail: [1] 111202012994@mhs.dinus.ac.id, [2] m.akrom@dsn.dinus.ac.id,
[3]gustina.alfa@dsn.dinus.ac.id

Abstrak

Penelitian ini merupakan studi eksperimen untuk melakukan penyelidikan inhibitor korosi oleh senyawa Benzimidazole dengan melakukan pendekatan machine learning (ML). Karena korosi menyebabkan banyak kerugian yang timbul karena kehilangan material konstruksi, keselamatan kerja dan pencemaran lingkungan akibat produk korosi dalam bentuk senyawa yang mencemarkan lingkungan. Melakukan pendekatan ML adalah untuk mendapatkan model akurasi yang terbaik sehingga dapat digunakan untuk memprediksi dengan relevan dan akurat terhadap suatu material. Dalam penelitian ini, kami mengevaluasi algoritma ML dengan metode linear dan nonlinear dengan menggunakan metode *k-fold cross-validation* untuk membantu dalam mengukur performa model ML. Mengacu pada metrik *coefficient of determination* (R^2) dan *root mean square error* (RMSE), kami menyimpulkan bahwa model AdaBoost regressor (ADA) merupakan model dengan performa prediksi terbaik dari eksperimen yang kami lakukan dari literatur untuk dataset senyawa benzimidazole. Keberhasilan model penelitian ini menawarkan perspektif baru tentang kemampuan model ML untuk memprediksi penghambat korosi yang efektif.

Kata Kunci: *Machine learning*, korosi, *cross-validation*, *benzimidazole*, *adaboost regressor*

Abstract

This research is an experimental study to investigate corrosion inhibitors by Benzimidazole compounds using a machine learning (ML) approach. Corrosion causes many losses arising from the loss of construction materials, work safety, and environmental pollution due to corrosion products in the form of compounds that pollute the environment. Taking an ML approach is to get the best accuracy model so that it can be used to make relevant and accurate predictions about a material. In this research, we evaluate ML algorithms with linear and non-linear methods using the k-fold cross-validation method to help measure the performance of the ML model. Referring to the coefficient of determination (R^2) and root mean square error (RMSE) metrics, we conclude that the AdaBoost regressor (ADA) model is the model with the best predictive performance from the experiments we conducted from the literature for the benzimidazole compound dataset. The success of this research model offers a new perspective on the ability of ML models to predict effective corrosion inhibitors.

Keywords: Machine learning, corrosion, cross-validation, benzimidazole, AdaBoost regressor

1. PENDAHULUAN

Korosi merupakan proses penurunan atau peluruhan material yang disebabkan oleh reaksi kimia antara logam dan lingkungan dimana terdapat banyak zat korosif yang menyebabkan terjadinya korosi sekitarnya [1]. Proses korosi melibatkan oksidasi logam oleh oksigen di udara atau zat korosif lainnya, yang menghasilkan produk korosi seperti oksida, hidroksida, atau garam logam. Reaksi korosi ini dapat mempengaruhi kualitas dan kinerja material, mengurangi umur

pakai, dan menyebabkan kerugian ekonomi yang signifikan [2], [3]. Beberapa faktor yang mempengaruhi laju korosi meliputi jenis logam yang terlibat, sifat lingkungan korosif (misalnya, kelembaban, pH, suhu, konsentrasi zat korosif), dan faktor-faktor lain seperti tegangan mekanis atau keausan gesekan [4]. Proses korosi juga dapat dipercepat oleh adanya galvanik (kontak antara dua logam yang berbeda dalam elektrolit), interaksi oleh mikroorganisme (misalnya bakteri), atau korosi terinduksi oleh tegangan [5], [6]. Studi korosi melibatkan pemahaman mekanisme korosi, pengembangan metode pengendalian korosi, dan evaluasi performa material dalam lingkungan korosif [7]. Pengendalian proses korosi sangat bermanfaat bagi berbagai industri seperti industri minyak dan gas, industri kimia, industri otomotif, dan konstruksi [8].

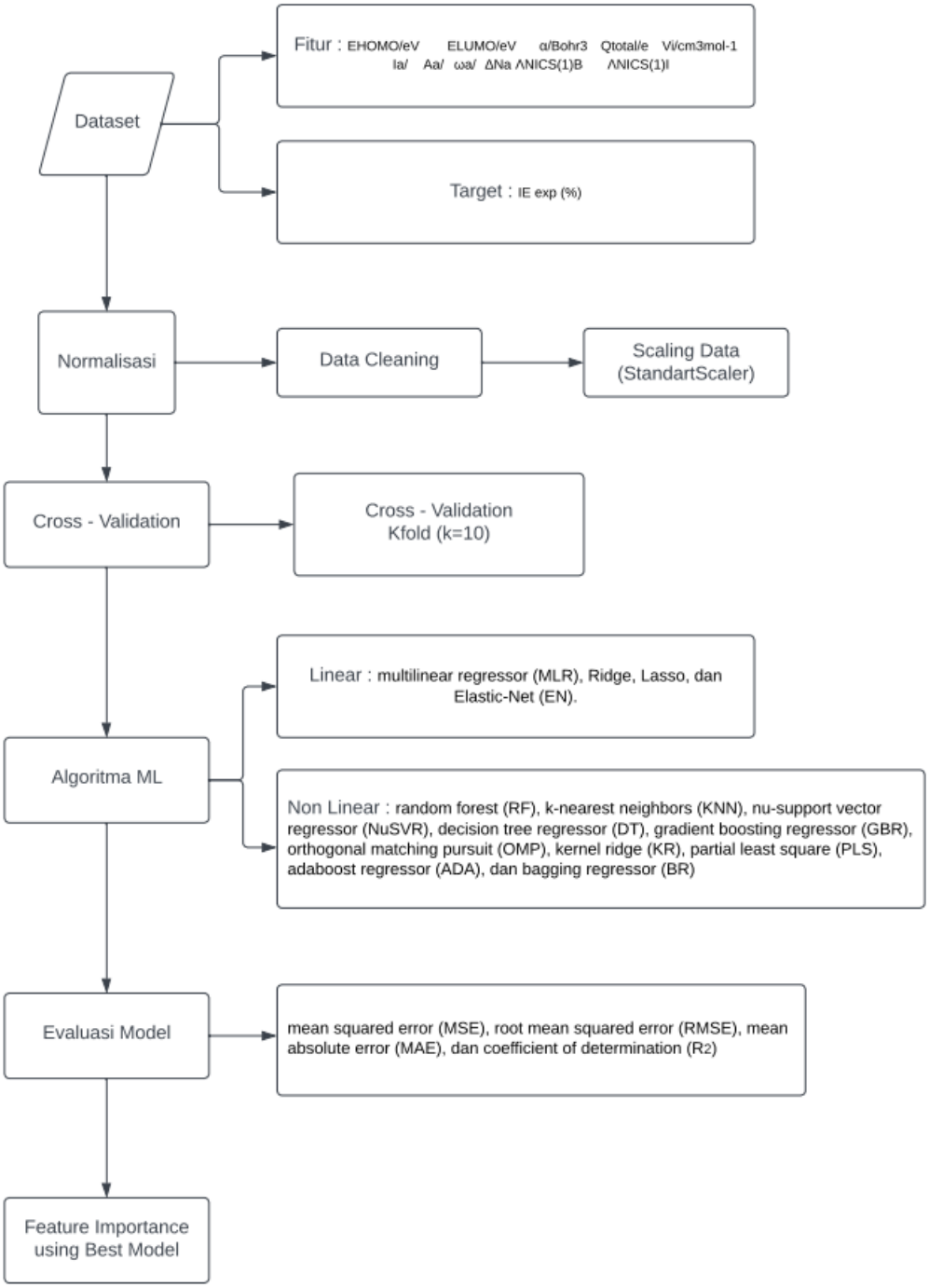
Senyawa benzimidazole ($C_7H_6N_2$) adalah senyawa organik yang terdiri dari cincin heterosiklik dengan struktur inti benzene (C_6H_5) dan imidazole ($C_3H_3N_2$) [9]. Senyawa benzimidazole memiliki berbagai turunan yang telah diteliti dan ditemukan memiliki aktivitas molekuler yang beragam [10]–[12], sehingga digunakan dalam berbagai bidang, termasuk farmasi, agrokimia, dan kimia material. Senyawa Benzimidazole digunakan dalam sintesis bahan kimia, sebagai pigmen organik, katalis, dan pengendali laju korosi [13]. Studi penggunaan benzimidazole sebagai penghambat (inhibitor) korosi secara eksperimental memakan banyak waktu, biaya, dan sumber daya yang intensif [14], [15].

Untuk mengatasi kesenjangan tersebut, baru-baru ini pendekatan *machine learning* (ML) berbasis model *quantitative structure-property relationship* (QSPR) digunakan dalam investigasi dan eksplorasi material baru antikorosi. Adanya hubungan antara struktur dan sifat molekuler dari senyawa, menjadikan QSPR efektif dan handal [16]. Lu Li dkk. [20] menggunakan algoritma ML yaitu *support vector machine* (SVM) untuk menginvestigasi senyawa benzimidazole sebagai inhibitor korosi. Hasilnya adalah model SVM memiliki nilai *coefficient of determination* (R^2) dan *root mean square error* (RMSE) masing-masing adalah 0.96 dan 6.79. Akrom dkk. [18] juga membandingkan beberapa model untuk melakukan pendekatan ML pada dataset senyawa pirimidin, dimana dihasilkan bahwa model *gradient boosting regressor* (GBR) memiliki akurasi terbaik (R^2 = 0.92, RMSE = 0,95) daripada model *support vector regression* (SVR), dan *k-nearest neighbor* (KNN).

Pada penelitian ini, kami bertujuan melakukan pengembangan model ML dalam memprediksi efisiensi penghambatan korosi/*corrosion inhibition efficiency* (CIE) senyawa turunan benzimidazole [17]. Hal yang membedakan penelitian ini dengan penelitian sebelumnya adalah implementasi teknik normalisasi data pada saat *preprocessing* dan teknik *k-fold cross-validation* dalam pengembangan model ML untuk meningkatkan akurasi prediksi. Hasil penelitian ini dapat memberikan wawasan dalam pengembangan model ML untuk desain senyawa inhibitor korosi potensial. Penelitian ini diharapkan juga bisa menjadi refrensi bagi peneliti lain dalam hal pengembangan model machine learning, terutama dalam meningkatkan akurasi performa model.

2. METODE PENELITIAN

Ilustrasi pengembangan model ML yang diusulkan dapat dilihat pada Gambar 1. Penjelasan detail dapat dilihat pada subbagian 2.1 sampai 2.4.



Gambar 1. Metode eksperimen model ML

2.1. Dataset

Dataset merupakan faktor mendasar untuk melakukan penelitian berbasis ML [19]. Pada penelitian ini, dataset yang digunakan diambil dari literatur yang telah dipublikasikan yang berisi 20 senyawa benzimidazole dengan 12 fitur dan 1 target [20]. Fitur-fitur yang digunakan merupakan sifat molekuler senyawa benzimidazole, diantaranya adalah energy of highest occupied molecular orbital (E-HOMO), energy of lowest unoccupied molecular orbital (E-LUMO), polarizability (α), total natural charges (Qtotal), molar volume (Vi), the adiabatic ionization potential (Ia), the adiabatic electron affinity (Aa), electrophilicity (ωa), the fraction electron shared (ΔN), indexes of aromaticity in the benzene (ΛNICS(1)B), dan indexes of

aromaticity in the imidazole (ΛNICS(1)I) yang merupakan variabel independen. Sementara targetnya adalah corrosion inhibition efficiency CIE (%) yang merupakan variabel dependen [21].

2.2. Normalisasi

Pada tahap preprocessing, dilakukan proses normalisasi data dengan teknik MinMax scaling untuk menghindasi sensitifitas data pada fitur tertentu. Selanjutnya, menerapkan cross-validation (CV). Teknik k-fold dipilih sebagai model CV untuk mengatasi bias dan varian pada data dengan melatih model secara berulang hingga menemukan kesalahan statistik yang paling kecil [27]. Kami menggunakan nilai k = 10, berati 1 fold digunakan sebagai set pengujian (test) dan 9 fold lainnya sebagai set pelatihan (train). Pemilihan nilai k-fold bergantung pada data yang digunakan, namun niali k = 5 atau k = 10 umum digunakan [28]–[31].

2.3. Algortima ML

Pada penelitian ini, kami membandingkan algoritma linear dan non-linear untuk mendapatkan model terbaik sebagai prediktor CIE senyawa benzimidazole. Model linear yang digunakan yaitu multilinear regressor (MLR), ridge, lasso, dan Elastic-Net (EN). Sedangkan model non-linear yang digunakan yaitu 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 (ADA), dan bagging regressor (BR) [22], [23]. Penggunaan ML merupakan sebuah metode statistika untuk melakukan identifikasi hubungan antara variabel independen (x) dan variabel dependen (y) [24], [25], [26]. Analisis fitur penting juga dilakukan untuk mengetahui sejauh mana fitur-fitur yang digunakan bertanggungjawab terhadap kinerja model ML.

2.4. Evaluasi model

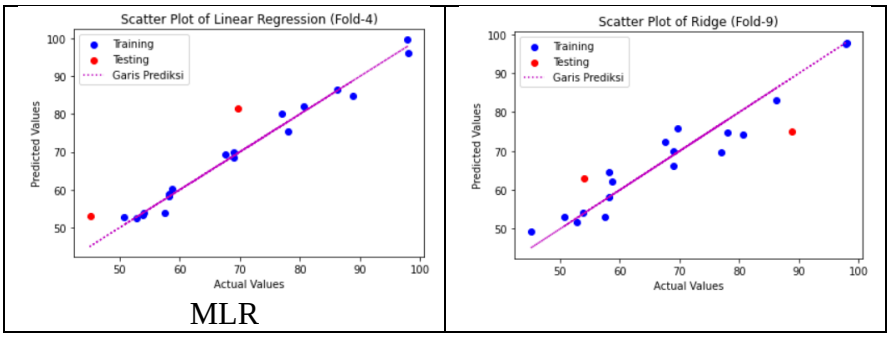
Untuk mengukur performa model prediksi, digunakan metrik regresi berupa mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), dan coefficient of determination (R^2). Model terbaik adalah yang memiliki nilai MSE, RMSE, dan MAE terendah, serta R^2 mendekati 1.

3. HASIL DAN PEMBAHASAN

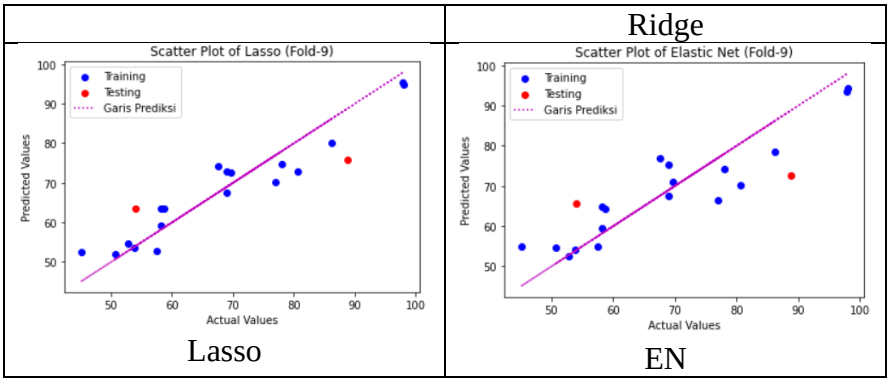
Langkah pertama yang dilakukan dalam penelitian ini adalah dengan menguji dataset Benzimidasole menggunakan algoritma linear yang terdapat pada library scikit-learn dengan bahasa python. Performa masing-masing model diukur dengan nilai R^2 dan RMSE. Tabel 1 dan Tabel 2 berikut masing-masing menyajikan performa model linier dan nonlinier.

Tabel 1. Performa model linier

Model	MSE	RMSE	MAE	R^2
MLR	3.664	1.914	1.498	0.984
Ridge	15.692	3.961	3.216	0.931
Lasso	20.763	4.557	3.939	0.909
EN	36.450	6.037	4.988	0.840



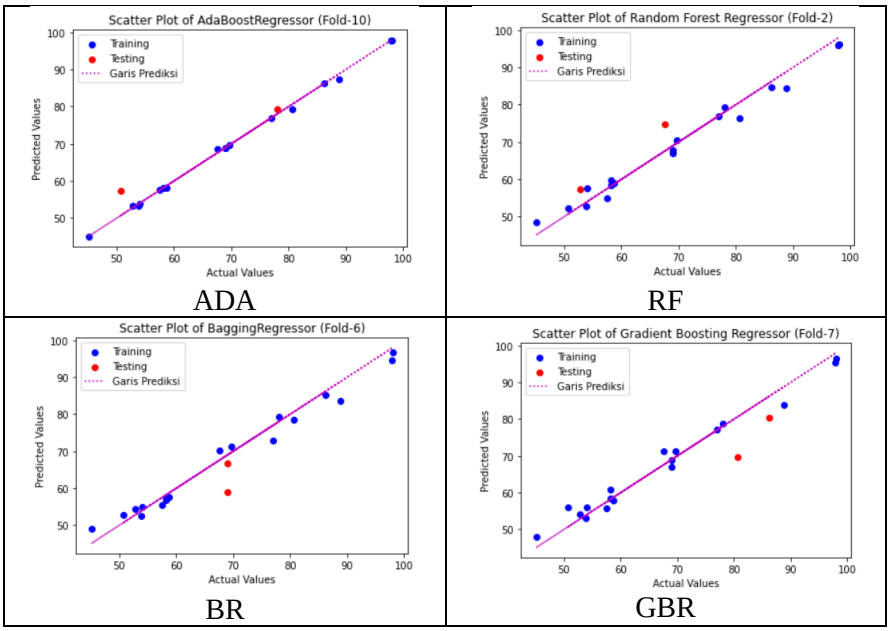
MLR



Gambar 2. Distribusi data poin model linier

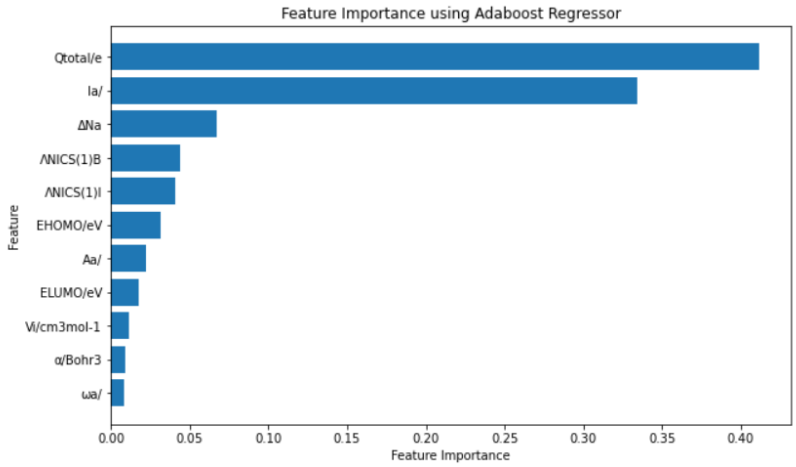
Tabel 2. Performa model nonlinier

Model	MSE	RMSE	MAE	R^2
ADA	0.298	0.546	0.316	0.999
RF	5.130	2.265	1.859	0.979
BR	5.804	2.409	2.066	0.977
GBR	5.980	2.445	1.955	0.974



Gambar 3. Distribusi data poin model nonlinier

Berdasarkan Tabel 1, diantara model linier, MLR tampil sebagai model yang memiliki kinerja prediksi lebih bagus dibandingkan ridge, lasso, dan EN berdasarkan nilai R^2, MAE, MSE, dan RMSE. MLR memiliki nilai R2 tertinggi (0.984), dan nilai MSE (3.664), RMSE (1.914), dan MAE (1.498) terendah. Hasil tersebut didukung oleh distribusi data pada Gambar 2, dimana data poin prediksi MLR cenderung paling dekat dengan garis prediksi dibandingkan model linier lainnya. Pada model nonlinier (Tabel 2), ADA menunjukkan kinerja prediksi yang paling unggul dibandingkan RF, BR, dan GBR berdasarkan metrik evaluasi yang digunakan (R2 = 0.999, MSE = 0.298, RMSE = 0.546, dan MAE = 0.316). Hasil tersebut terkonfirmasi oleh ditribusi data poin model ADA yang paling dekat dengan garis prediksi. Secara keseluruhan, model ADA muncul sebagai model terbaik dalam memprediksi nilai CIE senyawa benzimidazole sebagai inhibitor korosi. Selain itu, hasil prediksi ADA juga lebih baik dibandingkan hasil kajian serupa pada dataset yang sama dengan model SVR (RMSE = 6,79 dan R2 = 0.979) [20].



Gambar 5. Fitur Penting

Fitur penting merupakan fitur dalam dataset yang paling berpengaruh terhadap kemampuan interpretasi dan kestabilan algoritma, fitur penting juga dapat diketahui dengan memiliki korelasi yang tinggi [32], [33]. Dari Gambar 5 merupakan grafik hasil analisis fitur penting pada dataset Benzimidazole. Terlihat bahwa total natural charges (Qtotal) dan the adiabatic ionization potential (Ia) merupakan 2 fitur paling penting yang mempengaruhi performa model Adaboost Regressor. Namun pada grafik menunjukkan korelasi yang positif terhadap variabel y (CIE), maka dari itu semua fitur mendukung hasil dari performa kinerja model prediksi.

4. KESIMPULAN DAN SARAN

Investigasi model terbaik untuk memprediksi nilai CIE senyawa benzimidazole berbasis pendekatan ML telah dilakukan dengan membandingkan model linier dan nonlinier. Model nonlinier ADA terkonfirmasi sebagai model paling akurat dibandingkan 4 model linier dan 3 model nonlinier lainnya dengan nilai R^2 = 0.999 paling tinggi dan nilai MSE = 0.298, RMSE = 0.546, dan MAE = 0.316 paling rendah. Penelitian ini memberikan wawasan penting dalam pengembangan metode eksplorasi material secara efektif dan efisien sehingga dapat menjadi pertimbangan industri dalam perancangan material inhibitor korosi.

DAFTAR PUSTAKA

[1] AS Ariyanto, "Corrosion in Reinforcing Steel and its Prevention (Case Study of the Yos Sudarso Square Ruko Building, Semarang)," vol. 6, 2022.

[2] M. Sugeng, FM Ismail, and JP Utomo, "ANALYSIS OF DIFFERENCES IN CORROSION RATES FROM WEIGHT LOSS AND POLARIZATION TESTING ON PIPE WITH ASTM G59 AND ASTM G31 STANDARD CORROSION TESTING," *Tera Journal* , vol. 2, no. 1, Art. no. 1, March 2022.

[3] "Aesculus hippocastanum seeds extract as eco-friendly corrosion inhibitor for desalination plants: Experimental and theoretical studies - ScienceDirect." https://www.sciencedirect.com/science/article/abs/pii/S0167732222011321 (accessed 23 August 2023).

[4] A. MA'RUF, ""OPTIMIZATION OF MAIN DECK HANDLING AGAINST CORROSION ON BOARD THE SHIP MV. BELIC MAS," diploma, POLYTECHNIC OF SEALING SCIENCES SEMARANG, 2022. Accessed: 3 July 2023. [Online]. Available at: https://library.pip-semarang.ac.id

[5] M. Akrom, "INVESTIGATION OF NATURAL EXTRACTS AS GREEN CORROSION INHIBITORS IN STEEL USING DENSITY FUNCTIONAL THEORY," *Journal of Physics Theory and Applications* , vol. 10, no. 1, Art. no. 1, Jan 2022.

[6] MAS ISLAM, "ANALYSIS OF THE INFLUENCE OF VARIATIONS OF NICKEL COATING HOLDING TIME ON THE THICKNESS AND RATE OF CORROSION IN ASTM A36 STEEL," undergraduate_(S1), Nahdlatul Ulama Sunan Giri University, 2022. Accessed: 3 July 2023. [Online]. Available at: https://repository.unugiri.ac.id/id/eprint/1774/

[7] R. Napitupulu, J. Daely, R. Manurung, and C. Manurung, "EFFECT OF CHROM ELECTROPLATING TIME ON LOW CARBON STEEL ON HARDNESS, CORROSION RATE AND LAYER THICKNESS," *Citra Science Technology* , vol. 1, no. 2, Art. no. 2, Jan. 2022, doi: 10.2421/cisat.v1i2.38.

[8] A. Husodo, "OPTIMIZING THE PAINTING PROCESS ON SHIP DECKS TO SLOW DOWN CORROSION," *JPB: Jurnal Patria Bahari* , vol. 3, no. 1, Art. no. 1, May 2023, doi: 10.54017/jpb.v3i1.76.

[9] Bihan Musab, "ANALYSIS OF NITRO-BENZIMIDAZOLE SENSORS WITH CYANIDE ANIONS: CHEMISTRY MAGAZINE FOR SMA/MA AS A LEARNING SUPPLEMENT ON CHEMICAL BONDING MATERIALS," thesis, Mataram University, 2023. Accessed: 21 August 2023. [Online]. Available at: http://eprints.unram.ac.id/41978/

[10] S. Maharani, NY Haryono, and BD Mariana, "Analysis of the Effect of Concentration of Benomil Fungicide on Sterilization of Tawangmangu Citrus Plant Stem Tissue Culture," *Proceedings of Life and Applied Sciences* , vol. 1, no. 0, Art. no. 0, 2022, Accessed: 21 August 2023. [Online]. Available at: http://conference.um.ac.id/index.php/LAS/article/view/7846

[11] 2027011005 FENDI SETIAWAN, "ISOLATION AND CHARACTERIZATION OF BIOACTIVE COMPOUNDS RESULTING FROM FERMENTATION OF MARINE BIOTA ASSOCIATE ACTINOMYTETES AS FUNGICIDES," masters, LAMPUNG UNIVERSITY, 2022. Accessed: 23 August 2023. [Online]. Available at: http://digilib.unila.ac.id/64940/

[12] I. Panduwiguna, I. Hardiana, SA Ogiuma, and MS Latief, "GERD DISEASE PRESCRIBING PATTERNS IN THE INPATITION INSTALLATION OF XYZ HOSPITAL JAKARTA," *KRYONAUT PHARMACY JOURNAL* , vol. 2, no. 1, Art. no. 1, Jan. 2023, doi: 10.59969/jfk.v2i1.21.

[13] 1717011070 By MITA SEPTIANI, "SYNTHESIS, CHARACTERIZATION, AND APPLICATION OF SCHIFF BASE COMPLEX COMPOUNDS OF 4-(DIMETYLAMINO)BENZALDEHYDE AND ANYLYNE WITH Mn(II) COMPLEX AS SENSITIZER IN DYE SENSITIZED SOLAR CELL (DSSC) (Thesis)," 2021 http://digilib.unila.ac.id/60959/ (accessed 27 August 2023).

[14] MA Quraishi, DS Chauhan, and VS Saji, "Heterocyclic biomolecules as green corrosion inhibitors," *Journal of Molecular Liquids* , vol. 341, p. 117265, Nov 2021, doi: 10.1016/j.molliq.2021.117265.

[15] D. Prasad, R. Singh, Z. Safi, N. Wazzan, and L. Guo, "De-scaling, experimental, DFT, and MD-simulation studies of unwanted growing plant as natural corrosion inhibitor for SS-410 in acid medium," *Colloids and Surfaces A: Physicochemical and Engineering Aspects* , vol. 649, p. 129333, Jun 2022, doi: 10.1016/j.colsurfa.2022.129333.

[16] 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.

[17] "Machine Learning Understanding and How it Works - Nurdin Hamzah University." https://unh.ac.id/?p=1322 (accessed 26 December 2022).

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

[19] D. Leni, YP Kusuma, R. Sumiati, Muchlisinalahuddin, and Adriansyah, "Comparison of Machine Learning Algorithms for Predicting Mechanical Properties of Low Alloy Steel,"

Journal of Materials, Manufacturing and Energy Engineering , vol. 5, no. 2, Art. no. 2, Sep 2022, doi: 10.30596/rmme.v5i2.11407.

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

[21] 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," *International Journal of Molecular Sciences* , vol. 23, no. 9, Art. no. 9, Jan. 2022, doi: 10.3390/ijms23095086.

[22] D. Leni, "Selection of Optimal Machine Learning Algorithms for Predicting the Mechanical Properties of Aluminum," *Journal of Engines: Energy, Manufacturing, and Materials* , vol. 7, no. 1, Art. no. 1, May 2023, doi: 10.30588/jeemm.v7i1.1490.

[23] T. Sutojo, S. Rustad, M. Akrom, A. Syukur, GF Shidik, and HK Dipojono, "A machine learning approach for corrosion small datasets," *npj Mater Degrad* , vol. 7, no. 1, Art. no. 1, Mar 2023, doi: 10.1038/s41529-023-00336-7.

[24] LS Ihzaniah, "Comparison of the K-Nearest Neighbor Regression Method and Multiple Linear Regression Method on Boston Housing Data," Thesis, 2023. Accessed: 22 August 2023. [Online]. Available at: https://repository.uksw.edu//handle/123456789/29040

[25] D. Rahmawati, T. Kristanto, BFS Pratama, and DB Abiansa, "Prediction of Foreign Travelers During the COVID-19 Pandemic Using Simple Linear Regression Methods," *Journal of Information Systems Research (JOSH)* , vol. 3, no. 3, Art. no. 3, Apr. 2022, doi: 10.47065/josh.v3i3.1507.

[26] HD Panduwinata, S. Suyitno, and MN Huda, "Weibull Regression Model on Classified Continuous Data," *EXPONENTIAL* , vol. 13, no. 2, Art. no. 2, November 2022.

[27] "10 Fold-Cross Validation," *MTI* . https://mti.binus.ac.id/2017/11/24/10-fold-cross-validation/ (accessed 29 May 2023).

[28] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," *Journal of the American Statistical Association* , vol. 0, no. 0, p. 1–12, Apr 2023, doi: 10.1080/01621459.2023.2197686.

[29] 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, p. 63, Nov 2022, doi: 10.1007/s42773-022-00183-w.

[30] T. Zhu, S. Li, L. Li, and C. Tao, "A new perspective on predicting the reaction rate constants of hydrated electrons for organic contaminants: Exploring molecular structure characterization methods and ambient conditions," *Science of the Total Environment* , vol. 904, p. 166316, Dec 2023, doi: 10.1016/j.scitotenv.2023.166316.

[31] M. Akrom, S. Rustad, AG Saputro, A. Ramelan, F. Fathurrahman, and HK Dipojono, "A combination of machine learning model and density functional theory method to predict corrosion inhibition performance of new diazine derivative compounds," *Materials Today Communications* , vol. 35, p. 106402, Jun 2023, doi: 10.1016/j.mtcomm.2023.106402.

[32] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithms: A review," *Journal of King Saud University - Computer and Information Sciences* , vol. 34, no. 4, p. 1060–1073, Apr 2022, doi: 10.1016/j.jksuci.2019.06.012.

[33] "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data | Briefings in Bioinformatics | Oxford Academic." https://academic.oup.com/bib/article/23/1/bbab354/6366322 (accessed 11 September 2023).