Optimasi model *machine learning* untuk prediksi inhibitor korosi berbasis *augmentasi dataset* senyawa *n-heterocyclic* menggunakan *KDE*

Machine learning model optimization for corrosion inhibitor prediction based on n-heterocyclic compound dataset augmentation using KDE

Rizky Syah Gumelar[1)], Muhamad Akrom[2)], Gustina Alfa Trisnapradika[3)]

[1), 2), 3)] *Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro*
[2), 3)] *Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro*
Email : 111202113304@mhs.dinus.ac.id[1)], m.akrom@dsn.dinus.ac.id[2)], gustina.alfa@dsn.dinus.ac.id[3)]

Abstrak

Penelitian ini bertujuan mengoptimalkan model machine learning untuk memprediksi efektivitas inhibitor korosi dari senyawa N-Heterocyclic. Tantangan utama dalam pemodelan ini adalah terbatasnya dataset karena tingginya biaya dan waktu yang dibutuhkan untuk mengumpulkan data eksperimen. Untuk mengatasi masalah ini, penelitian ini menggunakan Kernel Density Estimation (KDE) sebagai teknik augmentasi data, menghasilkan sampel virtual yang meningkatkan keragaman dataset dan kinerja prediktif model. Dataset yang dikembangkan mencakup 11 fitur kimia yang relevan seperti HOMO, LUMO, dan Gap Energy. Model machine learning linier (MLR, Ridge, Lasso, dan ElasticNet) dan non-linier (KNR, Random Forest, Gradient Boosting, Adaboost, XGBoost) dievaluasi berdasarkan Root Mean Squared Error (RMSE) dan koefisien determinasi (R^2). Hasil penelitian menunjukkan bahwa augmentasi data menggunakan KDE meningkatkan akurasi dan stabilitas prediksi, terutama pada model non-linier seperti Random Forest dan XGBoost. Penerapan KDE terbukti efektif dalam meningkatkan performa model prediktif dan dapat direkomendasikan sebagai metode augmentasi dalam penelitian serupa yang memerlukan data tambahan untuk meningkatkan ketepatan prediksi.

Kata kunci: *Machine Learning, Kernel Density Estimator (KDE), Inhibitor Korosi, Augmentasi Dataset, N-Heterocyclic.*

Abstract

This study aims to optimize a machine learning model to predict the corrosion inhibitor effectiveness of N-Heterocyclic compounds. The main challenge in this modelling is the limited dataset due to the high cost and time required to collect experimental data. To overcome this problem, this research utilizes Kernel Density Estimation (KDE) as a data augmentation technique, generating virtual samples that improve dataset diversity and model predictive performance. The developed dataset includes 11 relevant chemical features such as HOMO, LUMO, and Gap Energy. Linear (MLR, Ridge, Lasso, and ElasticNet) and non-linear (KNR, Random Forest, Gradient Boosting, Adaboost, XGBoost) machine learning models were evaluated based on Root Mean Squared Error (RMSE) and coefficient of determination (R^2). The results show that data augmentation using KDE improves prediction accuracy and stability, especially in non-linear models like Random Forest and XGBoost. The application of KDE proved effective in improving the performance of predictive models. It can be recommended as an augmentation method in similar studies that require additional data to improve prediction accuracy.

Keywords: *Machine Learning, Kernel Density Estimator (KDE), Corrosion Inhibitor, Dataset Augmentation, N-Heterocyclic.*

1. PENDAHULUAN

Korosi adalah proses degradasi material khususnya logam, yang terjadi akibat interaksi kimiawi atau elektrokimiawi dengan lingkungan sekitarnya. Proses korosi dapat menyebabkan kerusakan yang signifikan pada infrastruktur dan peralatan industri, seperti pipa, jembatan, serta peralatan dalam industri kimia dan minyak [1], [2]. Proses ini menghasilkan konversi logam dari bentuk murninya menjadi bentuk yang lebih stabil secara kimiawi seperti oksida, sulfida, dan hidroksida [3]. Fenomena ini tidak hanya terbatas pada lingkungan tertentu, melainkan dapat terjadi dalam berbagai kondisi, baik itu padat, cair, maupun gas, dan dipandang sebagai proses yang universal [4]. Korosi dapat menyebabkan kerusakan signifikan pada infrastruktur dan peralatan industri, termasuk pipa, jembatan, serta peralatan dalam industri kimia dan minyak.

Akibatnya, kerusakan ini tidak hanya mempengaruhi struktur dan kinerja material, tetapi juga berdampak pada ekonomi dan keamanan, sering kali memicu kecelakaan serius jika tidak ditangani dengan baik [5], [6]. Langkah preventif yang umum dilakukan untuk mengatasi ini adalah penggunaan inhibitor korosi, yang berfungsi untuk memperlambat atau mencegah proses korosi [2], [7], [8].

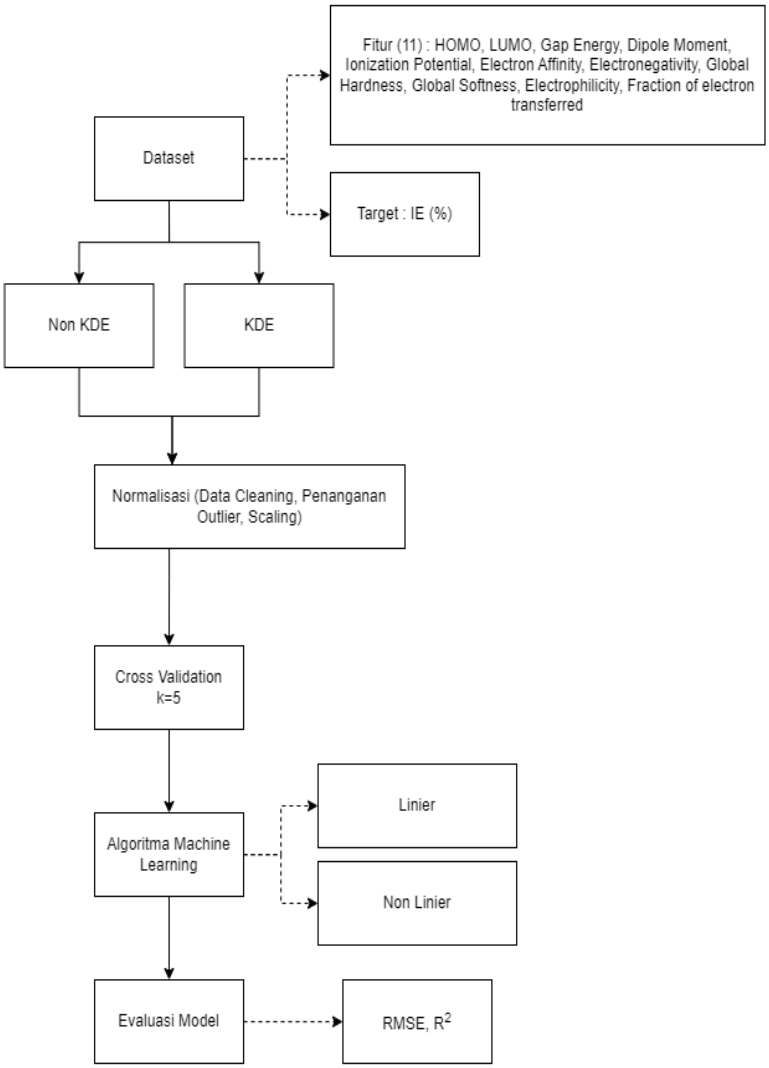
Inhibitor korosi adalah senyawa kimia yang dapat membentuk lapisan pelindung pada permukaan logam, sehingga mencegah reaksi korosif antara logam dan lingkungan sekitarnya [9]. Dalam penelitian ini, kumpulan turunan senyawa N-Heterocyclic digunakan sebagai objek studi untuk memprediksi efektivitas inhibitor korosi berdasarkan sejumlah fitur kimiawi. Sebanyak 11 fitur kimia digunakan sebagai variabel prediktor, yaitu HOMO (*Highest Occupied Molecular Orbital*), LUMO (*Lowest Unoccupied Molecular Orbital*), *Gap Energy*, *Dipole Moment*, *Ionization Potential*, dan *Electrophilicity*, serta beberapa fitur lain yang relevan. Target dari penelitian ini adalah memprediksi efisiensi inhibitor korosi yang dinyatakan dalam bentuk *Inhibition Efficiency* (IE %).

Namun, salah satu tantangan dalam mengembangkan model prediksi untuk inhibitor korosi adalah terbatasnya ukuran dataset. Pengumpulan data eksperimental untuk inhibitor korosi seringkali memerlukan biaya dan waktu yang tinggi, sehingga menyebabkan ketersediaan data menjadi terbatas [2], [10]. Dalam kasus inilah machine learning dapat menjadi alat yang efektif. Machine learning memungkinkan pemodelan hubungan antara fitur kimiawi dan efisiensi inhibitor korosi meskipun dengan data terbatas.

Untuk mengatasi masalah keterbatasan dataset, penelitian ini menggunakan Kernel *Density Estimation* (KDE) untuk mengatasi masalah keterbatasan dataset dengan cara menghasilkan sampel virtual. KDE adalah teknik yang digunakan untuk memperkirakan distribusi dari sampel yang ada, dan dengan menggunakan metode ini data virtual dapat dihasilkan untuk mengisi celah antara distribusi sampel, sehingga meningkatkan variasi dataset [11]. Dengan demikian, pendekatan ini tidak hanya mengatasi keterbatasan dataset, tetapi juga mengoptimalkan kinerja model dalam memprediksi efisiensi inhibitor korosi. Penelitian ini bertujuan untuk membuktikan bahwa augmentasi dataset dengan KDE dapat secara signifikan meningkatkan performa model *machine learning* dalam memprediksi efektivitas inhibitor korosi berbasis senyawa N-Heterocyclic.

2. DASAR TEORI

Gambar 1 menggambarkan proses pembuatan model *Machine Learning* (ML) untuk memprediksi efisiensi inhibitor korosi oleh senyawa N-Heterocyclic. Tahapan pertama adalah pengumpulan dataset yang mencakup fitur-fitur kimia seperti *HOMO, LUMO,* dan *Gap Energy*. Langkah berikutnya adalah normalisasi data, yang mencakup pembersihan data, penanganan outlier, dan scaling untuk memastikan model sensitif terhadap variasi data yang relevan. Setelah itu, model ML dibagi menjadi dua pendekatan: linier dan non-linier, untuk menangani hubungan antara variabel input dan output. Model ini kemudian dilatih menggunakan *cross-validation k-fold* (k=5) karena memiliki hasil yang cukup akurat dengan waktu komputasi yang efisien. Evaluasi model dilakukan dengan menggunakan metrik seperti *Root Mean Squared Error* (RMSE) dan koefisien determinasi (R^2), yang membantu dalam memperkirakan keakuratan prediksi model terhadap data baru. Proses ini secara keseluruhan bertujuan untuk memilih model ML yang optimal dan memastikan prediksi yang akurat tentang efisiensi inhibitor korosi oleh senyawa N-Heterocyclic dalam penelitian ini.



Gambar 1. *Pengembangan model ML*

3. METODOLOGI PENELITIAN

3.1. Dataset

Dataset yang digunakan dalam penelitian ini menghimpun senyawa N-Heterocyclic dari berbagai sumber literatur yang telah direview. Dataset ini mencakup 11 fitur yang penting untuk prediksi efisiensi inhibitor korosi, yaitu *HOMO, LUMO, Gap Energy, Dipole Moment, Ionization Potential, Electron Affinity, Electronegativity, Global Hardness, Global Softness, Electrophilicity,* dan *Fraction of electron transferred*. Fitur-fitur ini dipilih karena merefleksikan sifat-sifat molekuler serta karakteristik fisikokimia yang dapat mempengaruhi kemampuan senyawa sebagai inhibitor korosi. Variabel dependen dalam analisis ini adalah efisiensi inhibitor korosi (IE (%)), yang menjadi fokus utama untuk analisis dan prediksi dalam penelitian ini.

3.2. Kernel Density Estimator

Dalam penelitian ini*, Kernel Density Estimator* (KDE) diaplikasikan untuk menghasilkan sampel data sintetis dengan tujuan untuk meningkatkan keragaman dalam kumpulan data yang ada. KDE adalah teknik *non-parametrik* yang digunakan untuk mengestimasi fungsi kepadatan probabilitas dari variabel-variabel acak [12], [13]. Prinsip metode KDE adalah sebagai berikut: misalkan x1,x2,…,xn adalah data sampel yang independen dan identik terdistribusi yang diambil dari satu set satu dimensi X, di mana X memiliki fungsi kepadatan yang tidak diketahui f(x). Maka, estimasi *kernel* dari kepadatan ini, seperti pada persamaan 1, dimana:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \quad \text{.....(1)}$$

Di mana *K(·)* adalah fungsi kernel, *h* adalah *bandwidth* dari fungsi *kernel*, dan *n* adalah ukuran sampel. Pemilihan bandwidth h menentukan tingkat kehalusan dan akurasi dari fungsi kepadatan yang diestimasi. Dalam kasus ini, fungsi *kernel Tophat* digunakan sebagai fungsi *kernel*. Rumus untuk *kernel Tophat*, seperti persamaan 2 adalah:

$$K(x) = \begin{cases} \frac{1}{2h}, & |x| \le h \\ x, & |x| > h \end{cases} \quad \text{.....(2)}$$

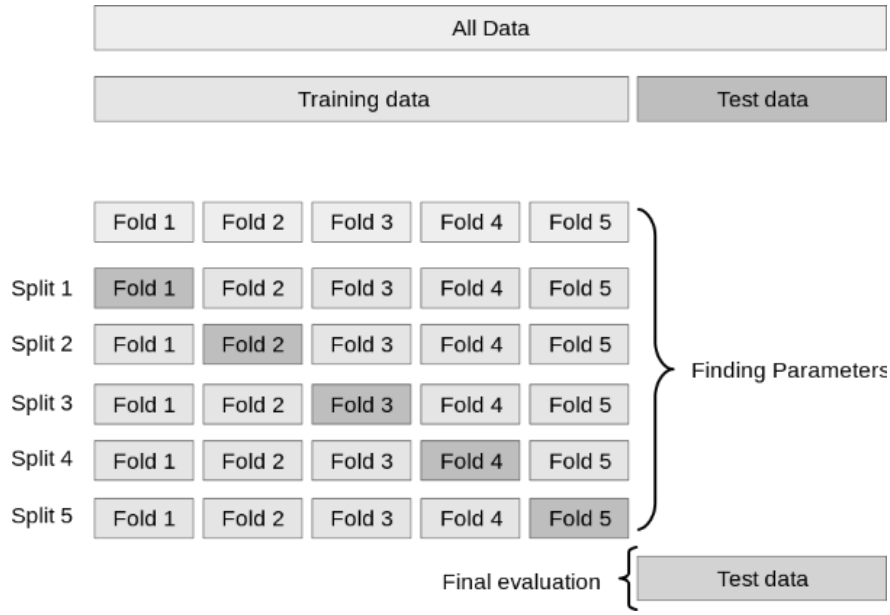
Di mana *h* merupakan *bandwidth*. KDE diimplementasikan dalam studi ini tidak hanya untuk meningkatkan ukuran *dataset*, tetapi juga untuk mengevaluasi stabilitas dan keandalan dari model prediktif yang dikembangkan, serta untuk memberikan analisis yang lebih komprehensif tentang kinerja model dalam berbagai kondisi data [12].

3.3. Normalisasi

Dalam proses *preprocessing*, normalisasi data dilakukan menggunakan *Robust Scaler* untuk mengurangi dampak *outlier* dan menyelaraskan rentang data tanpa mengasumsikan distribusi normal. Teknik ini sangat efektif untuk data yang sering mengandung nilai ekstrem, seperti yang sering ditemukan dalam *dataset* ekologis. Robust Scaler bekerja dengan menyesuaikan skala data berdasarkan nilai *interquartile range* (IQR), sehingga data dengan distribusi yang lebih bervariasi tetap dapat distandardisasi tanpa terpengaruh oleh *outlier*.

3.4. K-Fold Cross Validation

Pada pemodelan data, teknik *K-Fold Cross-Validation* digunakan sebagai metode evaluasi yang kuat untuk mengurangi bias dan varians dalam estimasi kesalahan model. Dengan menggunakan *K-fold cross-validation*, data dipisahkan menjadi data pelatihan dan pengujian, memungkinkan model untuk dinilai secara menyeluruh dan objektif [14], [15]. Proses ini juga bertujuan untuk meningkatkan akurasi prediksi model dan menurunkan risiko *overfitting*. Penelitian ini menggunakan teknik *K-Fold Cross-Validation* dengan $k = 5$ digunakan untuk menilai akurasi dan efisiensi model dengan lebih objektif, sambil mengurangi risiko bias dan varians dalam estimasi kesalahan.



Gambar 2. *Gambaran pembagian k-fold cross validation dengan k = 5*

1. Dataset dibagi menjadi 5 subset berukuran sama, dengan satu subset sebagai data pengujian dan sisanya untuk pelatihan. (*Gambar 2* menggambarkan pembagian data pada *K-Fold Cross-Validation*).
2. Pada setiap iterasi, *subset* yang berbeda digunakan sebagai data pengujian, sementara sisanya digunakan untuk pelatihan. Akurasi dihitung untuk setiap *fold*.
3. Proses ini diulang 5 kali, sehingga semua *subset* menjadi data pengujian sekali. Hasil akhir diperoleh dari rata-rata akurasi di semua *folds*, memberikan estimasi performa model yang lebih akurat.

3.5. Algoritma Machine Learning

Penelitian ini menggunakan berbagai algoritma *Machine Learning* baik linier maupun non-linier untuk mengidentifikasi model yang paling efektif. Algoritma yang diuji termasuk *Multilinear Regressor* (MLR), *Ridge, Lasso, ElasticNet* untuk regresi linier, dan *K-Nearest Neighbors Regressor* (KNR), *Random Forest*, *Gradient Boosting*, *Adaboost*, serta *XGBoost Regressor* (XGBR) untuk model regresi non-linier. Tujuan penggunaan algoritma-algoritma ini adalah untuk menguji dan membandingkan kinerja mereka dalam memprediksi variabel dependen berdasarkan hubungan antara variabel independen dan variabel dependen [16], [17].

3.6. Evaluasi Model

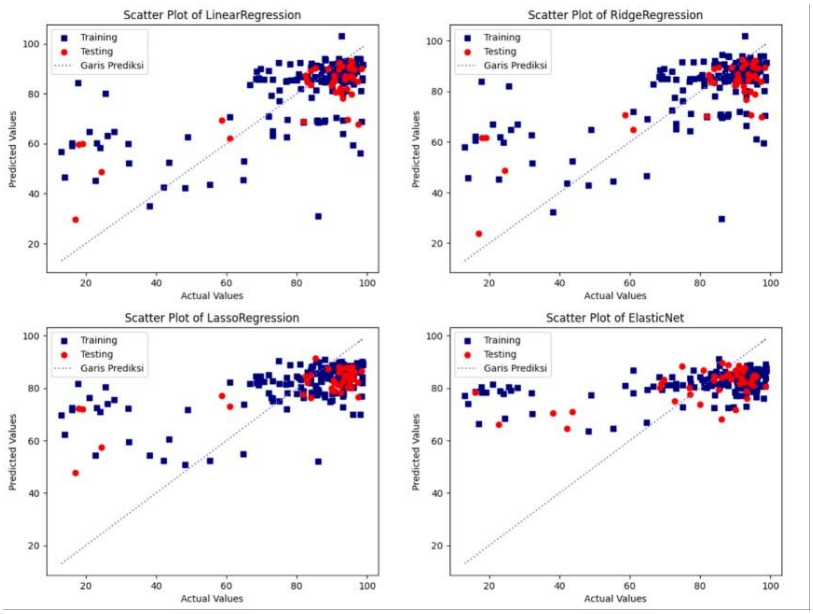
Evaluasi model dilakukan menggunakan dua metrik yaitu *Root Mean Squared Error* (RMSE) dan *Coefficient of Determination* (R^2). RMSE memberikan ukuran yang konsisten dari kesalahan model dalam unit yang sama dengan data, sedangkan R^2 menilai proporsi variabilitas dalam variabel dependen yang dapat dijelaskan oleh model independen [16], [18]. Model yang optimal ditandai dengan RMSE terendah dan R^2 yang mendekati satu, yang menunjukkan akurasi dan keandalan tinggi dalam prediksi model [19], [20]. Penambahan analisis sensitivitas untuk setiap algoritma juga akan membantu dalam memahami bagaimana variasi dalam data input mempengaruhi output model, memberikan wawasan tambahan tentang stabilitas dan *robustness* model yang dipilih.

4. PENGUJIAN DAN PEMBAHASAN

Pada penelitian ini, langkah pertama yang dilakukan adalah melakukan pengujian terhadap dataset N-*Heterocyclic* dengan menggunakan algoritma *machine learning* linier dan non-linier lalu membandingkannya berdasarkan metrik evaluasi R^2 (koefisien determinasi) dan RMSE (*Root Mean Squared Error*).

Tabel 1. *Kinerja model Prediksi Linier*

Model Linier	Training		Testing	
	R^2	RMSE	R^2	RMSE
MLR	0.424	0.157	0.621	0.155
Ridge	0.417	0.158	0.609	0.137
Lasso	0.322	0.170	0.430	0.166
ElasticNet	0.177	0.188	0.179	0.199

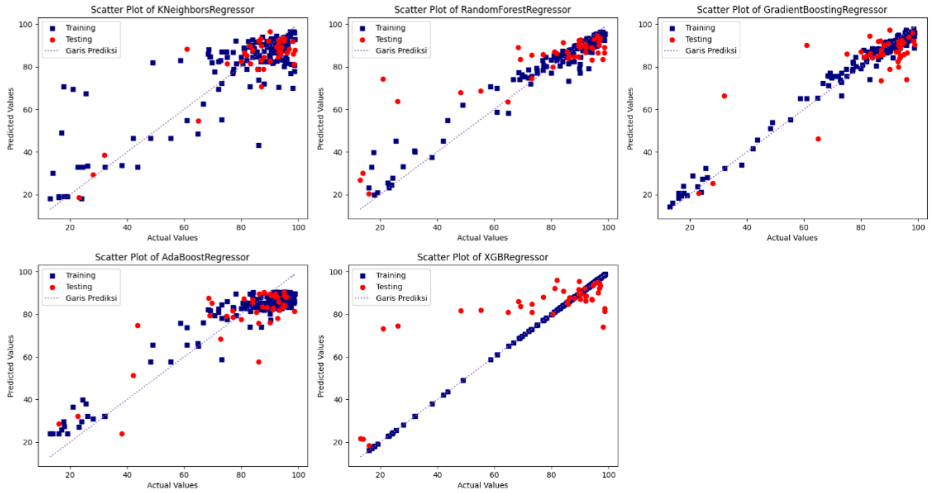


Gambar 3. *Scatter plot Model Linier*

Tabel 1 dan gambar 3 menunjukkan kinerja model linier: MLR, *Ridge*, *Lasso*, dan *ElasticNet* berdasarkan R^2 dan RMSE pada data *training* dan *testing*. MLR menghasilkan performa terbaik dengan R^2 sebesar 0,424 pada *training* dan 0,621 pada *testing*, namun sedikit menunjukkan tanda *overfitting* karena perbedaan kinerja antara kedua data. Model *Ridge* memberikan hasil hampir sebanding dengan MLR (R^2 0,417 pada *training* dan 0,609 pada *testing*) dan menunjukkan prediksi yang lebih baik dengan RMSE lebih rendah pada data uji (0,137). Sementara itu, *Lasso* dan *ElasticNet* menunjukkan performa yang lebih rendah pada kedua data dengan penurunan akurasi yang lebih signifikan. Secara keseluruhan, model MLR dan *Ridge* menunjukkan kinerja yang lebih baik dibandingkan Lasso dan *ElasticNet*, baik pada data *training* maupun *testing*. Meskipun begitu, model *Ridge* memberikan keseimbangan yang baik antara *training* dan *testing*, dengan sedikit penurunan *overfitting* dibandingkan dengan MLR, yang terlihat dari perbedaan nilai R^2 dan RMSE antara data latih dan uji.

Tabel 2. *Kinerja model Prediksi Non-Linier*

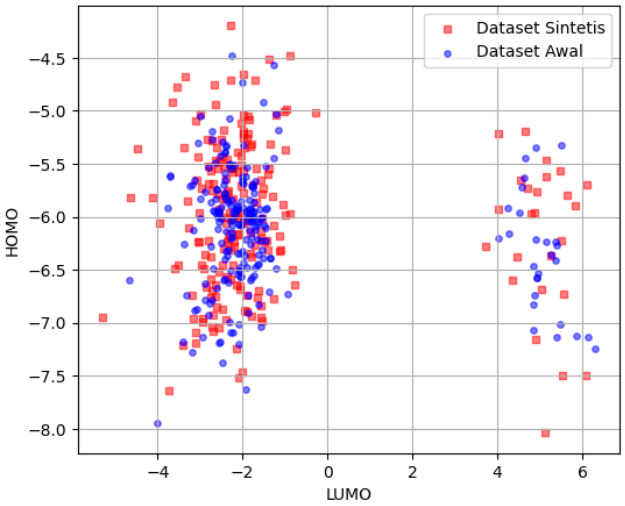
Model Non-Linier	**Training**		**Testing**	
	R^2	**RMSE**	**R^2**	**RMSE**
KNR	0.762	0.101	0.724	0.115
RFR	0.942	0.049	0.636	0.133
GBR	0.978	0.030	0.717	0.116
ABR	0.851	0.079	0.714	0.117
XGBR	0.999	0.002	0.578	0.142



Gambar 4. *Scatter plot Model Non-Linier*

Tabel 2 dan gambar 4, menunjukkan kinerja lima model regresi non-linier pada data *training* dan *testing*. XGBR menunjukkan performa terbaik pada data latih dengan R^2 hampir sempurna (0,999) dan RMSE sangat kecil (0,002), namun kinerjanya menurun signifikan pada data uji (R^2 0,578 dan RMSE 0,142), mengindikasikan *overfitting*. Model KNR memberikan prediksi terbaik dengan R^2 sebesar 0,724 dan RMSE 0,115 pada data uji, diikuti oleh GBR dan ABR yang juga konsisten. Secara keseluruhan, meskipun XGBR dan RFR unggul pada data latih, KNR, GBR, dan ABR lebih andal dalam generalisasi, dengan KNR menunjukkan kinerja paling seimbang.

Dari kedua tabel di atas, terlihat bahwa beberapa model memberikan prediksi yang kurang akurat, dengan indikasi kuat terjadinya *overfitting*, di mana model sangat cocok dengan data pelatihan namun performa menurun pada data pengujian. Untuk mengatasi masalah ini, metode *Kernel Density Estimator* (KDE) digunakan. Dalam metode ini, kernel *tophat* digunakan untuk menghasilkan sampel data virtual yang merata, meningkatkan keragaman dalam dataset pelatihan. Selain kernel *tophat*, metode KDE juga mendukung penggunaan kernel lainnya seperti *Gaussian, Epanechnikov,* dan *linear*. Pemilihan kernel bergantung pada karakteristik data dan tujuan analisis. Misalnya, kernel *Gaussian* sering digunakan karena sifatnya yang halus dan dapat menangkap distribusi data dengan baik, sedangkan kernel *Epanechnikov* lebih efisien dalam kasus dengan data berdimensi rendah.



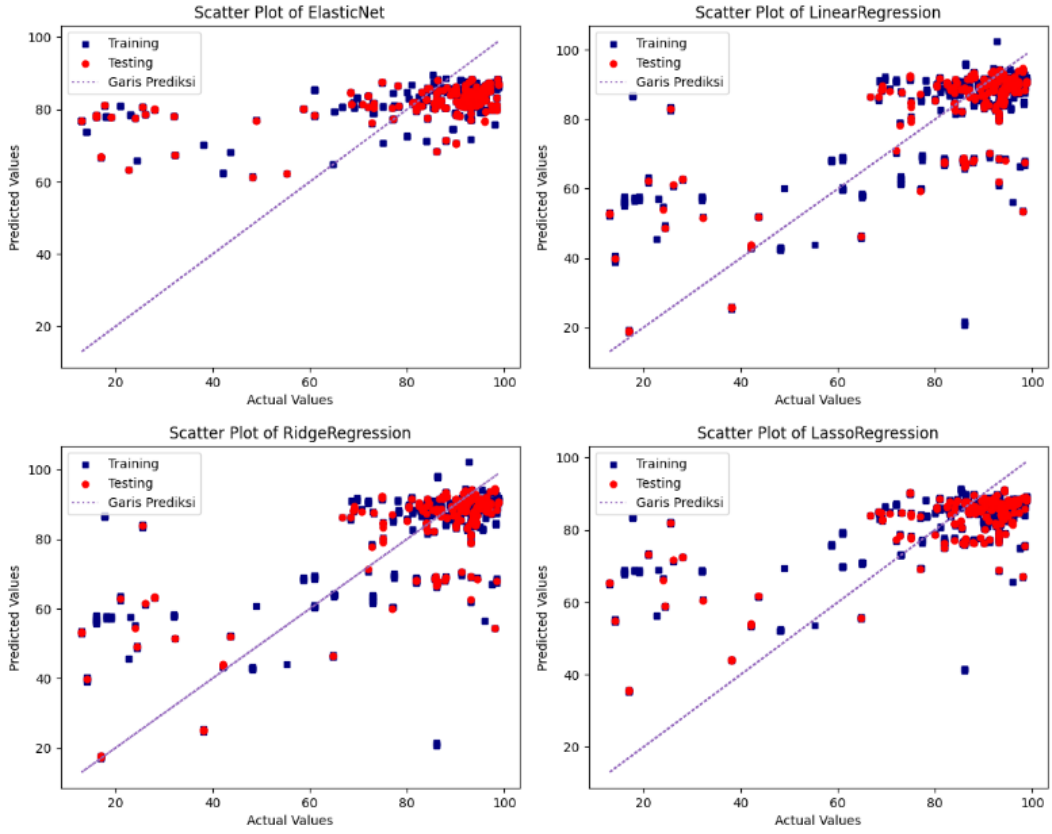
Gambar 5. *Sebaran data setelah penggunaan KDE*

Gambar 5 menunjukkan sebaran nilai variabel LUMO (*Lowest Unoccupied Molecular Orbital*) pada sumbu horizontal dan HOMO (*Highest Occupied Molecular Orbital*) pada sumbu

vertikal yang terlihat lebih beragam dan luas. Dengan menambahkan sampel virtual ini ke dalam *dataset* awal, diupayakan peningkatan kemampuan model sehingga bisa memberikan prediksi yang lebih akurat dan stabil di berbagai kondisi data.

Tabel 3. *Kinerja model Prediksi Linier dengan KDE*

Model Linier	Training		Testing	
	R^2	RMSE	R^2	RMSE
MLR	0.485	0.150	0.540	0.158
Ridge	0.481	0.151	0.540	0.158
Lasso	0.384	0.164	0.403	0.183
ElasticNet	0.195	0.188	0.171	0.213

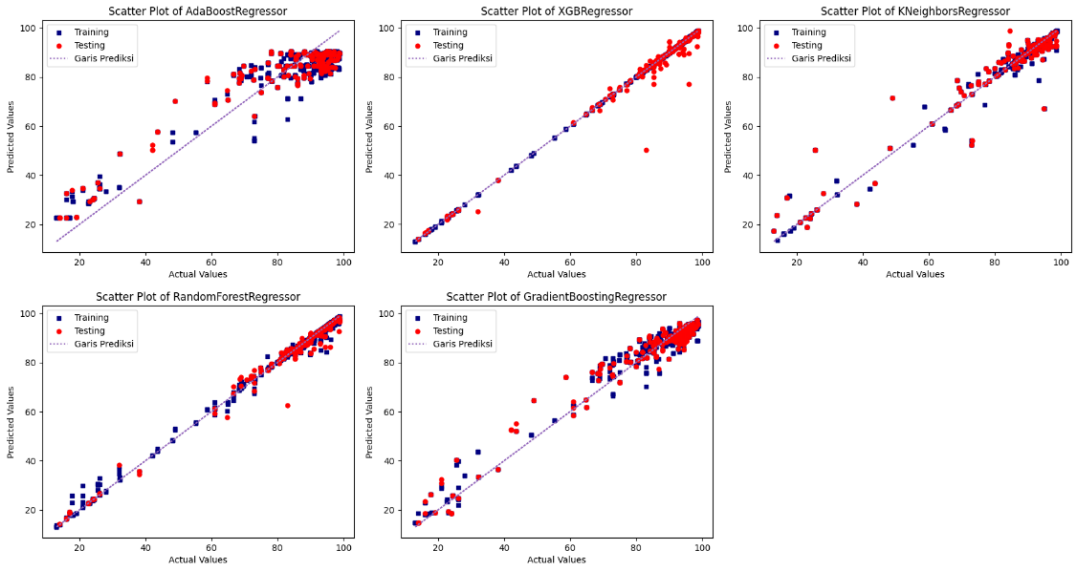


Gambar 6. *Scatter plot Model Linier menggunakan KDE*

Hasil model linier (MLR, *Ridge*, *Lasso*, dan *ElasticNet*) pada Tabel 3 dan gambar 6, menunjukkan peningkatan performa yang minimal. Pada data uji, nilai R^2 model linier terbaik, yaitu MLR dan *Ridge*, hanya mencapai 0,540 dengan RMSE sebesar 0,158. Model *Lasso* menunjukkan performa yang sedikit lebih rendah dengan R^2 sebesar 0,403 dan RMSE sebesar 0,183, sementara *ElasticNet* tetap memiliki performa terendah dengan R^2 sebesar 0,171 dan RMSE sebesar 0,213. Performa yang terbatas ini menunjukkan bahwa model linier tidak banyak merespons terhadap penambahan variasi data melalui KDE. Hal ini kemungkinan besar disebabkan oleh keterbatasan model linier dalam menangkap pola non-linier yang kompleks dalam data prediksi *inhibitor* korosi (IE%).

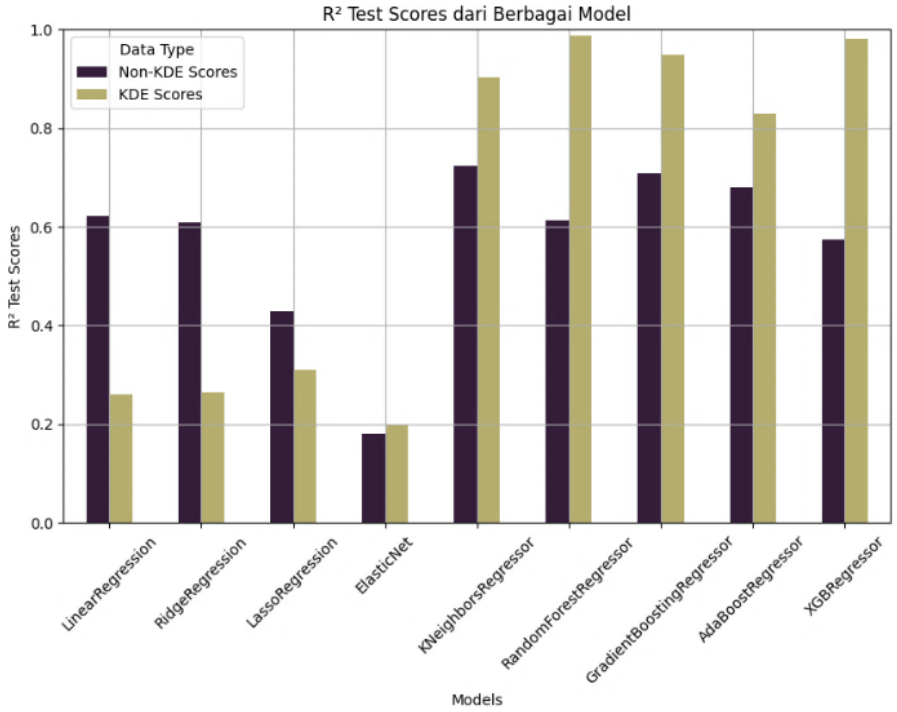
Tabel 4. *Kinerja model Prediksi Non-Linier dengan KDE*

Model Non-Linier	Training		Testing	
	R^2	RMSE	R^2	RMSE
KNR	0.948	0.045	0.918	0.058
RFR	0.995	0.013	0.975	0.032
GBR	0.966	0.038	0.944	0.048
ABR	0.843	0.082	0.821	0.087
XGBR	0.999	0.001	0.967	0.037



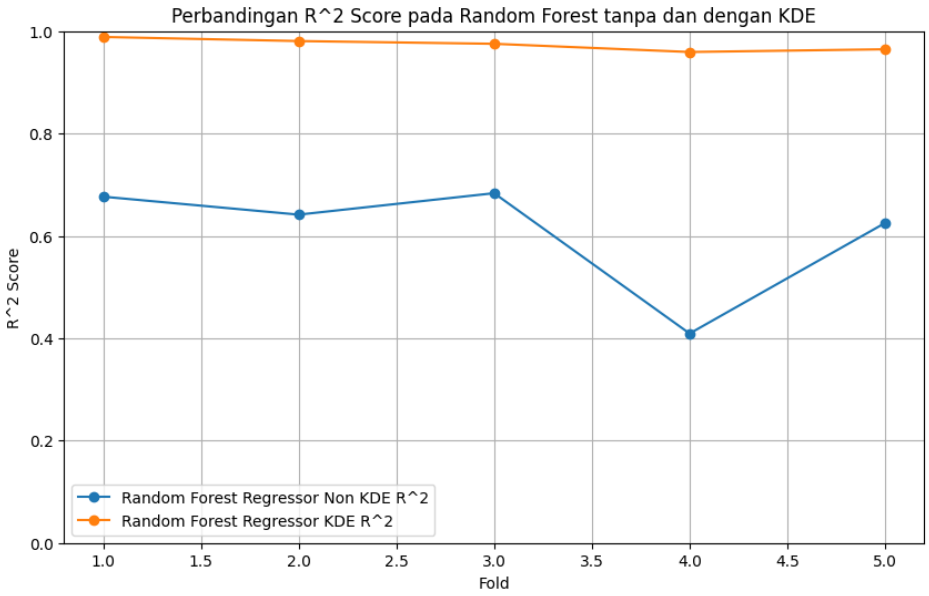
Gambar 7. *Scatter plot Model Non-Linier menggunakan KDE*

Setelah penggunaan KDE R^2 dan RMSE pada model non-linier meningkat secara drastis, dapat dilihat pada tabel 4 dan gambar 7. Misalnya, RFR dan GBR yang awalnya memiliki nilai R^2 sebesar 0,636 dan 0,717 pada data uji, naik menjadi 0,975 dan 0,944, dengan RMSE yang menurun drastis ke 0,032 dan 0,048. Model XGBR juga menunjukkan peningkatan performa yang signifikan, dengan R^2 naik menjadi 0,967 dan RMSE menurun ke 0,037. Hal ini menunjukkan bahwa penggunaan KDE berkontribusi besar dalam meningkatkan keakuratan dan ketepatan model non-linier melalui perluasan data sampel, yang membantu model mengenali pola lebih baik.



Gambar 8. *Perbandingan R2 Score sebelum dan sesudah menggunakan KDE*

Grafik pada gambar 8, ini membandingkan skor R² berbagai model regresi pada data dengan dan tanpa penerapan *Kernel Density Estimation* (KDE). Model linier seperti *LinearRegression* dan RidgeRegression justru mengalami penurunan performa dengan KDE, sedangkan model ensemble seperti *RandomForest*, *GradientBoosting*, dan *XGBRegressor* menunjukkan peningkatan yang signifikan, mencapai skor mendekati 1. Hasil ini menunjukkan bahwa KDE lebih efektif dalam meningkatkan akurasi model-model kompleks, terutama model ensemble, sementara efeknya pada model linier relatif minimal.



Gambar 9. *Sebaran data setelah penggunaan KDE*

Dari hasil sebelumnya didapatkan bahwa *Random Forest Regressor* mendapatkan hasil R2 paling seimbang dan nilai RMSE terendah. Gambar 9 menampilkan perbandingan skor R² untuk model *Random Forest Regressor* dengan dan tanpa penggunaan *Kernel Density Estimator*

(KDE) dalam proses 5-fold cross-validation. Sumbu vertikal menampilkan skor R^2 yang mengukur seberapa baik model sesuai dengan data sebenarnya, dan pada sumbu horizontal ditampilkan jumlah fold dalam *cross-validation*. Garis oranye menunjukkan skor R^2 untuk model dengan KDE, yang secara konsisten mendekati nilai 1,0 di setiap fold. Sebaliknya, garis biru menggambarkan performa model tanpa KDE, dengan nilai R^2 yang lebih rendah, berkisar antara 0,5 hingga 0,7, dan menunjukkan fluktuasi di tiap fold. Hasil ini mengindikasikan bahwa penerapan KDE secara signifikan meningkatkan performa model *Random Forest*, menghasilkan prediksi yang lebih akurat dan konsisten. Peningkatan ini menunjukkan bahwa KDE berhasil memperkaya variasi data, membantu model mengenali pola lebih baik dan meningkatkan stabilitas prediksi pada IE% di seluruh *fold cross-validation*.

5. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan *Kernel Density Estimator* (KDE) dapat meningkatkan performa model prediksi inhibitor korosi (IE%) dari senyawa N-Heterocyclic, khususnya pada model non-linier. Tanpa KDE, model non-linier seperti *Random Forest Regressor* (RFR) dan *Gradient Boosting Regressor* (GBR) sudah menunjukkan performa yang lebih baik dibandingkan model linier (MLR, *Ridge*, *Lasso*, dan *ElasticNet*), tetapi masih mengalami fluktuasi dan keterbatasan akurasi. Penggunaan KDE untuk menghasilkan data sintetik terbukti memperluas distribusi data, yang berdampak pada peningkatan nilai R^2 dan penurunan RMSE, terutama pada model non-linier.

Selain itu, model RFR dengan KDE mencapai skor R^2 yang mendekati 1,0 secara konsisten, menunjukkan akurasi dan stabilitas prediksi yang tinggi, sementara model tanpa KDE masih menunjukkan variasi skor R^2 yang signifikan. Penelitian ini membuktikan bahwa penggunaan KDE sebagai metode augmentasi data sangat efektif dalam meningkatkan akurasi dan kemampuan generalisasi model non-linier. Hal ini menjadikan KDE strategi yang bermanfaat dalam pemodelan *machine learning* untuk data yang memerlukan pola kompleks. Oleh karena itu, KDE direkomendasikan untuk diintegrasikan dalam *pipeline* pemodelan untuk optimasi performa, terutama pada aplikasi serupa yang memerlukan ketepatan prediksi tinggi.

DAFTAR PUSTAKA

[1] N. V. Putranto, M. Akrom, and G. A. Trinapradika, "Implementasi Fungsi Polinomial pada Algoritma Gradient Boosting Regressor: Studi Regresi pada Dataset Obat-Obatan Kadaluarsa Sebagai Material Antikorosi," JTMI, vol. 9, no. 2, pp. 172-182, Dec. 2023, doi: https://doi.org/10.26905/jtmi.v9i2.11192

[2] T. Sutojo, S. Rustad, M. Akrom, A. Syukur, G. F. Shidik, and H. K. Dipojono, "A machine learning approach for corrosion small datasets," npj Mater Degrad, vol. 7, no. 1, p. 18, Mar. 2023, doi: https://doi.org/10.1038/s41529-023-00336-7

[3] M. Akrom, "DFT Investigation of Syzygium Aromaticum and Nicotiana Tabacum Extracts as Corrosion Inhibitor," Science Tech: Jurnal Ilmu Pengetahuan dan Teknologi, vol. 8, no. 1, pp. 42-48, Feb. 2022, doi: https://doi.org/10.30738/st.vol8.no1.a11775

[4] S. Harsimran, K. Santosh, and K. Rakesh, "Overview Of Corrosion And Its Control: A Critical Review," PES, vol. 3, no. 1, pp. 13-24, Mar. 2021, doi: https://doi.org/10.24874/PES03.01.002

[5] 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," Computational and Theoretical Chemistry, vol. 1229, p. 114307, 2023, doi: https://doi.org/10.1016/j.comptc.2023.114307.

[6] N. Islami, M. Ihsan, T. Hafli, R. Putra, and M. Muhammad, "Pengaruh Lingkungan Korosif dan Beban Mekanis Terhadap Perilaku Korosi pada Material Stainless Steel AISI-304," MJMST, vol. 5, no. 2, p. 28, Oct. 2021, doi: https://doi.org/10.29103/mjmst.v5i2.6025

[7] W. Wibowo and M. N. Ilman, "Studi Eksperimental Pengendalian Korosi pada Aluminium 2024-T3 di Lingkungan Air Laut Melalui Penambahan Inhibitor Kalium

Kromat (K2CrO4)," Jurnal Rekayasa Proses, vol. 5, no. 1, 2011, doi: https://doi.org/10.22146/jrekpros.1893.

[8] G. Priyotomo, S.T., M.Si., H. Sumada Sitepu, and Y. Dwiyanti, "Pengaruh Penambahan Konsentrasi Inhibitor Ekstrak Daun Talas Terhadap Laju Korosi Pada Baja Api 5L X-52 Dengan Media Korosif H2SO4 0,5 M," j. widyariset, vol. 5, no. 1, p. 30, Mar. 2020, doi: https://doi.org/10.14203/widyariset.5.1.2019.30-36

[9] A. Miralrio and A. Espinoza Vázquez, "Plant Extracts as Green Corrosion Inhibitors for Different Metal Surfaces and Corrosive Media: A Review," Processes, vol. 8, no. 8, p. 942, Aug. 2020, doi: https://doi.org/10.3390/pr8080942

[10] A. Hu et al., "A new framework for predicting tensile stress of natural rubber based on data augmentation and molecular dynamics simulation data," J Mater Inf, vol. 4, no. 3, Aug. 2024, doi: https://doi.org/10.20517/jmi.2024.11

[11] Q.-X. Zhu, Z.-H. Wang, Y.-L. He, and Y. Xu, "A Monte Carlo and Kernel Density Estimation based virtual sample generation method for small data modeling problem," in 2020 Chinese Automation Congress (CAC), Shanghai, China: IEEE, Nov. 2020, pp. 1123-1128. doi: https://doi.org/10.1109/CAC51589.2020.9326486

[12] L. Zhang et al., "Probability prediction of short-term user-level load based on random forest and kernel density estimation," Energy Reports, vol. 8, pp. 1130-1138, Aug. 2022, doi: https://doi.org/10.1016/j.egyr.2022.02.256

[13] N. A. Matar, W. Matar, and T. AlMalahmeh, "Predictive Model for Students Admission Uncertainty Using Naïve Bayes Classifier and Kernel Density Estimation (KDE)," Int. J. Emerg. Technol. Learn., vol. 17, no. 08, pp. 75-96, Apr. 2022, doi: https://doi.org/10.3991/ijet.v17i08.29827

[14] F. Novianti and N. Ulinnuha, "Seleksi Fitur Algoritma Genetika Dalam Klasifikasi Data Rekam Medis PCOS Menggunakan SVM," vol. 9, no. 1, 2024, doi: https://doi.org/10.21107/nero.v9i1.25399

[15] C. A. P. Sumarjono, M. Akrom, and G. A. Trisnapradika, "Perbandingan Model Machine Learning Terbaik untuk Memprediksi Kemampuan Penghambatan Korosi oleh Senyawa Benzimidazole," tc, vol. 22, no. 4, pp. 973-980, Nov. 2023, doi: https://doi.org/10.33633/tc.v22i4.9201

[16] M. Akrom, S. Rustad, and H. Kresno Dipojono, "Machine learning investigation to predict corrosion inhibition capacity of new amino acid compounds as corrosion inhibitors," Results in Chemistry, vol. 6, p. 101126, Dec. 2023, doi: https://doi.org/10.1016/j.rechem.2023.101126

[17] M. Akrom and T. Sutojo, "Investigasi Model Machine Learning Berbasis QSPR pada Inhibitor Korosi Pirimidin," Eksergi, vol. 20, no. 2, p. 107, Jul. 2023, doi: https://doi.org/10.31315/e.v20i2.9864

[18] S. Budi et al., "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," KEG, Mar. 2024, doi:. https://doi.org/10.18502/keg.v6i1.15351

[19] M. Akrom, "Green Corrosion Inhibitors for Iron Alloys: A Comprehensive Review of Integrating Data-Driven Forecasting, Density Functional Theory Simulations, and Experimental Investigation," JIMAT, vol. 1, no. 1, pp. 22-37, Apr. 2024, doi: https://doi.org/10.62411/jimat.v1i1.10495

[20] W. Herowati et al., "Prediction of Corrosion Inhibition Efficiency Based on Machine Learning for Pyrimidine Compounds: A Comparative Study of Linear and Non-linear Algorithms," KnE Engineering, vol. 6, no. 1, pp. 68-77, Mar. 2024, doi: https://doi.org/10.18502/keg.v6i1.15350