Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.2, April 2025, pp. 382~389
e-ISSN: 2548-6861

Effect of Virtual Sample Generation in Predicting Corrosion Inhibition Efficiency on Pyridazine

Ilham Pratama Aldiansah [1]*, Muhamad Akrom [2]**
* Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
** Research Center for Quantum Computing and Materials Informatics, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
111202113885@mhs.dinus.ac.id [1], m.akrom@dsn.dinus.ac.id [2]

Article Info

Article history:

Received 2025-01-30
Revised 2025-02-20
Accepted 2025-02-21

Keyword:

Corrosion,
Inhibitor,
Linear Interpolation,
Gaussian Noise Augmentation,
Virtual Sample Generation.

ABSTRACT

The purpose of this research is to study how the application of virtual sample generation using the linear interpolation and gaussian noise augmentation method impacts the improvement of prediction model performance in the case of corrosion inhibition efficiency using pyridazine. Random Forest Regressor, Gradient Boosting Regressor, and Bagging Regressor are the models used. The coefficient of determination (R^2) values for each model are -0.06, 0.05, and 0.12 on the initial data; the RMSE values are 34.80, 32.90, and 31.65, respectively. After the use of virtual sample development, the R^2 values significantly increased to 0.99, 0.96, and 0.99, while the RMSE values significantly decreased to 1.59, 2.88, and 1.25. The research results show that the linear interpolation method can enrich the dataset without altering the data distribution pattern, this method significantly improves the model's accuracy. This performance improvement demonstrates the ability of virtual sample generation to overcome the limitations of the original data; ultimately, this results in a more accurate and reliable predictive model. In the field of material efficiency prediction especially for material technology applications and corrosion control this research helps develop data augmentation methods for similar cases.

I. PENDAHULUAN

Korosi adalah fenomena alam yang terjadi ketika material, khususnya logam, terdegradasi ketika berinteraksi dengan hal-hal di sekitarnya, seperti air, oksigen, atau gas korosif lainnya[1] .Proses ini dapat menyebabkan kerusakan yang signifikan pada struktur material, yang mengakibatkan penurunan kualitas dan daya tahan komponen[2]. Dalam lingkungan industri, korosi juga dapat menyebabkan biaya perawatan yang lebih tinggi, umur komponen yang diperpanjang, dan penurunan kinerja operasional[3].

Salah satu cara yang telah lama digunakan untuk mengurangi dampak korosi adalah dengan menggunakan inhibitor korosi[4]. Inhibitor korosi bekerja dengan membentuk lapisan pelindung pada permukaan logam atau dengan mengubah sifat kimia lingkungan material tersebut[5]. Untuk potensi mereka sebagai inhibitor korosi yang efektif, berbagai senyawa kimia[6], termasuk senyawa heterosiklik seperti pyridazine, telah diteliti. Senyawa dengan cincin heterosiklik ini diketahui memiliki sifat adsorpsi yang baik pada permukaan logam, yang memungkinkannya untuk menghambat proses korosi secara signifikan.

Namun, penelitian eksperimental sering kali memerlukan banyak waktu dan sumber daya untuk menemukan inhibitor korosi yang efektif[7]. Metode tradisional yang bergantung pada pengujian fisik tidak hanya mahal tetapi juga memakan waktu lama, terutama jika dilakukan pada berbagai konsentrasi senyawa atau kondisi lingkungan[8], [9]. Di sinilah simulasi komputasi dan sample virtual menjadi sangat penting. Model komputasional yang menggunakan data simulasi memungkinkan kita untuk mengeksplorasi berbagai kondisi dan senyawa dalam waktu singkat tanpa harus melakukan percobaan fisik yang mahal[10].

Data atau sampel virtual dapat dibuat dengan menggunakan simulasi komputer dari kondisi atau objek dunia nyata. Dalam konteks ini, virtual sample digunakan untuk memodelkan interaksi antara senyawa yang menghambat korosi (seperti pyridazine) dengan permukaan

logam dalam kondisi tertentu[11]. Teknik ini dapat mempercepat proses eksperimen karena peneliti dapat dengan cepat mengidentifikasi senyawa yang dapat berfungsi sebagai inhibitor korosi yang efektif bahkan sebelum uji laboratorium dimulai[12]. Ini tidak hanya menghemat waktu dan uang, tetapi juga memungkinkan untuk memprediksi bagaimana inhibitor berfungsi dalam berbagai kondisi[13].

Dataset pyridazine memiliki potensi besar untuk mengembangkan inhibitor korosi berbasis virtual sample. Kita dapat memodelkan pengaruh struktur kimia pyridazine terhadap kemampuan inhibisinya terhadap korosi dengan menggunakan teknik *machine learning* dan algoritma simulasi[14]. Metode berbasis data ini dapat meningkatkan prediksi dan memungkinkan peneliti menemukan senyawa dengan efisiensi yang lebih tinggi daripada metode konvensional. Pemodelan prediktif berbasis algoritma pengajaran mesin adalah salah satu metode yang dapat digunakan. Metode ini menganalisis kumpulan data yang ada untuk menemukan pola yang dapat digunakan untuk memprediksi efektivitas inhibitor korosi pada material tertentu[15].

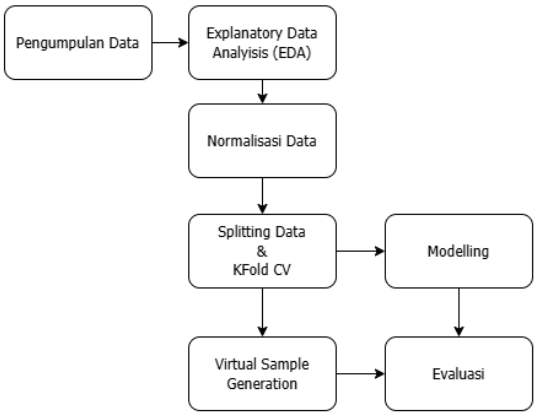
Tujuan utama penelitian ini adalah untuk mengetahui bagaimana penerapan sampel virtual terhadap kumpulan pyridazine dapat membantu memprediksi dan merancang inhibitor korosi yang lebih baik[16]. Diharapkan bahwa dengan menggunakan metode ini, akan ditemukan hubungan yang lebih mendalam antara struktur kimia pyridazine dan kemampuan untuk menghambat korosi pada material logam[17]. Penelitian ini juga bertujuan untuk menemukan cara untuk mempercepat pengujian inhibitor korosi, yang sebelumnya membutuhkan banyak percobaan eksperimental, dengan lebih hemat biaya dan efisien[18].

Dengan menggunakan sampel virtual pada kumpulan data pyridazine[19], simulasi menjadi lebih mudah dan formulasi inhibitor korosi dapat dioptimalkan dengan data yang lebih luas dan tepat. Dengan cara ini, pembuatan bahan yang lebih tahan terhadap korosi dapat dilakukan dengan lebih cepat dan dengan biaya yang lebih rendah. Ini membantu meningkatkan daya tahan dan efisiensi material yang digunakan dalam berbagai industri[20].

Oleh karena itu, penelitian ini bertujuan untuk mengetahui bagaimana pengaruh Virtual Sample Generation dalam meningkatkan performa model dalam memprediksi efisiensi penghambatan korosi pada senyawa pyridazine[14]. Dengan menggunakan 2 metode *interpolasi linear* dan *gaussian noise augmentation* untuk membuktikan lebih efektif mana dari 2 metode tersebut dalam meningkatkan performa model terhadap senyawa pyridazine[21].

II. METODE

Pada penelitian ini menggunakan beberapa alur pengembangan model machine learning, dimulai dari pengumpulan data, *explanatory data analysis* (EDA), normalisasi data, splitting data, modelling, lanjut implementasi *virtual sample generation* (VSG), serta evaluasi untuk menghitung hasil performa dari masing-masing model.



Gambar 1. Pengembangan Model ML

A. *Pengumpulan Data*

Pada penelitian ini, kami menggunakan dataset senyawa pyridazine yang berjumlah 21 yang digunakan sebagai inhibitor korosi. Dataset ini dikumpulkan dari literatur terkait yang berfokus pada peningkatan efisiensi inhibitor [1]. Dataset ini mempunyai 11 fitur (variabel independent) dan 1 target yang digunakan. Fitur pada molekul tersebut yaitu HOMO (eV), LUMO (eV), L-H (energi gap), μ (momen dipol), IP (potensial ionisasi), EA (afinitas elektron), χ (elektronegativitas), η (global hardness), σ (global softness), ΔN (fraksi elektron yang ditransfer), ω (Elektrofilisitas). Dan *Inhibitor efficiency* (IE) sebagai nilai kasus prediksi peningkatan efisiensi penghambatan korosi.

B. *Explanatory Data Analysis (EDA)*

Kemudian dilanjutkan dengan melakukan pre-processing data dengan *Explanatory Data Analysis* (EDA). *Explanatory Data Analysis* (EDA) adalah bagian penting dari analisis data. Tujuan EDA adalah untuk mempelajari dan memahami pola-pola yang ada dalam dataset sebelum melanjutkan ke pemodelan atau analisis statistik lebih lanjut[21]. Untuk memberikan gambaran yang lebih jelas tentang atribut data yang sedang diteliti, EDA dapat dilakukan dengan menggunakan berbagai teknik deskriptif statistik dan visual. Ada beberapa elemen utama yang dilakukan untuk EDA yaitu pembersihan data dan persiapan data, statistik deskriptif, visualisasi data, deteksi outliers, hubungan antar variabel.

C. *Normalisasi Data*

Normalisasi data sangat penting untuk analisis regresi yang melibatkan variabel dengan distribusi yang tidak normal atau nilai pencilan. *Robust Scaler* adalah salah satu metode normalisasi yang efektif untuk menangani data dengan outliers[22]. Normalisasi pada dasarnya bertujuan untuk mengubah skala variabel sehingga semua variabel atau fitur dataset sama atau memiliki skala yang sama. Ini sangat penting karena algoritma regresi sensitif terhadap perbedaan skala antara variabel. Metode normalisasi yang dikenal sebagai Robust Scaler mengubah skala data dengan menggunakan median dan rentang *interquartile* (IQR)[23]. Dibandingkan dengan metode seperti *MinMax Scaling*, yang

bergantung pada nilai minimum dan maksimum, atau *Standard Scaling*, yang menggunakan rata-rata dan deviasi standar, *Robust Scaler* lebih robust (tahan) terhadap keberadaan *outliers*.

D. Splitting Data

Pada tahap ini dilakukan pembagian data menjadi training dan testing. Langkah penting dalam pengembangan model regresi adalah pembagian data pengujian. Tujuan utama pembagian ini adalah untuk melatih model menggunakan data *training* (80 persen dari dataset) dan menguji model pada data *testing* (20 persen dari dataset) untuk menilai kinerja model secara objektif dan memastikan bahwa model tidak melebihi data pelatihan. Dengan memastikan bahwa model dilatih pada data pelatihan dan diuji pada data pengujian yang terpisah, kita dapat memperoleh hasil yang lebih objektif dan menghindari *overfitting*, sehingga model yang dihasilkan lebih mampu menggeneralisasi dan memberikan prediksi yang akurat pada data baru. Serta melakukan validasi tambahan yaitu menggunakan K-Fold Cross Validation membagi dataset menjadi bagian K atau lipatan, juga dikenal sebagai folds, dengan ukuran yang hampir sama. K iterasi digunakan untuk melakukan validasi, dengan satu fold digunakan sebagai data uji dan sisa fold digunakan sebagai data latih. Sampai setiap fold menjadi data uji sekali, proses ini berulang[18]. Menghitung rata-rata performa model untuk setiap iterasi menghasilkan hasil evaluasi akhir. Pada penelitian ini kami menggunakan beberapa nilai k yaitu 2, 5, 8,9,10 kemudian mendapatkan hasil terbaik pada nilai k = 5.

E. Virtual Sample Generation (VSG)

Virtual sample generation adalah metode untuk menghasilkan sampel data sintetis yang dapat digunakan untuk memperluas ukuran dataset saat ini atau untuk menggali pola tertentu dalam data[24]. Dalam konteks prediksi peningkatan penghambatan korosi dengan dataset pyridazine, Pada penelitian ini kami menggunakan 2 metode yaitu *linear interpolation* dan *Gaussian Noise Agumentation* dimana untuk digunakan lebih bagus mana dalam segi efektivitasnya. dapat dilakukan dengan menginterpretasikan pola linear yang ada dalam hubungan antara variabel-variabel, seperti konsentrasi pyridazine (variabel independen) dan penghambatan korosi[25]. Tujuan dari interpretasi pola linear adalah untuk mendapatkan pemahaman tentang hubungan fungsional antara dua variabel. Jika hubungan antara variabel independen dan dependen bersifat linear, kita dapat membuat prediksi dan menghasilkan data sintetis yang menggambarkan hubungan tersebut. Untuk eksperimen tambahan atau mengatasi keterbatasan data, metode ini dapat sangat bermanfaat.

F. Modelling

Pada tahap modelling ini, kami menggunakan beberapa model ensemble learning seperti Random Forest, Gradient Boosting Regressor dan Bagging Regressor. Serta proses ini menggunakan variabel independent (fitur X) dan variabel dependen (target Y) untuk menguji model seberapa bagus pada dataset tersebut.

G. Evaluasi

Untuk memastikan bahwa model regresi dapat membuat prediksi yang akurat dan dapat diandalkan, sangat penting untuk melakukan evaluasi kinerjanya. *Root Mean Squared Error* (RMSE) dan *R-Squared* (R^2) adalah dua metrik evaluasi yang sering digunakan untuk menilai kinerja model regresi. Seberapa baik model memprediksi nilai yang sebenarnya ditunjukkan oleh dua metrik ini. Kedua metrik ini akan menunjukkan kemampuan model untuk menjelaskan hubungan antara konsentrasi pyridazine dan penghambatan korosi dalam konteks regresi peningkatan penghambatan korosi dengan dataset pyridazine[26]. *Root Mean Squared Error* (RMSE): Nilai RMSE yang tinggi menunjukkan bahwa model kurang akurat dalam memprediksi penghambatan korosi, dan nilai RMSE yang rendah menunjukkan bahwa model memiliki kesalahan prediksi yang kecil. *R-squared* (R^2): Mengukur sejauh mana variabilitas data dapat dijelaskan oleh model. Nilai R^2 yang lebih tinggi menunjukkan bahwa model lebih baik dalam menjelaskan variabilitas data penghambatan korosi berdasarkan konsentrasi pyridazine. Pada evaluasi ini kami juga menambahkan uji statistik dengan *Kolmogrov-Smirnov* (KS test) Digunakan untuk membandingkan dua distribusi sampel atau untuk menguji kesesuaian distribusi data dengan distribusi referensi, uji *Kolmogorov-Smirnov* (KS) dapat diterapkan pada data kontinu dan memiliki keunggulan karena tidak bergantung pada asumsi distribusi tertentu. Artikel ini membahas konsep dasar, teknik, dan bagaimana uji *Kolmogorov-Smirnov* dapat digunakan dalam berbagai bidang penelitian. Pada penelitian ini digunakan dalam melihat distribusi data asli dengan *augmentasi* apakah memiliki perbedaan yang signifikan atau tidak.

III. HASIL DAN PEMBAHASAN

A. Hasil

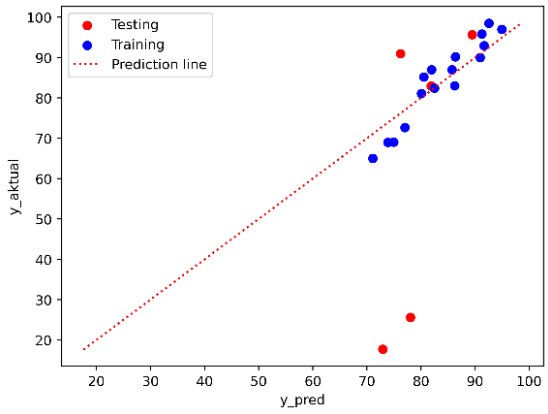
Pada perbandingan kinerja model prediksi Random Forest, Gradient Boosting Regressor dan Bagging Regressor yang ditampilkan pada tabel dibawah ini dengan menggunakan matriks evaluasi nilai dari R^2 dan RMSE. Model yang mempunyai hasil yang bagus adalah yang memiliki nilai R^2 mendekati 1 dan nilai RMSE semakin rendah semakin baik.

TABEL I.
PERFORMA PREDIKSI MODEL

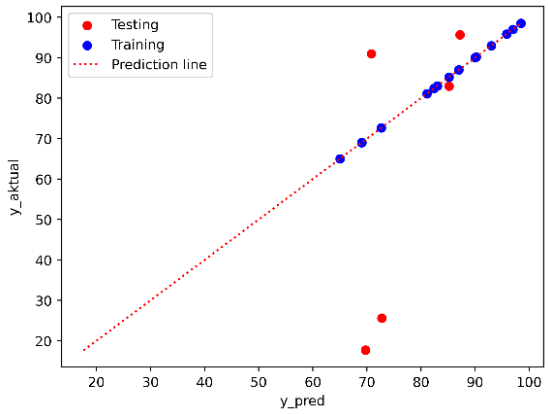
Model	Performa	
	R^2	RMSE
Random Forest	-0.06	34.80
Gradient Boosting Regressor	0.05	32.90
Bagging Regressor	0.12	31.65

Tabel I Performa prediksi model diatas menunjukkan bahwa data yang digunakan tidak cocok dengan model tersebut. Akan tetapi masih bisa ditingkatkan dengan beberapa metode lain-nya. Kemudian, pada gambar dibawah

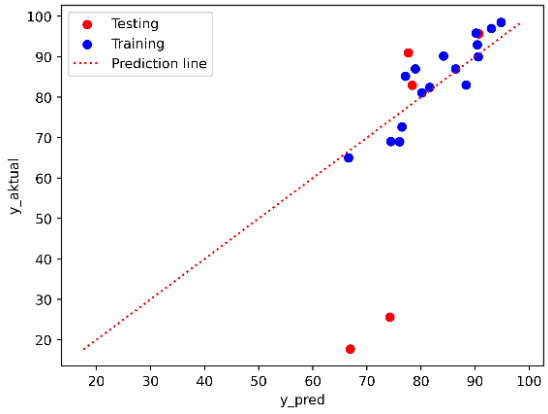
ini merupakan hasil plot dari masing-masing model sebelum diberikan Virtual Sample Generation dengan interpretasi pola linear yang saling menghubungkan titik satu dengan titik lain-nya. Dataset yang hanya memiliki jumlah sedikit juga dapat mempengaruhi hasil performa terhadap masing-masing model yang digunakan, seperti hal-nya gambar visualiasi dari hasil plotting line dibawah ini. dari gambar tersebut menujukkan bahwa dari ketiga model tersebut data training lebih mendekati nilai sebenarnya.



Gambar 2. Sebaran data pada model random forest



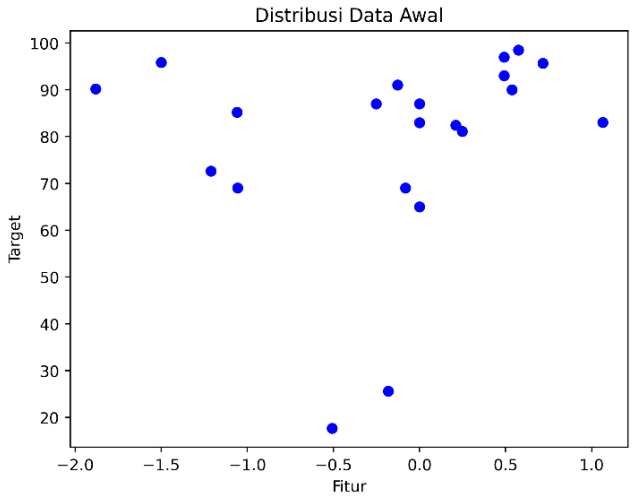
Gambar 3. Sebaran data pada model GBR



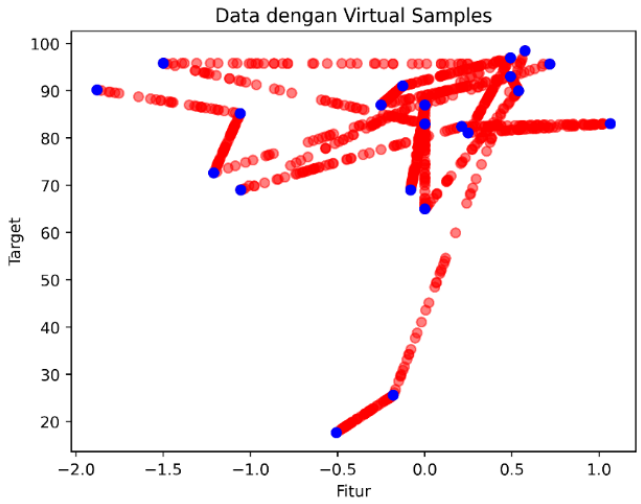
Gambar 4. Sebaran data pada model bagging regressor

Dari ketiga model tersebut setelah mendapat hasil prediksi pada nilai sebenarnya masi menunjukkan bahwa model belum mampu memberikan kinerja yang cukup baik, oleh karena itu kami mencoba untuk melakukan virtual sample generation dengan melakukan penambahan data sintetik pada dataset pyridazine dengan tujuan mampu memberikan pengaruh dalam kinerja model memprediksi nilai sebenarnya dengan hasil yang lebih baik dari sebelumnya. Dengan menggunakan interpolasi linier, jumlah data bertambah secara signifikan, tergantung pada parameter (*n_samples*) yang ditentukan. Hal ini memberikan dataset yang lebih kaya dan lebih representatif untuk pelatihan model. Pada penelitian ini, dengan menambahkan data sintetik menjadi 1000 sebagai (*n_samples*). Karena interpolasi linier mempertahankan hubungan linier antara fitur dan target, data sintetik yang dihasilkan tidak menambahkan noise berlebih atau mengubah struktur dasar data. Hal ini penting untuk menjaga interpretabilitas model. Dengan menambahkan data sintetik, model machine learning yang dilatih di atas dataset ini cenderung memiliki generalisasi yang lebih baik karena data tambahan mengurangi efek *overfitting*, terutama jika data asli terbatas. Data tambahan ini membantu meningkatkan kemampuan model untuk belajar, terutama jika data asli memiliki jumlah sampel kecil. Jika fitur (variabel X) memiliki lebih dari satu fitur, interpolasi dilakukan dalam ruang multidimensi. Hal ini memungkinkan data sintetik untuk mencakup keseluruhan manifold data asli, membuat model lebih robust terhadap variasi. Dalam contoh-nya Interpolasi linier adalah metode augmentasi yang mudah diterapkan tanpa memerlukan algoritma kompleks. Teknik ini tanpa menambah kompleksitas yang signifikan, mempertahankan pola struktural, meningkatkan jumlah data, dan membantu model belajar lebih baik. Namun, agar performa model tetap optimal, metode ini harus dikombinasikan dengan metode augmentasi lain untuk data dengan pola yang lebih kompleks. Dari hasil proses pembuatan data sintetik yang kemudian digabungkan lagi menjadi satu dengan dataset yang semula. Berikut visualisasi data setelah dilakukan virtual sample generation dengan interpolasi linear.

Data Asli yang menunjukkan warna biru menampilkan distribusi data awal yang menjadi dasar *interpolasi*. Pola data asli menunjukkan hubungan linier antara fitur (sumbu x) dan target (sumbu y). Data Sintetik yang berwana merah menunjukkan data tambahan hasil interpolasi linier. Distribusi titik-titik merah terlihat sejajar dengan pola data asli, mencerminkan bahwa metode interpolasi linier mempertahankan hubungan linier dalam dataset.



Gambar 5. Visualisasi data awal



Gambar 6. Visualisasi data setelah interpolasi linear

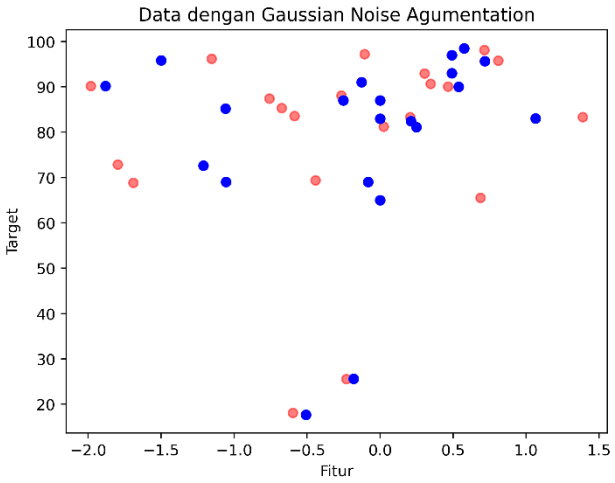
Setelah melakukan *Virtual Sample Generation* dengan menggunakan interpolasi yang dapat dilihat dari gambar diatas. Virtual sample generation dengan interpolasi linier memberikan manfaat signifikan dalam meningkatkan jumlah data dan melestarikan hubungan linier dalam dataset. Hal ini dapat membantu model machine learning dalam meningkatkan performa prediksi dan memperbaiki generalisasi model. Setelah melakukan penambahan data sintetis, maka melakukan uji kinerja dari ketiga model yang sudah digunakan sebelumnya. Dengan konfigurasi perbandingan data *training* dan *testing* yang sama seperti sebelumnya, setelah melakukan proses yang sama dengan data yang sudah ditambah dengan 1000 data sintetis dan mendapatkan hasil dengan melihat nilai matriks evaluasi sama seperti diatas. Nilai R^2 dan RMSE sebagai nilai evaluasi yang dapat dilihat dibawah ini.

TABEL II.
PERFORMA MODEL SETELAH LINEAR INTERPOLATION

Model	Performa	
	R2	RMSE
Random Forest	0.99	1.59
Gradient Boosting Regressor	0.96	2.88
Bagging Regressor	0.99	1.25

Dapat dilihat bahwa virtual sample generation dengan interpolasi linear mampu meningkatkan performa dari ketiga model yaitu nilai R^2 Random Forest, Gradient Boosting Reg dan Bagging Reg yang semula -0.06, 0.05 dan 0.12 meningkat menjadi 0.99, 0.96. dan 0.99 menunjukkan bahwa metode diatas cukup baik dalam memberikan pengaruh kinerja model *machine learning* pada dataset pyridazine dengan kasus prediksi meningkatkan efisiensi inhibitor korosi, dikarenakan data yang semula hanya sedikit dan membuat hasil uji kinerja sebelumnya cukup buruk dengan nilai R^2 dibawah 0.2 serta nilai RMSE yang melebihi dari 10. Karena interpolasi linier menghasilkan data baru yang sesuai dengan pola linier asli, model regresi linear dapat mempelajari hubungan ini dengan lebih baik.

Kemudian dari ketiga model yang di atas, kami mencoba menggunakan metode yang lain yaitu *Gaussian noise augmentation* untuk kita bandingkan hasilnya dengan melihat hasil dari performa model tersebut. Dengan menambahkan noise acak dari distribusi Gaussian (Normal) ke dalam data, metode augmentasi suara Gaussian bertujuan untuk meningkatkan ketahanan model terhadap variasi kecil dalam data, mencegah overfitting, dan membantu model umumkan pola yang lebih umum. Pada penelitian ini kami menggunakan noise gaussian dengan *mean* = 0 dan standar deviasi = 0.3 untuk fitur X, standar deviasi kecil (0.3) dipilih agar noise tidak merusak pola asli dalam data. Serta pada variabel target yaitu *noise* dengan standar deviasi = 0.5 ditambahkan ke target agar model belajar untuk menangani variabilitas nyata.



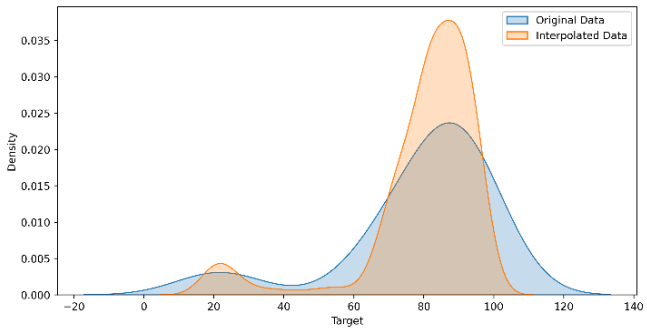
Gambar 7. Visualisasi data setelah gaussian noise augmentation

Gambar diatas merupakan hasil visualisasi distribusi data setelah melakukan virtual sample generation dengan menggunakan *gaussisan noise augmentation.* Dari visualisasi diatas dapat diketahui bahwa titik yang berwarna biru merupakan data asli sebelum augmentasi, sedangkan titik yang berwarna merah merupakan data augmentasi yang telah diberikan *gaussian noise* yang tersebar disekitar titik biru atau data asli. Dari gambar diatas menunjukkan bahwa gaussian noise augmentasi tidak merubah pola utama data, hanya memberikan beberapa variasi kecil dengan distribusi yang tetap mirip dengan distribusi awal serta berarti augmentasi ini berhasil tanpa menyebabkan perubahan drastis pada karakteristik dataset.

TABEL III.
PERFORMA MODEL SETELAH GAUSSIAN NOISE AUGMENTATION

Model	Performa	
	R2	RMSE
Random Forest	-0.25	37.72
Gradient Boosting Regressor	-0.07	34.82
Bagging Regressor	-0.17	36.36

Dari table III. Dapat dilihat bahwa virtual sample generation dengan gaussian noise augmentation hanya mendapatkan hasil performa model pada ketiga model dengan nilai R^2 -0.25, -0.07, -0.17 yang hampir sama dengan performa model yang awal dengan menggunakan data asli, berarti dapat kita simpulkan bahwa metode gaussian noise augmentation tidak cukup berpengaruh pada penelitian ini. Jika dibandingkan dengan metode interpolasi linear dan gaussian noise augmentation menunjukkan bahwa dataset ini lebih di unggulkan metode interpolasi linear yang mampu meningkatkan performa model dari ketiga model yang digunakan.



Gambar 8. Visualisasi distribusi data asli vs data interpolasi

Dari Gambar di atas menunjukkan plot distribusi kepadatan kernel (KDE) yang membandingkan distribusi data asli dengan data hasil interpolasi. Sumbu X merepresentasikan nilai target, yang dalam konteks penelitian ini dapat berupa efisiensi penghambatan korosi, sedangkan sumbu Y menunjukkan kepadatan probabilitas dari distribusi nilai target. Garis biru dengan isian transparan menggambarkan distribusi data asli, sementara garis oranye dengan isian transparan merepresentasikan distribusi data hasil interpolasi.

B. Pembahasan

Dari tabel I. Menujukkan bahwa uji performa model pada machine learning dengan algoritma ensemble yaitu Random Forest, Gradient Boosting Regressor dan Bagging Regressor mendapatkan hasil dari nilai R^2 yang cukup rendah serta nilai dari RMSE yang cukup tinggi. Hal tersebut terjadi dikarenakan data yang digunakan terlalu sedikit dan mempengaruhi nilai dari matriks evaluasi diatas. Dari awal sudah dilakukan untuk *Explanatory Data Analysis* serta sudah melakukan normalisasi dara menggunakan RobustScaler agar tidak terdapat perbedaan rentang nilai dalam fitur yang cukup jauh dan mengakibatkan hasil evaluasi yang kurang baik. Hal tersebut bisa disimpulkan karena data testing dan training hanya sedikit. Dengan validasi tambahan menggunakan *KFold Cross Validation* untuk mengetahui apakah data mengalami overfitting atau kebocoran data (data leakage) dari hasil KFold dengan nilai k=5 mendapatkan hasil performa model yang cukup baik dan dikombinasikan dengan *virtual sample generation.*

Pada Gambar 2, 3, dan 4 Menujukkan sebaran data prediksi dengan nilai sebenarnya yang dimana data testing cukup jauh dari fitting line atau nilai sebenarnya. Dari hasil visualisasi tersebut berarti model tidak cukup baik dalam memprediksi efisiensi inhibitor korosi pada dataset pyridazine. Pada Gambar 5 yang merupakan visualisasi data awal kemudian kami mengambil langkah untuk menambah data sintetik ke dalam dataset tersebut agar ukuran data tidak underfitting. Dengan Teknik Virtual Sample Generation metode *interpolasi linear* yang menambahkan sampel data sejumlah 1000, dengan visualisasi pada Gambar 6. dengan titik warna merah merupakan data sintetik yang sudah ditambahkan dengan menyambungkan antara titik biru satu dengan titik biru lainnya, serta titik biru menunjukkan bahwa tersebut adalah nilai yang sebenarnya. Hal tersebut bertujuan untuk menjaga interpebilitas dan mengubah struktur data agar ketika dilakukan uji ulang performa pada model machine learning dapat menghasilkan hasil matriks evaluasi yang cukup bagus dan menandakan bahwa Teknik *Virtual Sample Generation* mampu memberikan pengaruh dalam meningkatkan suatu hasil uji performa, dengan syarat bahwa dataset yang digunakan mengalami underfitting. Serta pada gambar 7 merupakan hasil visualisasi dari virtual sample generation dengan metode *gaussian noise augmentation* dengan Data asli dan data hasil augmentasi ditunjukkan dalam scatter plot yang digunakan Gaussian Noise. Dalam dataset, nilai fitur ditunjukkan pada sumbu horizontal (x-axis), dan nilai target ditunjukkan pada sumbu vertikal (y-axis). Scatter plot ini memiliki titik biru dan merah. Titik biru menunjukkan data asli, sedangkan titik merah menunjukkan data yang telah dimodifikasi dengan gangguan gaussian. Gaussian Noise adalah teknik augmentasi data yang menambahkan noise acak ke distribusi normal data asli, menghasilkan variasi baru yang dapat membantu meningkatkan generalisasi model dalam pembelajaran mesin.

Setelah melalui proses Virtual Sample Generation menjadi 1000 kemudian digabungkan dengan data awal. Dan

dilakukan proses evaluasi model machine learning seperti sebelumnya, pada tabel II merupakan hasil uji performa setelah melakuakan penambahan data, dan matriks evaluasi menunjukkan nilai dari R^2 dan RMSE yang cukup signifika dari sebelumnya, yaitu model Random Forest, Gradient Boosting Regressor dan Bagging Regressor mendapatkan nilai semula -0.06, 0.05 dan 0.12 menjadi 0.99, 0.96. dan 0.99, serta nilai dari RMSE yang semula 34.80, 32.90 dan 31.65 menurun menjadi 1.59, 2.88 dan 1.25 yang menunjukkan bahwa Teknik Virtual Sample Generation cukup memberikan pengaruh pada kasus prediksi dalam meningkatkan efisiensi inhibitor korosi pada dataset pyridazine. Sedaangkan pada tabel III menujukkan hasil uji performa setelah melakukan gaussian noise augmentation dan menunjukkan nilai R^2 dan RMSE yang hampir sama dengan nilai awal yang menggunakan data asli sebelum di berikan visual sample generation dengan 2 metode tersebut. Dapat diambil Keputusan bahwa metode *gaussian noise augmentation* tidak mempengaruhi kinerja performa dari ke 3 model yang digunakan dibandingkan menggunakan metode interpolasi linear.

Pada Gambar 8. tersebut, terlihat bahwa kurva interpolasi bertumpuk dengan kurva data asli, menunjukkan bahwa teknik tersebut tidak mengubah distribusi secara signifikan. Namun, distribusi interpolasi memiliki puncak yang lebih tajam, terutama di antara nilai target 80 dan 100, yang menunjukkan bahwa interpolasi mungkin telah menambahkan lebih banyak sampel di sekitarnya. Sebagian besar area di bawah kurva interpolasi bertumpuk dengan kurva data asli, menunjukkan bahwa data yang dihasilkan dari interpolasi tidak jauh berbeda dari distribusi awal. Ini berarti interpolasi tidak menyebabkan perubahan besar pada distribusi data target. Kemudian pada tahap akhir melakukan uji statistik dengan menggunakan *Kolmogorov Smirnov* (KS test) untuk membuktikan bahwa tidak ada perbedaan distribusi yang signifikan. Dari pengujian tersebut menggunakan toleransi nilai signifikan sebesar 0.05, apabila nilai *p_value > 0.05* maka distribusi data asli dan augmentasi tidak berbeda signifikan dan apabila nilai *p_value < 0.05* maka distribusi data asli dan augmentasi terdapat perbedaan nilai yang signifikan atau berbeda dari data asli. Dari pengujian tersebut menunjukkan nilai dari *p_value* sebesar 0.85 yang menunjukkan bahwa pada visrtual sample generation dengan metode interpolasi linear tidak mengubah distribusi augmentasi tidak berbeda dengan signifikan dari data asli.

Jika hubungan antar fitur linear atau hampir linear, interpolasi linear adalah teknik sederhana yang efektif. Namun, jika hubungan antar variabel lebih kompleks, interpolasi linear dapat menghasilkan data yang tidak mencerminkan pola sebenarnya. Oleh karena itu, sebelum menggunakan interpolasi linear, perlu dilakukan analisis awal terhadap hubungan antar fitur. Untuk menghasilkan data tambahan yang lebih representatif, metode interpolasi yang lebih canggih seperti spline interpolasi atau peningkatan berbasis ML dapat dipertimbangkan jika hubungan yang ditemukan bersifat non-linear.

Interpolasi linear tidak efektif jika ada variabel eksternal yang tidak terlihat tetapi berdampak besar pada hubungan antar fitur. Metode yang lebih maju, seperti model regresi berbasis faktor laten, teknik pembelajaran mesin yang dapat mendeteksi pola kompleks, atau teknik augmentasi yang mempertimbangkan distribusi data secara lebih menyeluruh, diperlukan untuk mengatasi masalah ini. Mengidentifikasi dan mengukur variabel eksternal juga dapat meningkatkan kualitas augmentasi data dan hasil analisis jika memungkinkan. Jika material lain memiliki hubungan yang mirip dengan pyridazine, seperti efek linear antara konsentrasi dan efisiensi penghambatan korosi, maka metode ini bisa diterapkan dengan baik.

IV. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa *virtual sample generation* dengan metode *interpolasi linear* dapat memberikan pengaruh dibandingkan metode *gaussian noise augmentation* terhadap dataset pyridazine dalam memprediksi peningkatan efiesiensi penghambatan korosi yang mengalami *underfitting*. Dengan melakukan beberapa tahapan seperti KFold Cross validation serta uji statistik *Kolmogorov Smirnov* untuk memastikan bahwa penelitian tersebut tidak mengubah distribusi data asli .Dengan menggunakan model algoritma ensemble yaitu Random Forest, Gradient Boosting Regressor dan Bagging Regressor yang menghasilkan uji performa dengan nilai R^2 dan RMSE sebesar 0.99, 0.96. dan 0.99 dan 1.59, 2.88 dan 1.25.

DAFTAR PUSTAKA

[1] E. H. El Assiri *et al.*, "Development and validation of QSPR models for corrosion inhibition of carbon steel by some pyridazine derivatives in acidic medium," *Heliyon*, vol. 6, no. 10, Oct. 2020, doi: 10.1016/j.heliyon.2020.e05067.

[2] K. Rasheeda, V. D. P. Alva, P. A. Krishnaprasad, and S. Samshuddin, "Pyrimidine derivatives as potential corrosion inhibitors for steel in acid medium – An overview," *International Journal of Corrosion and Scale Inhibition*, vol. 7, no. 1, pp. 48–61, 2018, doi: 10.17675/2305-6894-2018-7-1-5.

[3] O. S. I. Fayomi, I. G. Akande, and S. Odigie, "Economic Impact of Corrosion in Oil Sectors and Prevention: An Overview," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Dec. 2019. doi: 10.1088/1742-6596/1378/2/022037.

[4] R. S. Iyer, N. S. Iyer, R. A. P, and A. Joseph, "Harnessing machine learning and virtual sample generation for corrosion studies of 2-alkyl benzimidazole scaffold small dataset with an experimental validation," *J Mol Struct*, vol. 1306, p. 137767, 2024, doi: https://doi.org/10.1016/j.molstruc.2024.137767.

[5] M. Akrom, "Investigation Of Natural Extracts As Green Corrosion Inhibitors In Steel Using Density Functional Theory," 2022.

[6] T. W. Quadri *et al.*, "Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors," *Mater Today Commun*, vol. 30, p. 103163, 2022, doi: https://doi.org/10.1016/j.mtcomm.2022.103163.

[7] 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, no. 1, pp. 22–37, Apr. 2024, doi: 10.62411/jimat.v1i1.10495.

[8] M. Akrom, S. Rustad, and H. K. Dipojono, "Variational quantum circuit-based quantum machine learning approach for predicting corrosion inhibition efficiency of pyridine-quinoline compounds," *Materials Today Quantum*, vol. 2, p. 100007, Jun. 2024, doi: 10.1016/j.mtquan.2024.100007.

[9] T. W. Quadri *et al.*, "Predicting protection capacities of pyrimidine-based corrosion inhibitors for mild steel/HCl interface using linear and nonlinear QSPR models," *J Mol Model*, vol. 28, no. 9, p. 254, 2022, doi: 10.1007/s00894-022-05245-1.

[10] D.-C. Li and I.-H. Wen, "A genetic algorithm-based virtual sample generation technique to improve small data set learning," *Neurocomputing*, vol. 143, pp. 222–230, 2014, doi: https://doi.org/10.1016/j.neucom.2014.06.004.

[11] S. Wu, B. Wang, J. Zhao, M. Zhao, K. Zhong, and Y. Guo, "Virtual sample generation and ensemble learning based image source identification with small training samples," *International Journal of Digital Crime and Forensics*, vol. 13, no. 3, pp. 34–46, May 2021, doi: 10.4018/IJDCF.20210501.oa3.

[12] M. R. Rosyid, L. Mawaddah, A. P. Santosa, M. Akrom, S. Rustad, and H. K. Dipojono, "Implementation of quantum machine learning in predicting corrosion inhibition efficiency of expired drugs," *Mater Today Commun*, vol. 40, p. 109830, 2024, doi: https://doi.org/10.1016/j.mtcomm.2024.109830.

[13] 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, 2023, doi: 10.1038/s41529-023-00336-7.

[14] 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)*, 2020, pp. 1123–1128. doi: 10.1109/CAC51589.2020.9326486.

[15] 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*, Mar. 2024, doi: 10.18502/keg.v6i1.15350.

[16] S. Budi, M. Akrom, G. A. Trisnapradika, T. Sutojo, and 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*, vol. 10, no. 2, pp. 151–158, May 2023, doi: 10.15294/sji.v10i2.43929.

[17] M. Akrom, S. Rustad, and H. K. Dipojono, "SMILES-based machine learning enables the prediction of corrosion inhibition capacity," *MRS Commun*, vol. 14, no. 3, pp. 379–387, 2024, doi: 10.1557/s43579-024-00551-6.

[18] 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 Chem*, vol. 6, Dec. 2023, doi: 10.1016/j.rechem.2023.101126.

[19] C. Cui, J. Tang, H. Xia, J. Qiao, and W. Yu, "Virtual sample generation method based on generative adversarial fuzzy neural network," *Neural Comput Appl*, vol. 35, no. 9, pp. 6979–7001, 2023, doi: 10.1007/s00521-022-08104-5.

[20] M. Akrom *et al.*, "DFT and microkinetic investigation of oxygen reduction reaction on corrosion inhibition mechanism of iron surface by Syzygium Aromaticum extract," *Appl Surf Sci*, vol. 615, p. 156319, 2023, doi: https://doi.org/10.1016/j.apsusc.2022.156319.

[21] M. Akrom, S. Rustad, H. K. Dipojono, and R. Maezono, "A comprehensive approach utilizing quantum machine learning in the study of corrosion inhibition on quinoxaline compounds," *Artificial Intelligence Chemistry*, vol. 2, no. 2, p. 100073, 2024, doi: https://doi.org/10.1016/j.aichem.2024.100073.

[22] M. Akrom, S. Rustad, and H. K. Dipojono, "Development of quantum machine learning to evaluate the corrosion inhibition capability of pyrimidine compounds," *Mater Today Commun*, vol. 39, p. 108758, 2024, doi: https://doi.org/10.1016/j.mtcomm.2024.108758.

[23] M. Akrom, S. Rustad, and H. K. Dipojono, "Investigation of Corrosion Inhibition Capability of Pyridazine Compounds via Ensemble Learning," *J Mater Eng Perform*, 2024, doi: 10.1007/s11665-024-10129-x.

[24] J. Yang, X. Yu, Z.-Q. Xie, and J.-P. Zhang, "A novel virtual sample generation method based on Gaussian distribution," *Knowl Based Syst*, vol. 24, no. 6, pp. 740–748, 2011, doi: https://doi.org/10.1016/j.knosys.2010.12.010.

[25] 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, doi: https://doi.org/10.1016/j.comptc.2023.114307.

[26] T. W. Quadri *et al.*, "Computational insights into quinoxaline-based corrosion inhibitors of steel in HCl: Quantum chemical analysis and QSPR-ANN studies," *Arabian Journal of Chemistry*, vol. 15, no. 7, p. 103870, 2022, doi: https://doi.org/10.1016/j.arabjc.2022.103870.