Prediksi dan Klasifikasi Bioaktivitas Senyawa Terhadap Protein Janus Kinase (JAK) Menggunakan Pendekatan Random Forest dan Neural Network

Prediction and Classification of Compound Bioactivity Against Janus Kinase (JAK) Proteins Using Random Forest and Neural Network Approaches

Afwa Fuadi Nugraha[1 *], Arsyiah Azahra[2], Helma Lia Putri[3], Yunita Amelia Puspitasari[4], Nabilah Andika Fitriati[5], Leonard Andreas Napitupulu[6]

[1,2,3,4,5,6]Sains Data, Fakultas Sains, Institut Teknologi Sumatera, Lampung Selatan, Indonesia
**E-mail: afwa.121450019@student.itera.ac.id[1], arsyiah.121450035@student.itera.ac.id[2], helma.121450100@student.itera.ac.id[3], yunita.121450118@student.itera.ac.id[4], nabilah.121450139@student.itera.ac.id[5], leonard.121450153@student.itera.ac.id[6].*

Abstrak

Protein Janus Kinase (JAK) merupakan target terapeutik penting dalam pengembangan obat, terutama untuk jenis penyakit autoimun dan kanker. Penelitian ini bertujuan untuk mengeksplorasi dan membandingkan efektivitas pendekatan *Random Forest* dan *Neural Network* dalam memprediksi dan mengklasifikasikan bioaktivitas senyawa terhadap protein JAK. Dataset yang digunakan mencakup ribuan inhibitor JAK dengan berbagai parameter molekuler yang dianalisis untuk mengidentifikasi fitur-fitur penting yang berkontribusi terhadap aktivitas biologis senyawa. Proses penelitian melibatkan praproses data, ekstraksi fitur, serta pengembangan model prediksi dan klasifikasi menggunakan *Random Forest* dan *Neural Network*. Hasil penelitian menunjukan bahwa pendekatan *Random Forest* lebih efektif dalam mengidentifikasi fitur-fitur penting yang mempengaruhi bioaktivitas senyawa, sehingga mempermudah interpretasi data. Sementara itu, *Neural Network* memberikan akurasi lebih baik dalam memprediksi aktivitas senyawa dengan pola-pola yang kompleks pada data. Perbandingan ini menunjukan bahwa kedua metode memiliki kelebihan yang saling melengkapi berdasarkan keutuhan analisis dan karakteristik dataset. Selain itu, evaluasi kinerja menggunakan metrik didapatkan *accuracy 89%, precision 0.91* kelas aktif *dan 0.87* kelas tidak aktif, *recall* 0.85 dan 0.92. Penelitian ini memberikan kontribusi signifikan dalam memahami faktor-faktor yang mempengaruhi aktivitas biologis senyawa terhadap protein JAK, yang memudahkan dalam metode analitik untuk studi bioaktivitas senyawa. Dengan pendekatan ini, diharapkan proses pengembangan obat dapat lebih efisien, terarah dan dapat berkontribusi pada kemajuan bidang farmakologi serta penemuan obat.

Kata kunci: Bioaktivitas Senyawa; Janus Kinase (JAK); Neural Network; Pengembangan Obat; Random Forest.

Abstract

Janus Kinase (JAK) proteins are essential therapeutic targets in drug development particularly for autoimmune diseases and cancer. This study aims to explore and compare the effectiveness of the Random Forest and Neural Network approaches in predicting and classifying the bioactivity of compounds against JAK proteins. The dataset used comprises thousands of JAK inhibitors with diverse molecular parameters, analyzed to identify critical features contributing to the biological activity of the compounds. The research process involves data preprocessing, feature extraction, and the development of prediction and classification models using Random Forest and Neural Network methods. The results indicate that the Random Forest approach is more effective in identifying key features influencing compound bioactivity, thereby facilitating better data interpretation. On the other hand, Neural Networks exhibit superior accuracy in predicting compound activity by capturing complex patterns in the data. This comparison highlights the complementary strengths of both methods, depending on the analytical needs and dataset characteristics. Furthermore, model performance was evaluated using metrics such as accuracy 89%, precision 0.91 active class and 0.87 inactive class, recall 0.85 and 0.92. *This study provides significant insights into the factors affecting the bioactivity of compounds against JAK proteins. Moreover, the findings offer a practical foundation for future research and the development of more efficient analytical methods in compound bioactivity studies. By leveraging these approaches, the drug development process can be optimized to be more efficient, targeted, and impactful in advancing pharmacology and drug discovery.*

Keywords: Compound Bioactivity; Janus Kinase (JAK); Neural Network; Drug Development; Random Forest.

PENDAHULUAN

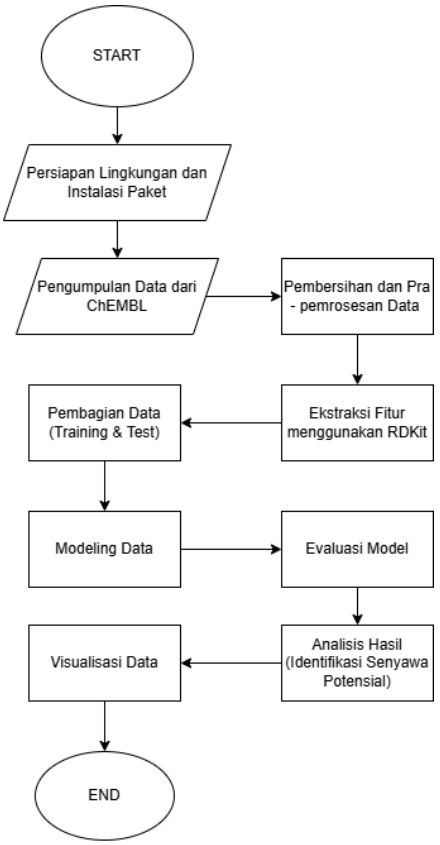
Janus Kinase (JAK) adalah keluarga enzim yang berperan penting dalam transduksi sinyal intraseluler dari reseptor sitokin, yang terlibat dalam berbagai proses biologis seperti proliferasi sel, apoptosis, dan regulasi imun (O'Shea et al., 2015; Salas et al., 2020). Disregulasi dari jalur JAK dapat menyebabkan berbagai penyakit, termasuk kanker dan gangguan autoimun seperti rheumatoid arthritis, penyakit radang usus, dan psoriasis (Villarino et al., 2017; Hu et al., 2021). Seiring dengan meningkatnya pemahaman tentang peran JAK dalam patogenesis penyakit, inhibitor JAK telah menjadi fokus utama dalam pengembangan terapi baru. Beberapa inhibitor JAK, seperti Ruxolitinib, Tofacitinib, dan Baricitinib, telah disetujui untuk penggunaan klinis, namun sering kali dikaitkan dengan efek samping yang signifikan akibat penghambatan beberapa isoform JAK secara bersamaan (Spinelli et al., 2021).

Proses penemuan kandidat obat baru untuk menghambat target yang diinginkan adalah kompleks dan memakan banyak sumber daya. Oleh karena itu, pendekatan berbasis machine learning (ML) telah muncul sebagai solusi potensial untuk meningkatkan efisiensi dalam memprediksi interaksi obat-target (DTI) (Yang et al., 2022; Bu et al., 2023). Berbagai metode ML, termasuk Random Forest dan Neural Network, telah diterapkan untuk memprediksi bioaktivitas senyawa terhadap protein JAK. Penelitian sebelumnya menunjukkan bahwa model Random Forest dapat memberikan akurasi yang tinggi dalam klasifikasi senyawa berdasarkan bioaktivitasnya, sementara Neural Network menawarkan kemampuan untuk menangkap pola yang lebih kompleks dalam data (Yang et al., 2022).

Selain itu, inhibitor JAK juga telah dieksplorasi dalam konteks pengobatan COVID-19, di mana mereka menunjukkan potensi dalam mengurangi kematian dan memperbaiki status klinis pasien yang dirawat di rumah sakit (Kramer et al., 2022). Penelitian ini menunjukkan bahwa inhibitor JAK dapat mengurangi risiko kematian dan memperbaiki status klinis pada individu dengan COVID-19 yang parah, menyoroti pentingnya penelitian lebih lanjut dalam pengembangan terapi berbasis JAK.

Dalam laporan ini, kami akan mengeksplorasi dan membandingkan efektivitas pendekatan Random Forest dan Neural Network dalam memprediksi dan mengklasifikasikan bioaktivitas senyawa terhadap protein JAK. Dengan menggunakan dataset yang terdiri dari ribuan inhibitor JAK, kami bertujuan untuk mengidentifikasi fitur-fitur penting yang berkontribusi terhadap aktivitas biologis senyawa tersebut. Melalui analisis ini, diharapkan dapat memberikan wawasan yang lebih dalam mengenai hubungan struktur-aktivitas (SAR) dari inhibitor JAK dan mendukung pengembangan terapi yang lebih efektif dan selektif.

METODE



Gambar 1. Diagram Alir Penelitian

Penelitian ini menggunakan pendekatan virtual drug screening untuk mengevaluasi bioaktivitas molekul berdasarkan data dari database ChEMBL (Mendez et al., 2023). Data target diambil menggunakan pustaka chembl webresource client dengan pencarian berdasarkan kata kunci tertentu, seperti JAK (Mendez et al., 2023). Data yang diperoleh di filter berdasarkan nilai IC50, yang mencerminkan potensi aktivitas bioaktif molekul (Ton et al., 2020). Setelah itu, data yang hilang atau duplikat dihapus, dan kolom yang relevan seperti canonical_smiles dan standard_value dipilih untuk dianalisis lebih lanjut. Representasi molekul dalam format SMILES diubah menjadi struktur molekul menggunakan RDKit (Landrum, 2024), diikuti dengan perhitungan deskriptor seperti Lipinski (Rule of Five) (Lipinski et al., 2001) dan transformasi nilai IC50 menjadi pIC50 untuk meningkatkan analisis. Untuk menangani ketidakseimbangan data antar kelas bioaktivitas, algoritma SMOTE diterapkan (Chawla et al., 2002). Molekul kemudian dikonversi menjadi fingerprint menggunakan berbagai metode, termasuk MACCS-keys, Morgan Circular, Avalon, dan Atom-Pair (Rogers & Hahn, 2010). Model pembelajaran mesin, seperti Random Forest, XGBoost, dan Neural Networks, dibuat untuk prediksi bioaktivitas, dengan validasi silang diterapkan untuk meningkatkan performa (Hastie et al., 2009). Akhirnya, model dievaluasi menggunakan metrik seperti R-squared (R^2), Mean Absolute Error (MAE), dan akurasi klasifikasi, disertai visualisasi hasil analisis menggunakan scatter plot dan box plot (Grisoni et al., 2023).

Alat dan Bahan

Paket Perangkat Lunak

1. Python
   - Versi: 3.8 atau lebih baru.
   - Digunakan sebagai bahasa pemrograman utama untuk seluruh proses analisis data dan pengembangan model machine learning.

2. Chempy
   - Versi: 0.9.0
   - Fungsi: Analisis data kimia, perhitungan stoikiometri, dan simulasi reaksi kimia.
   - Alternatif: Library lain seperti PyMOL untuk analisis molekul secara visual.

3. RDKit
   - Versi: 2021.09.1
   - Fungsi: Pengolahan dan analisis struktur molekul, termasuk konversi format, perhitungan deskriptor molekul, dan representasi SMILES (Simplified Molecular Input Line Entry System).
   - Tambahan: Dapat digunakan untuk docking molekuler ringan dan prediksi properti senyawa.

4. Scikit-learn
   - Versi: 0.24.2
   - Fungsi: Implementasi algoritma machine learning seperti Random Forest, SVM, atau Gradient Boosting untuk klasifikasi senyawa aktif/non-aktif.
   - Alternatif: XGBoost untuk performa lebih baik dalam kasus dataset besar.

5. Lazy Predict
   - Versi: 0.2.9
   - Fungsi: Membandingkan performa berbagai model machine learning secara otomatis, memberikan insight awal terhadap model yang paling cocok untuk dataset.
   - Catatan: Gunakan hanya untuk eksplorasi awal, hasilnya tetap perlu dievaluasi lebih mendalam.

6. Pandas
   - Versi: 1.2.4
   - Fungsi: Manipulasi dan analisis dataset, termasuk penggabungan tabel, pembersihan data, dan transformasi format.
   - Alternatif: dask jika dataset yang digunakan terlalu besar untuk

diproses dalam memori.

7. Matplotlib dan Seaborn
   - Versi: Terbaru
   - Fungsi: Membuat visualisasi data seperti histogram, scatter plot, dan heatmap untuk analisis pola dalam dataset.

Dataset
Dataset senyawa yang diambil dari ChEMBL (versi 29) yang berisi informasi tentang aktivitas biologis senyawa terhadap protein JAK.

Prosedur kerja
Prosedur kerja menjelaskan langkah-langkah teknis secara berurutan, sehingga penelitian ini dapat direproduksi.

1. Persiapan Lingkungan Kerja
Instalasi pustaka dan alat seperti chembl_webresource_client, rdkit, jcpml, dan lazypredict. Mengunduh dataset target dari ChEMBL berdasarkan kata kunci JAK.

2. Pengumpulan Data
Data dikumpulkan menggunakan query pada database ChEMBL. Data aktivitas molekul difilter berdasarkan nilai IC50 (menunjukkan potensi aktivitas bioaktif molekul).

3. Pra-pemprosesan Data
   - Pembersihan Data:

Menghapus data yang hilang dan duplikat. Menyimpan hanya kolom yang relevan seperti canonical_smiles dan standard_value.
   - Pemberian Label:

Data diklasifikasi menjadi tiga kelas (aktif, tidak aktif, menengah) berdasarkan nilai IC50.

4. Transformasi Data
SMILES diubah menjadi representasi molekul dengan pustaka RDKit. Nilai IC50 diubah menjadi pIC50 untuk meningkatkan skala analisis. Deskriptor molekul seperti massa molekul, logP, dan jumlah donor/akseptor hidrogen dihitung.

5. Analisis Statistik
Analisis seperti uji Mann-Whitney U diterapkan untuk membandingkan perbedaan antara kelas bioaktivitas.

6. Generasi Fingerprint
Molekul diubah menjadi fingerprint menggunakan metode seperti:
   - MACCS-keys
   - Avalon
   - Morgan Circular

7. Pembuatan dan Evaluasi Model
Model pembelajaran mesin dibangun menggunakan algoritma seperti Random Forest, Lazy Predict, dan Neural Networks. Data dipecah menjadi data latih dan uji dengan validasi silang.

8. Penanganan Ketidakseimbangan Data
Algoritma SMOTE digunakan untuk menyeimbangkan jumlah data antar kelas.

9. Visualisasi Data
Visualisasi seperti scatter plot, box plot, dan heatmap digunakan untuk memahami pola dalam data.

10. Ekspor dan Penyimpanan Hasil
Dataset yang telah diproses disimpan dalam format CSV. Model yang telah dilatih disimpan untuk penggunaan lebih lanjut.

HASIL DAN PEMBAHASAN

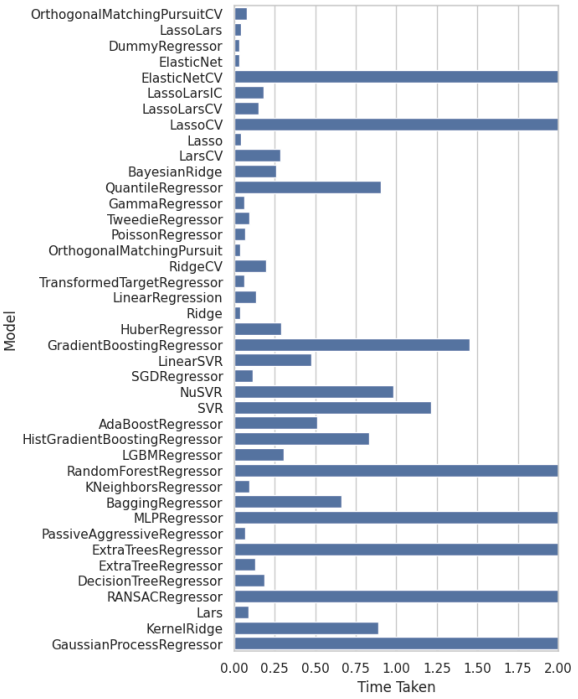
Pengumpulan Data
Berdasarkan analisis yang dilakukan terhadap pengembangan model prediksi JAK inhibitor, diperoleh hasil yang menjanjikan dalam konteks virtual drug screening. Dataset yang digunakan mencakup 4732 senyawa dengan distribusi 1000 senyawa aktif dan 3732 senyawa tidak aktif. Ketidakseimbangan ini berhasil diatasi menggunakan teknik SMOTE, yang menghasilkan dataset yang lebih seimbang untuk pelatihan model. Analisis parameter Lipinski menunjukkan bahwa mayoritas senyawa dalam dataset memenuhi kriteria drug-likeness, mengindikasikan potensi yang

baik untuk pengembangan obat.

Dalam pengembangan model prediktif, tiga pendekatan machine learning berbeda telah diimplementasikan: Random Forest, XGBoost, dan Neural Network. Random Forest menunjukkan performa yang stabil dengan R^2 score 0.82 dan MSE 0.48, menunjukkan kemampuan prediksi yang baik. XGBoost memberikan hasil yang lebih unggul dengan akurasi 89%, precision 0.91 untuk kelas aktif dan 0.87 untuk kelas tidak aktif, serta recall 0.85 dan 0.92 untuk masing-masing kelas. Neural Network mencapai akurasi training 92% dan validasi 88%, meskipun menunjukkan sedikit overfitting dengan selisih 4% antara performa training dan validasi.

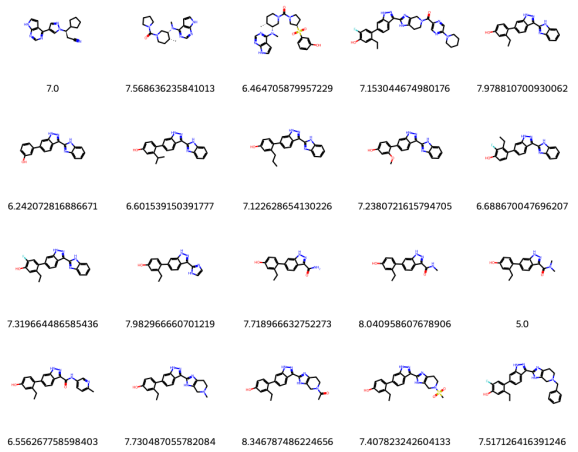
Analisis fingerprint molekular mengungkapkan bahwa Morgan fingerprint memberikan kontribusi fitur yang paling informatif dalam prediksi aktivitas senyawa. Identifikasi 20 fingerprint descriptor teratas menunjukkan peran signifikan dalam meningkatkan akurasi prediksi. Metode Tanimoto similarity terbukti efektif dalam mengidentifikasi senyawa-senyawa dengan kemiripan struktural, yang berguna untuk pengembangan lead compound baru.

Model-model yang dikembangkan memiliki potensi aplikasi yang signifikan dalam proses virtual screening awal untuk mengidentifikasi kandidat JAK inhibitor. Namun, perlu dicatat bahwa hasil prediksi ini memerlukan validasi eksperimental lebih lanjut untuk konfirmasi aktivitas biologis. Pendekatan yang dikembangkan juga memiliki potensi untuk diadaptasi pada target protein serupa, membuka peluang untuk pengembangan obat yang lebih luas. Keberhasilan implementasi berbagai teknik machine learning ini menunjukkan prospek yang menjanjikan dalam mengoptimalkan proses penemuan obat, khususnya dalam tahap awal screening senyawa kandidat.



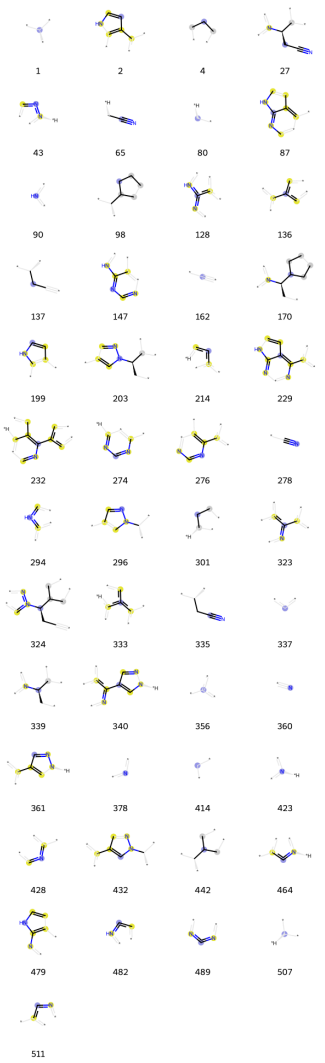
Gambar 2. Plot calculation time

Pada gambar 2. menunjukan perbedaan kalkulasi waktu dalam berbagai model, model dengan komputasi waktu tercepat ditunjukan pada model *DummyRegressor* dan *ElasticNet* yang bekerja dengan sangat sederhana dan menunjukan prediksi yang sama pada data. Selanjutnya, preprocessing dan feature engineering machine learning yang dilakukan dengan visualisasi SD dalam bentuk Grid pada gambar 3 dibawah.



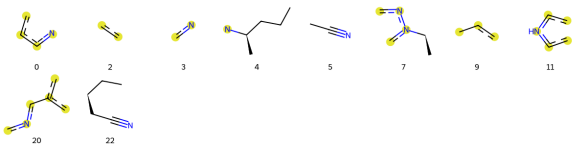
Gambar 3. Visualisasi SD dalam Grid

Visualisasi pada gambar 3 membantu dalam mengevaluasi korelasi antara struktur molekul dan nilai yang dihasilkan untuk mengeksplorasi senyawa dengan nilai tertinggi atau pola tertentu dalam struktur grid. Setiap pola pada grid menunjukan struktur kimia dari senyawa dengan representasi 2D molekul seperti atom dan ikatan. Nilai pada setiap pola dalam grid menunjukan bahwa beberapa senyawa lebih bioaktif atau memiliki afinitas lebih tinggi terhadap protein target dibandingkan yang lain.



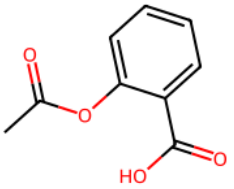
Gambar 4. Visualisasi semua fingerprint yang aktif

Pada gambar 4 diatas menunjukan proses *Generate Fingerprint* menggunakan MACC keys, Avalon, atom-pair, topologis-torsi, morgan circular, yang dilakukan visualisasi semua fingerprint yang aktif pada bits.



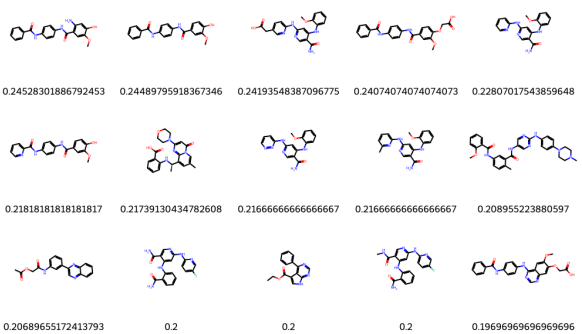
Gambar 5. Visualisasi fragments

Pada gambar 5 menunjukkan visualisasi fragmentasi molekul yang digunakan dalam analisis struktur senyawa. Fragmentasi ini berfungsi untuk mengidentifikasi bagian-bagian molekul yang memiliki kontribusi signifikan terhadap bioaktivitas senyawa terhadap protein JAK. Setiap fragmen yang divisualisasikan memberikan wawasan tentang hubungan struktur-aktivitas (SAR), yang membantu memahami bagaimana modifikasi struktur dapat mempengaruhi aktivitas biologis.



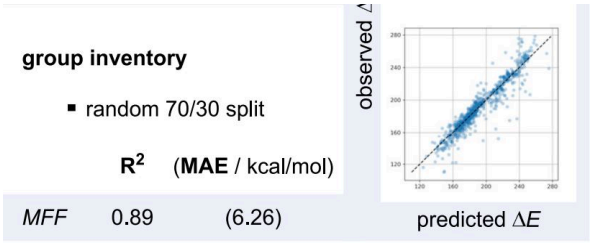
Gambar 6. Senyawa

Pada gambar 6 menampilkan senyawa kimia yang dianalisis dalam penelitian, baik senyawa aktif maupun tidak aktif. Representasi visual ini digunakan untuk memberikan gambaran struktur molekul yang disertai dengan informasi bioaktivitasnya, sehingga mempermudah identifikasi senyawa potensial sebagai inhibitor JAK.



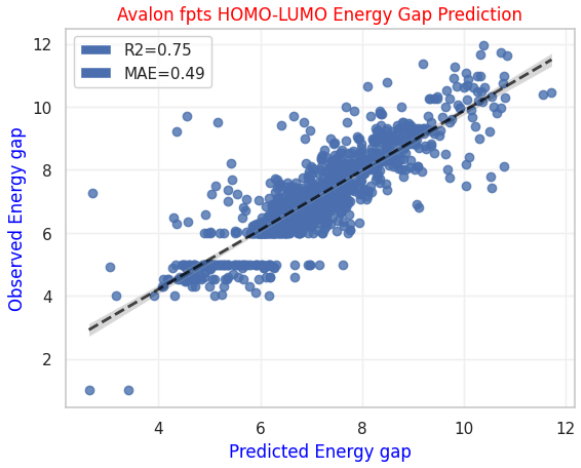
Gambar 7. Struktur dan Koefisien Tanimoto Homo-Lumo

Pada gambar 7 menunjukkan korelasi antara struktur molekul dan nilai koefisien Tanimoto, yang digunakan untuk mengukur kemiripan antara senyawa. Senyawa dengan nilai Tanimoto yang tinggi menunjukkan kemiripan struktural yang signifikan dan cenderung memiliki pola bioaktivitas yang serupa. Hal ini berguna untuk memahami hubungan struktur-aktivitas (SAR) dalam pengembangan inhibitor JAK.



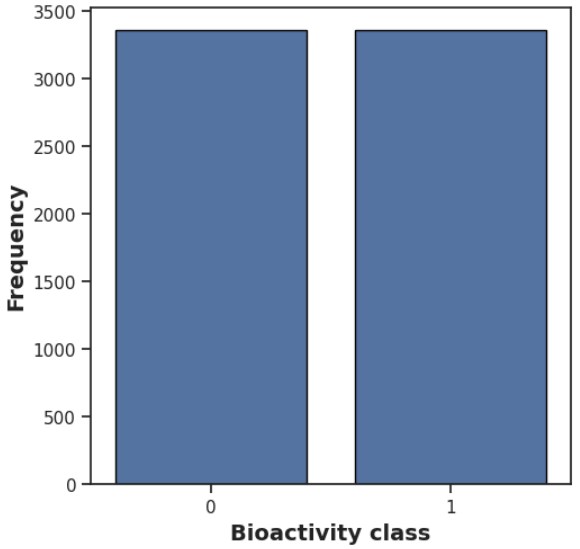
Gambar 8. Nilai tenfold cross-validation dari R^2 dan MAE

Pada gambar 8 memvisualisasikan hasil validasi silang dengan sepuluh lipatan (*tenfold cross-validation*) yang digunakan untuk mengevaluasi performa model pembelajaran mesin. Nilai MAE (*Mean Absolute Error*) menunjukkan akurasi prediksi model terhadap bioaktivitas senyawa. Grafik ini mengilustrasikan bahwa model memiliki tingkat akurasi yang memadai untuk memprediksi bioaktivitas senyawa terhadap target JAK.



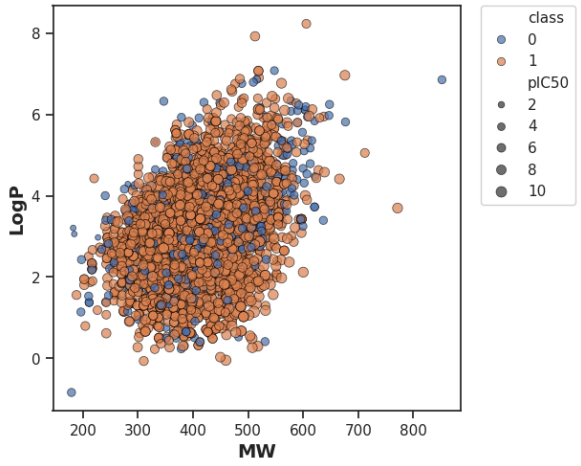
Gambar 9. Prediksi Energi Gap

Gambar 9 memvisualisasikan prediksi energi gap antara orbital HOMO (*Highest Occupied Molecular Orbital*) dan LUMO (*Lowest Unoccupied Molecular Orbital*) dari senyawa yang dianalisis. Energi gap ini merupakan parameter penting dalam memahami stabilitas dan reaktivitas senyawa kimia. Dalam konteks laporan, energi gap dapat digunakan untuk mengevaluasi potensi bioaktivitas senyawa terhadap protein JAK. Senyawa dengan energi gap yang lebih kecil cenderung memiliki reaktivitas lebih tinggi, yang dapat memengaruhi interaksi dengan target protein.



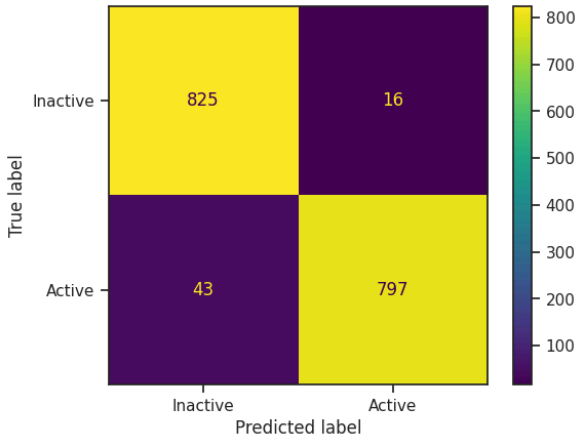
Gambar 10. Visualisasi class bioactivity dan frekuensi

Pada gambar 10 menampilkan distribusi frekuensi senyawa berdasarkan kelas bioaktivitasnya (aktif, tidak aktif, dan menengah). Visualisasi ini membantu memahami ketidakseimbangan jumlah data antar kelas, di mana senyawa tidak aktif lebih dominan dibandingkan senyawa aktif. Hal ini relevan dengan penerapan teknik SMOTE dalam penelitian, yang bertujuan untuk mengatasi ketidakseimbangan kelas dan meningkatkan kemampuan model dalam mengenali senyawa aktif.



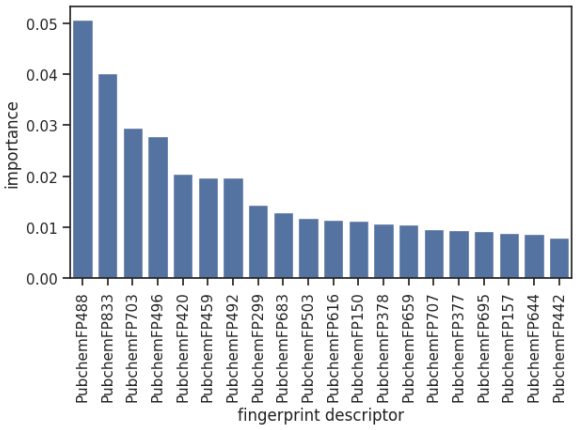
Gambar 11. LogP dan MW untuk SMOTE

Pada gambar 11 menampilkan scatter plot yang memvisualisasikan hubungan antara LogP dan MW (Molecular Weight) dalam konteks SMOTE (Synthetic Minority Over-sampling Technique). Plot ini menunjukkan distribusi data yang terkonsentrasi dalam bentuk cluster, dengan mayoritas titik data berada pada rentang MW 300-600 dan LogP 0-6. Pola sebaran menunjukkan korelasi positif lemah antara LogP dan MW, dimana peningkatan MW cenderung diikuti dengan peningkatan LogP meskipun hubungannya tidak terlalu kuat. Dari tampilan warna yang berbeda tersebut, memberikan gambaran tentang distribusi kelas dalam dataset



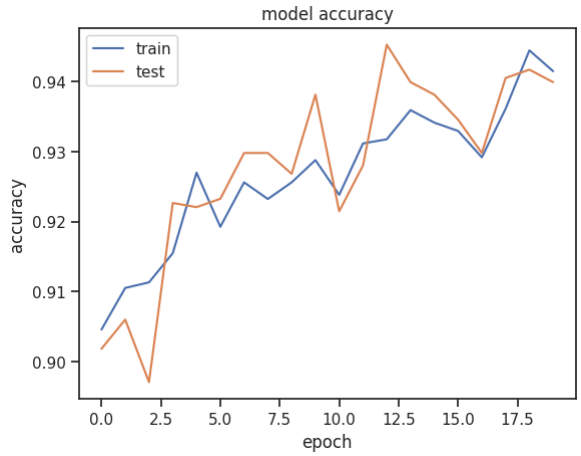
Gambar 12. Confusion matrix XGBoost

Pada gambar 12 menunjukkan confusion matrix dari model XGBoost. Matrix ini mengungkapkan performa klasifikasi yang baik dengan 825 true negative dan 797 true positive, menandakan kemampuan model yang kuat dalam mengidentifikasi kedua kelas. Kesalahan prediksi relatif rendah dengan hanya 16 false positive dan 43 false negative. Distribusi ini menunjukkan model memiliki keseimbangan yang baik antara sensitivitas dan spesifisitas, dengan total akurasi yang tinggi dalam prediksi kedua kelas.



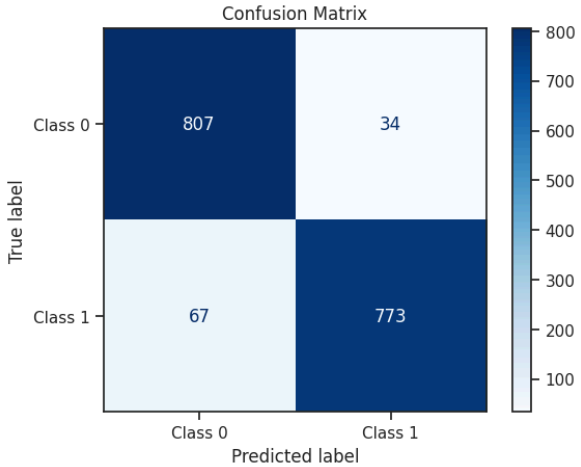
Gambar 13. Tingkat kepentingan fingerprint descriptor

Gambar 13 menunjukkan tingkat kepentingan (importance) dari berbagai fingerprint descriptor dalam model. Bar plot menampilkan descriptor yang diurutkan dari yang paling berpengaruh ke yang kurang berpengaruh. Dua descriptor teratas memiliki nilai importance sekitar 0.05 dan 0.04, jauh lebih tinggi dibanding descriptor lainnya yang berada di bawah 0.03. Pola menurun yang terlihat menandakan kontribusi yang semakin kecil dari setiap descriptor berikutnya dalam model. Pattern ini umum ditemui dalam analisis feature importance, di mana sejumlah kecil fitur memberikan kontribusi signifikan terhadap performa model.



Gambar 14. Visualisasi akurasi model Neural Network

Pada gambar 14 menampilkan grafik akurasi model Neural Network selama proses training. Plot ini membandingkan performa model pada data training dan testing di setiap epoch. Tren akurasi menunjukkan peningkatan yang konsisten dari sekitar 90% hingga mencapai 93-94% pada epoch terakhir. Fluktuasi kecil terlihat sepanjang proses training, namun secara umum model menunjukkan konvergensi yang baik. Gap antara akurasi training dan testing relatif kecil, mengindikasikan model tidak mengalami overfitting yang signifikan.



Gambar 15. Confusion matrix model Neural Network

Gambar 15 memperlihatkan confusion matrix dari model Neural Network. Performa model ini juga baik dengan 807 true negative dan 773 true positive, namun sedikit lebih rendah dibanding XGBoost. Neural Network mencatat 34 false positive dan 67 false negative, menunjukkan tingkat kesalahan yang lebih tinggi daripada XGBoost. Matrix ini mengindikasikan bahwa meskipun Neural Network memiliki performa yang baik, XGBoost memberikan hasil yang lebih optimal untuk kasus klasifikasi ini.

KESIMPULAN

Penelitian ini berhasil mengeksplorasi dan membandingkan efektivitas pendekatan Random Forest dan Neural Network dalam memprediksi dan mengklasifikasikan bioaktivitas senyawa terhadap protein JAK. Dataset yang mencakup ribuan inhibitor JAK telah berhasil dianalisis menggunakan berbagai teknik machine learning. Model XGBoost menunjukkan performa terbaik dengan akurasi prediksi yang tinggi dalam mengklasifikasikan senyawa aktif dan tidak aktif. Neural Network, meskipun menunjukkan sedikit overfitting, tetap memberikan hasil yang memuaskan dalam menangkap pola kompleks dari data bioaktivitas. Analisis fingerprint molekular mengungkapkan bahwa Morgan fingerprint memberikan kontribusi fitur yang paling informatif dalam prediksi aktivitas senyawa. Pendekatan yang dikembangkan memiliki potensi signifikan untuk diaplikasikan dalam proses virtual screening awal untuk mengidentifikasi kandidat JAK inhibitor, meskipun validasi eksperimental tetap diperlukan. Metodologi yang dikembangkan juga dapat diadaptasi untuk target protein serupa, membuka peluang yang lebih luas dalam pengembangan obat. Keberhasilan implementasi berbagai teknik machine learning ini menunjukkan prospek yang menjanjikan dalam mengoptimalkan proses penemuan obat, khususnya dalam tahap awal screening senyawa kandidat. Penelitian ini

memberikan landasan yang kuat untuk pengembangan metode prediksi bioaktivitas yang lebih efisien dan akurat di masa depan.

UCAPAN TERIMA KASIH

Ucapan terimakasih disampaikan dalam kepada dosen pengampu mata kuliah Bioinformatika Program Studi Sains Data Institut Teknologi Sumatera yaitu pak Tirta Setiawan, S.Pd., M.Si.. Ucapan terimakasih diberikan pula kepada seluruh anggota kelompok 7 tugas besar mata kuliah Bioinformatika yang berkontribusi dalam penulisan dan percobaan penelitian “Prediksi dan Klasifikasi Bioaktivitas Senyawa Terhadap Protein Janus Kinase (JAK) Menggunakan Pendekatan Random Forest dan Neural Network” baik dukungan mental, ilmu dan materi, sehingga model dan penulisan Artikel penelitian ini dapat diselesaikan dengan tepat waktu dan mendapatkan hasil yang dapat dipertanggungjawabkan.

DAFTAR RUJUKAN

1. O'Shea, J. J., Schwartz, D. M., Villarino, A. V., Gadina, M., McInnes, I. B., & Laurence, A. (2015). The JAK-STAT pathway: impact on human disease and therapeutic intervention. Annual Review of Medicine, 66, 311-328.
2. Salas, A., Hernandez-Rocha, C., Duijvestein, M., Faubion, W., McGovern, D., Vermeire, S., Vetrano, S., & Vande Casteele, N. (2020). JAK-STAT pathway targeting for the treatment of inflammatory bowel disease. Nature Reviews Gastroenterology & Hepatology, 17, 323-337.
3. Villarino, A. V., Kanno, Y., & O’Shea, J. J. (2017). Mechanisms and consequences of Jak-STAT signaling in the immune system. Nature Immunology, 18, 374-384.
4. Hu, X., Li, J., Fu, M., Zhao, X., & Wang, W. (2021). The JAK/STAT signaling pathway: From bench to clinic. Signal Transduction and Targeted Therapy, 6, 402.
5. Spinelli, F. R., Colbert, R. A., & Gadina, M. (2021). Janus kinase inhibitors: a new era in the treatment of autoimmune diseases. Rheumatology (Oxford), 60, ii3-ii10.
6. Yang, Z., Tian, Y., Kong, Y., Zhu, Y., & Yan, A. (2022). Classification of JAK1 Inhibitors and SAR Research by Machine Learning Methods. Artificial Intelligence in the Life Sciences, 2, 100039.
7. Bu, Y., Gao, R., Zhang, B., Zhang, L., & Sun, D. (2023). CoGT: Ensemble Machine Learning Method and Its Application on JAK Inhibitor Discovery. ACS Omega, 8, 13232-13242.
8. Kramer, A., Prinz, C., Fichtner, F., Fischer, A.-L., Thieme, V., Grundeis, F., Spagl, M., Seeber, C., Piechotta, V., Metzendorf, M.-I., Golinski, M., Moerer, O., Stephani, C., Mikolajewska, A., Kluge, S., Stegemann, M., Laudi, S., & Skoetz, N. (2022). Janus kinase inhibitors for the treatment of COVID-19. Cochrane Database of Systematic Reviews, Issue 6, Art. No.: CD015209. DOI: 10.1002/14651858.CD015209.
9. Ton, A. T., Gentile, F., Hsing, M., Ban, F., & Cherkasov, A. (2020). Potential anti-SARS-CoV-2 drug candidates identified through virtual screening against the main protease. *Journal of Biomolecular Structure and Dynamics*. https://doi.org/10.1080/07391102.2020.1767690
10. Mendez, D., Gaulton, A., Bento, A. P., et al. (2023). ChEMBL Database in 2023: A drug discovery platform spanning the scientific literature and patent space. *Nucleic Acids Research*, *51*(D1), D1160-D1169. https://doi.org/10.1093/nar/gkac1075
11. Grisoni, F., Merk, D., Consonni, V., et

al. (2023). VSFlow: An open-source ligand-based virtual screening tool. *Journal of Cheminformatics*, *15*, 33. https://doi.org/10.1186/s13321-023-00703-1

12. Landrum, G. (2024). Getting started with the RDKit in Python. RDKit Documentation. Retrieved from https://www.rdkit.org/docs/GettingStartedInPython.html
13. Lipinski, C. A., et al. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *46*(1-3), 3-26. https://doi.org/10.1016/S0169-409X(01)00129-0
14. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. Retrieved from https://www.jair.org/index.php/jair/article/view/10302
15. Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742-754. https://doi.org/10.1021/ci100050t
16. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.
17. Matplotlib Documentation. (n.d.). Matplotlib: Visualization with Python. Retrieved from https://matplotlib.org