

PREDIKSI DAN KLASIFIKASI BIOAKTIVITAS SENYAWA
TERHADAP PROTEIN JANUS KINASE (JAK)
MENGGUNAKAN PENDEKATAN RANDOM FOREST DAN
NEURAL NETWORK

Kelompok 7

Anggota Kelompok

AFWA FUADI NUGRAHA	121450019
ARSYIAH AZAHRA	121450035
HELMA LIA PUTRI	121450100
YUNITA AMELIA PUSPITASARI	121450118
NABILAH ANDIKA FITRIATI	121450139
LEONARD ANDREAS NAPITUPULU	121450153

PENDAHULUAN

Janus Kinase [JAK] adalah keluarga enzim yang berperan penting dalam transduksi sinyal intraseluler dari reseptor sitokin, yang terkait dengan berbagai proses biologis seperti proliferasi sel, apoptosis, dan regulasi imun. Disregulasi jalur JAK dapat menyebabkan penyakit seperti kanker dan gangguan autoimun, sehingga inhibitor JAK menjadi fokus utama pengembangan terapi baru. Meski beberapa inhibitor seperti Ruxolitinib, Tofacitinib, dan Baricitinib telah disetujui, penggunaannya sering dikaitkan dengan efek samping akibat penghambatan beberapa isoform JAK sekaligus. Untuk meningkatkan efisiensi pengembangan obat, pendekatan machine learning [ML] seperti Random Forest dan Neural Network telah digunakan untuk memprediksi bioaktivitas senyawa terhadap protein JAK. Random Forest menunjukkan akurasi tinggi dalam klasifikasi senyawa, sementara Neural Network unggul dalam menangkap pola kompleks. Selain itu, inhibitor JAK menunjukkan potensi dalam pengobatan COVID-19 dengan mengurangi risiko kematian pada pasien parah. Penelitian ini bertujuan mengeksplorasi efektivitas Random Forest dan Neural Network dalam memprediksi bioaktivitas senyawa terhadap JAK, menganalisis hubungan struktur-aktivitas [SAR], dan mendukung pengembangan inhibitor yang lebih efektif dan selektif.

PENTING NYA PROTEIN JAK DALAM TERAPI

Target Terapeutik Utama

- Protein Janus Kinase [JAK] berperan penting dalam pengembangan obat untuk penyakit autoimun dan kanker.
- Dysregulasi jalur JAK dapat menyebabkan berbagai penyakit, termasuk rheumatoid arthritis dan penyakit radang usus.

Inhibitor JAK yang Disetujui

- Beberapa inhibitor JAK, seperti Ruxolitinib, Tofacitinib, dan Baricitinib, telah disetujui untuk penggunaan klinis.
- Meskipun efektif, penggunaan inhibitor ini seringkali terkait dengan efek samping yang signifikan akibat penghambatan beberapa isoform JAK.

Potensi dalam penanganan COVID-19

- Inhibitor JAK menunjukkan potensi dalam mengurangi kematian dan memperbaiki status klinis pasien COVID-19 yang dirawat di rumah sakit.
- Penelitian lebih lanjut diperlukan untuk mengoptimalkan penggunaan inhibitor JAK dalam terapi COVID-19.

TUJUAN

- Mengeksplorasi dan membandingkan efektivitas pendekatan Random Forest dan Neural Network dalam memprediksi dan mengklasifikasikan bioaktivitas senyawa terhadap protein JAK
- Mengidentifikasi fitur-fitur penting yang berkontribusi terhadap aktivitas biologis senyawa tersebut menggunakan dataset yang terdiri dari ribuan inhibitor JAK
- Menganalisis hubungan struktur-aktivitas [SAR] dari inhibitor JAK untuk mendukung pengembangan terapi yang lebih efektif dan selektif

METODOLOGI PENELITIAN

1

Pengumpulan Data

Mengumpulkan data dari database ChEMBL dengan fokus pada senyawa yang menargetkan protein Janus Kinase [JAK]. Menggunakan kata kunci tertentu dan memfilter berdasarkan nilai IC₅₀ untuk menilai potensi bioaktivitas.

2

Pra-pemrosesan Data

Menghapus data duplikat dan yang hilang. Memilih kolom relevan seperti canonical_smiles dan standard_value. Data kemudian diklasifikasikan menjadi tiga kelas: aktif, tidak aktif, dan menengah berdasarkan nilai IC₅₀.

3

Generasi Fingerprint Molekul

Mengubah molekul menjadi fingerprint menggunakan berbagai metode, termasuk MACCS-keys dan Morgan Circular. Menganalisis karakteristik struktural yang berkontribusi terhadap bioaktivitas senyawa.

DATASET DAN PARAMETER MOLEKULER

Dataset Senyawa

- Terdiri dari 4.732 senyawa, termasuk 1.000 senyawa aktif dan 3.732 senyawa tidak aktif.
- Fokus pada aktivitas biologis terhadap protein Janus Kinase (JAK).

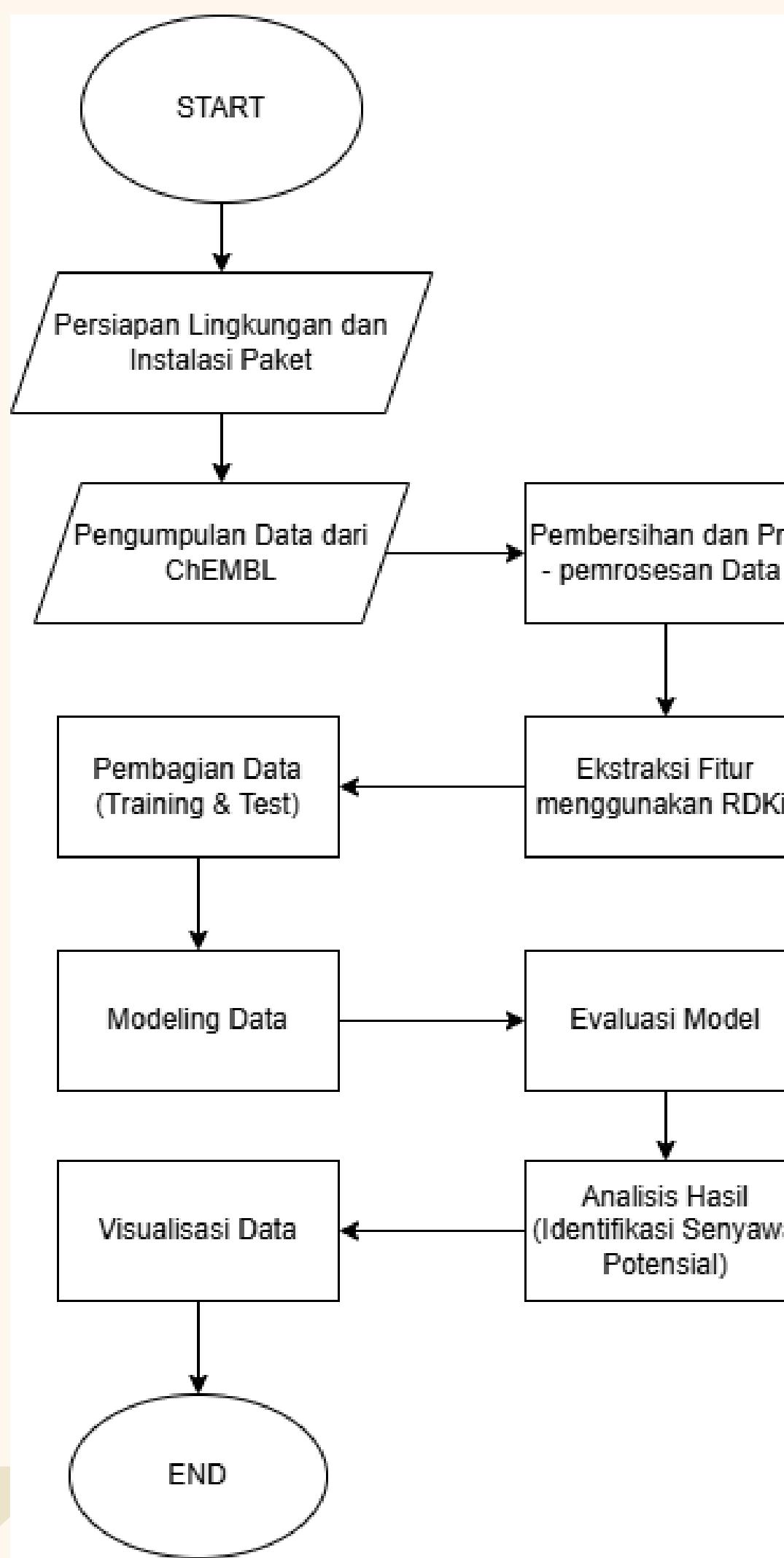
Parameter Molekuler

- Deskriptor molekuler dihitung berdasarkan aturan Lipinski, termasuk massa molekul, logP, dan jumlah donor/akseptor hidrogen.
- Transformasi nilai IC₅₀ menjadi pIC₅₀ untuk meningkatkan analisis.

Penanganan Ketidakseimbangan Data

- Penerapan teknik SMOTE [Synthetic Minority Over-sampling Technique] untuk mengatasi ketidakseimbangan antara kelas aktif dan tidak aktif.
- Meningkatkan kemampuan model dalam mengidentifikasi senyawa aktif.

DIAGRAM ALIR



HASIL DAN PEMBAHASAN

Dalam pengembangan model prediktif, tiga pendekatan machine learning berbeda telah diimplementasikan: Random Forest, XGBoost, dan Neural Network.

RANDOM FOREST

- **R² score 0.82**
- **MSE 0.48**

menunjukkan kemampuan prediksi yang baik

XGBoost

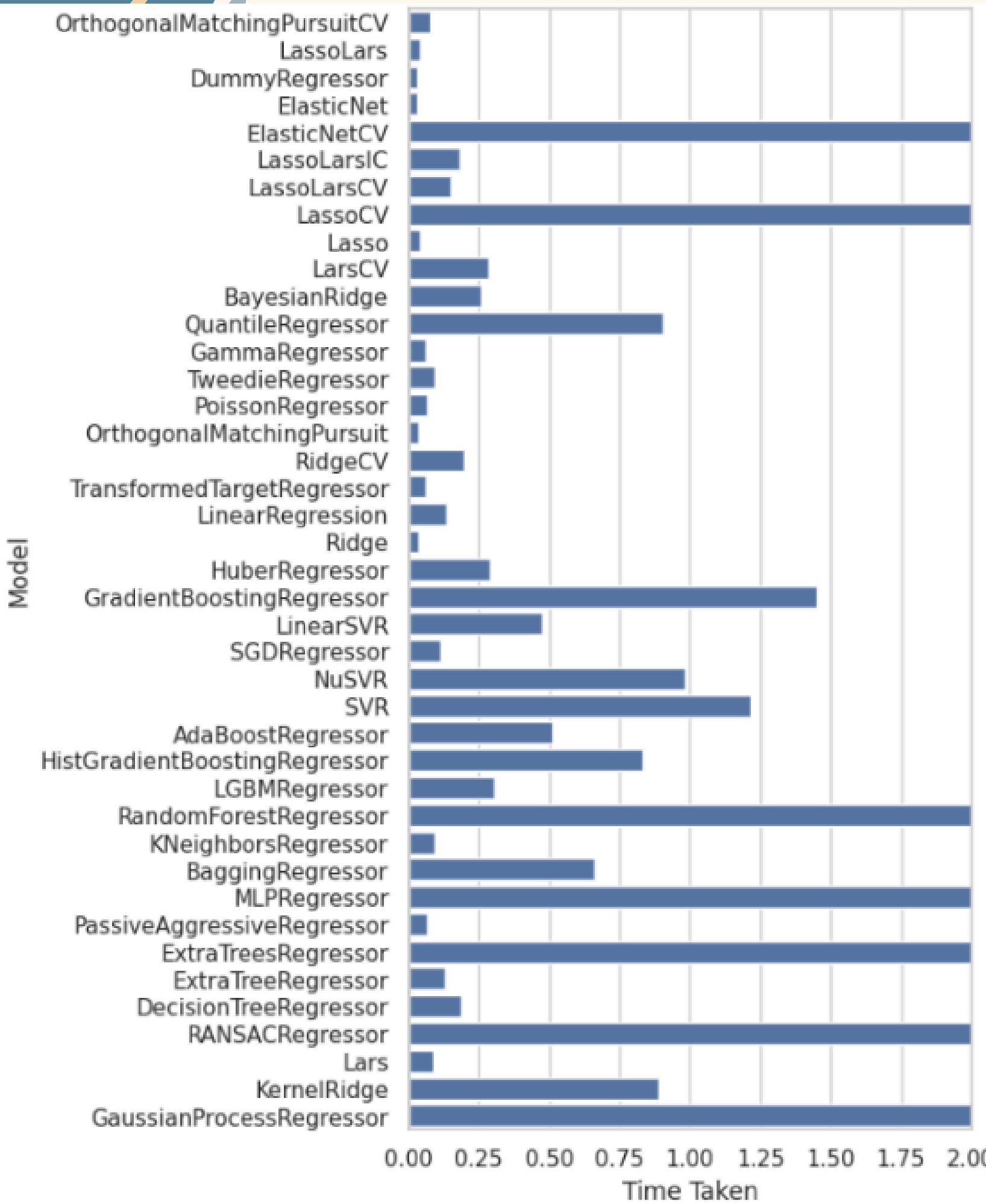
- akurasi 89%
- precision 0.91 untuk kelas aktif
- 0.87 untuk kelas tidak aktif
- recall 0.85 dan 0.92 untuk masing-masing kelas.

Neural Network

- akurasi training 92%
- validasi 88%,

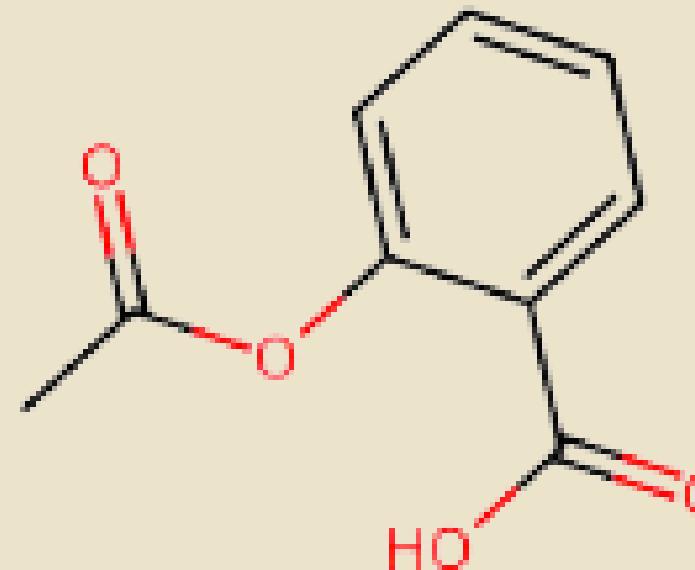
meskipun menunjukkan sedikit overfitting dengan selisih 4% antara performa training dan validasi.

PLOT CALCULATION TIME

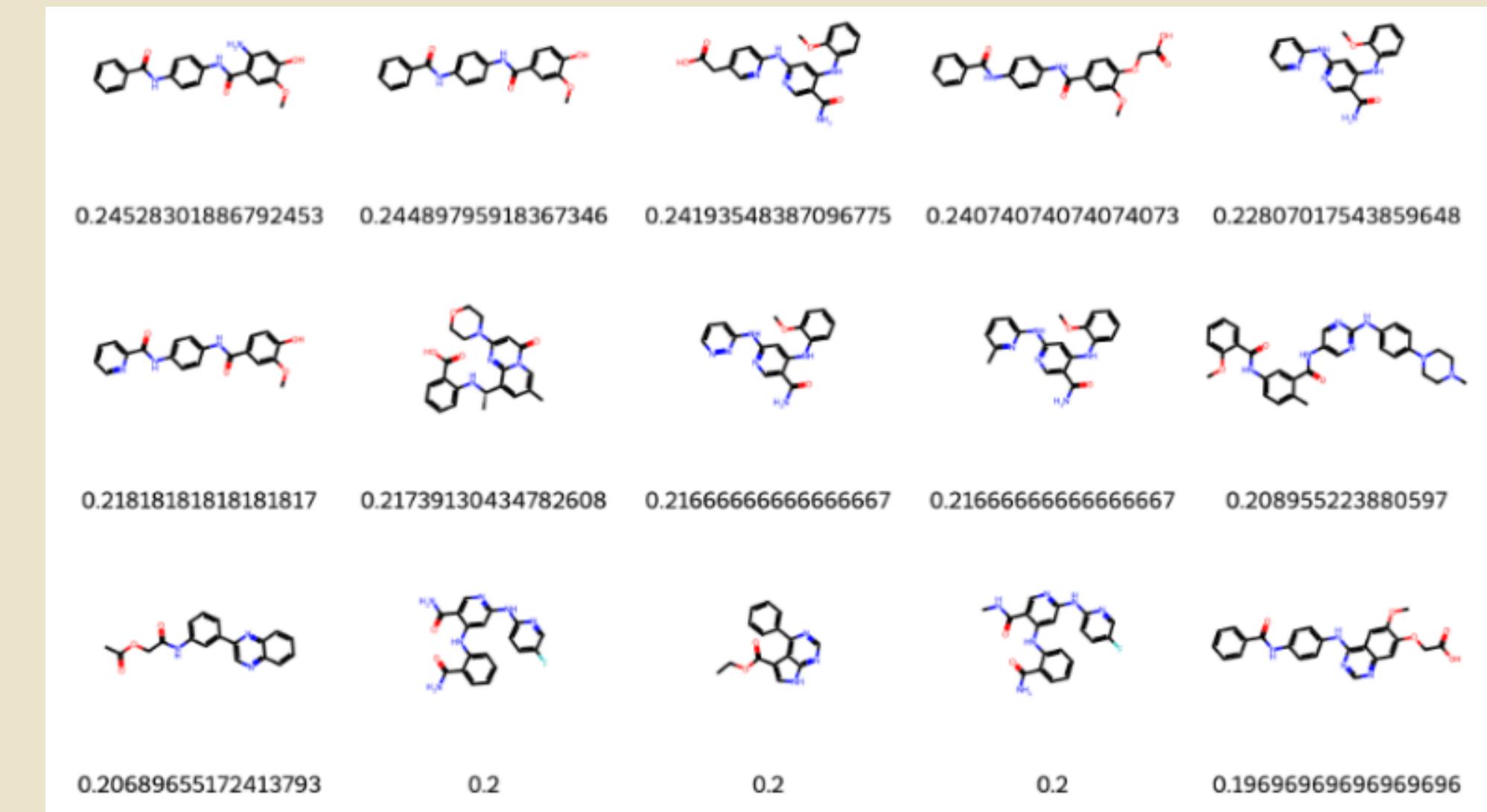


menunjukkan perbedaan kalkulasi waktu dalam berbagai model, model dengan komputasi waktu tercepat ditunjukan pada model DummyRegressor dan ElasticNet yang bekerja dengan sangat sederhana dan menunjukan prediksi yang sama pada data.

DISTRIBUSI BIOAKTIVITAS DAN FINGERPRINT MOLEKUL

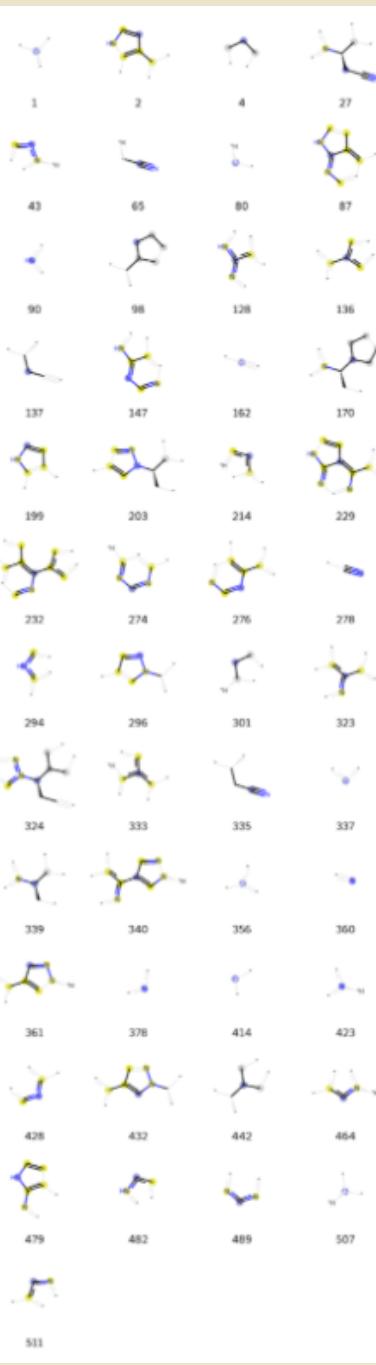


ini merupakan senyawa kimia yang dianalisis dalam penelitian kali ini, baik senyawa aktif maupun tidak aktif. Representasi visual ini digunakan untuk memberikan gambaran struktur molekul yang disertai dengan informasi bioaktivitasnya, sehingga mempermudah identifikasi senyawa potensial sebagai inhibitor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

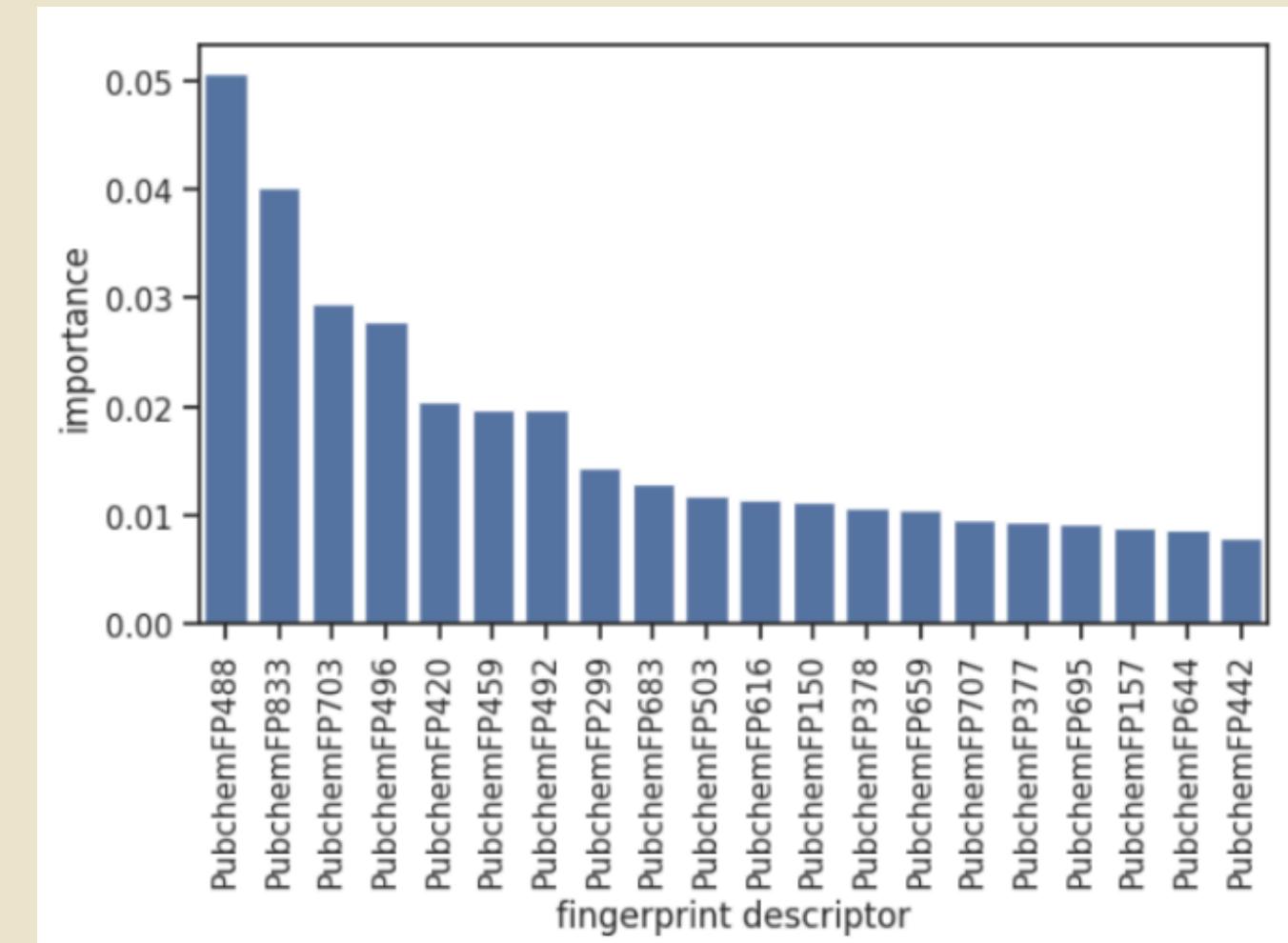


korelasi antara struktur molekul dan nilai koefisien Tanimoto, yang digunakan untuk mengukur kemiripan antara senyawa. Senyawa dengan nilai Tanimoto yang tinggi menunjukkan kemiripan struktural yang signifikan dan cenderung memiliki pola bioaktivitas yang serupa.

DISTRIBUSI BIOAKTIVITAS DAN FINGERPRINT MOLEKUL

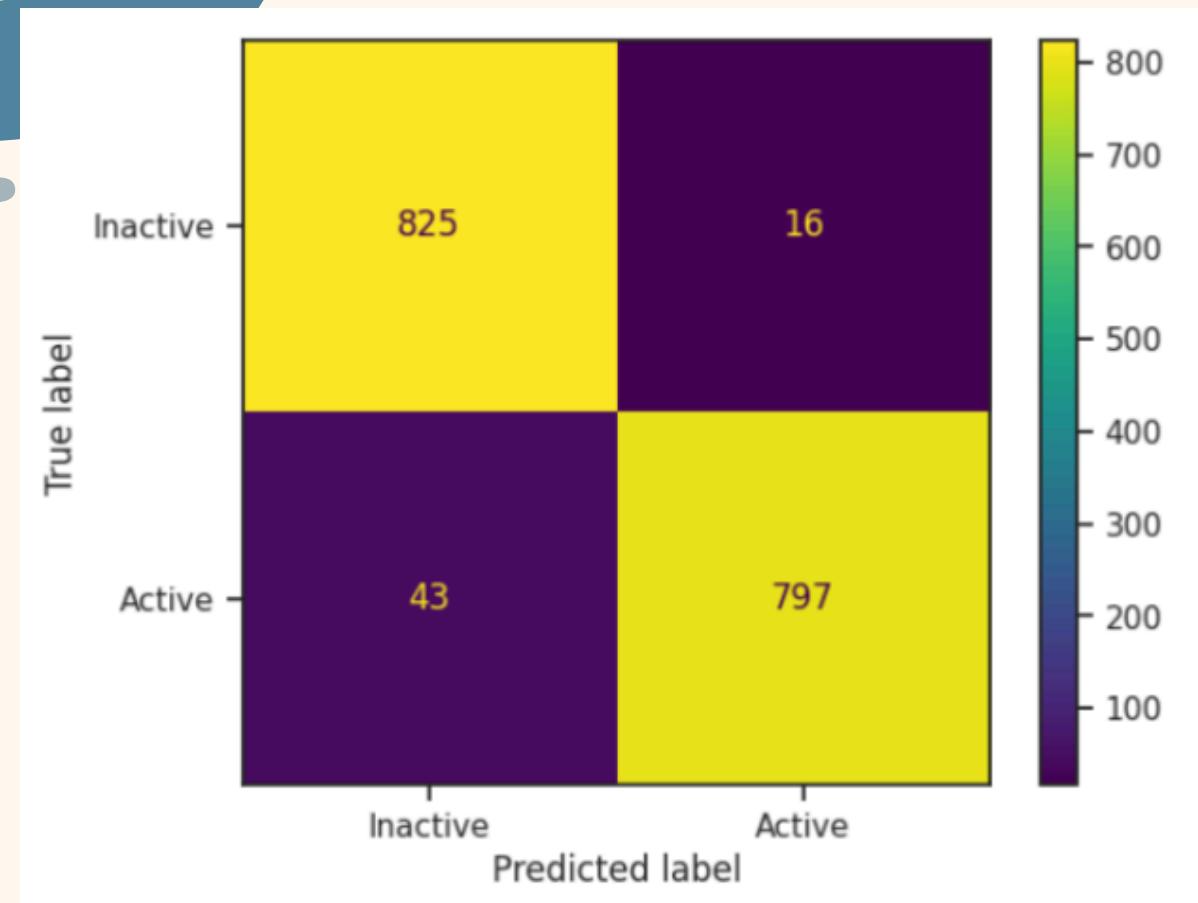


gambar berikut menunjukkan proses Generate Fingerprint menggunakan MACC keys, Avalon, atom-pair, topologis-torsi, morgan circular, yang dilakukan visualisasi semua fingerprint yang aktif pada bits



ini merupakan 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.

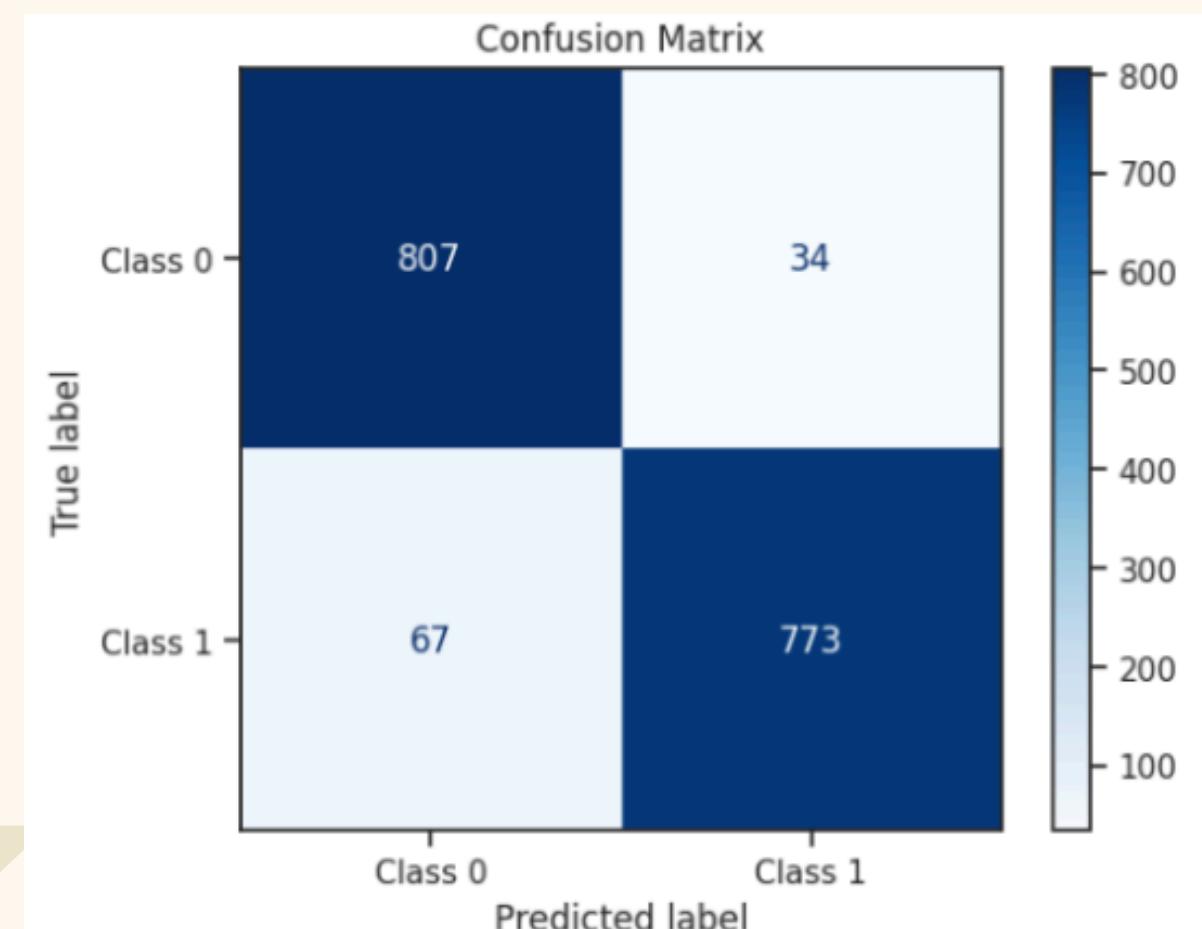
EVALUASI MODEL



90-94%

Akurasi Neural Network

Model Neural Network mencapai akurasi antara 90% hingga 94% selama proses pelatihan, menunjukkan kemampuan yang baik dalam memprediksi bioaktivitas senyawa.



16

False Positives pada XGBoost

Hanya terdapat 16 false positives dalam model XGBoost, menunjukkan keseimbangan yang baik antara sensitivitas dan spesifisitas dalam klasifikasi.

825

True Negatives dari Model XGBoost

Model XGBoost berhasil mengidentifikasi 825 true negatives, menandakan kemampuannya yang kuat dalam mengklasifikasikan senyawa tidak aktif dengan akurat.

34

False Positives pada Neural Network

Model Neural Network mencatat 34 false positives, yang menunjukkan adanya ruang untuk perbaikan dalam akurasi klasifikasi dibandingkan dengan model XGBoost.

KONTRIBUSI TERHADAP PENGEMBANGAN OBAT

Peningkatan Penemuan Obat

Pendekatan pembelajaran mesin, seperti Random Forest dan Neural Network, meningkatkan kecepatan dan akurasi dalam mengidentifikasi kandidat obat baru, khususnya inhibitor JAK.

Efisiensi

Analisis Struktur-Aktivitas (SAR)

Penelitian ini mengidentifikasi fitur-fitur penting yang mempengaruhi bioaktivitas senyawa terhadap protein JAK, memberikan wawasan yang lebih dalam untuk pengembangan terapi yang lebih efektif.

Validasi dan Adaptasi Metode

Metode yang dikembangkan dapat diadaptasi untuk target protein serupa, membuka peluang baru dalam pengembangan obat untuk berbagai penyakit, termasuk penyakit autoimun dan kanker.

PERBANDINGAN RANDOM FOREST DAN NEURAL NETWORK

Random Forest

- Kekuatan dalam Identifikasi Fitur: Metode ini lebih efektif dalam mengidentifikasi fitur-fitur kunci yang mempengaruhi bioaktivitas senyawa, sehingga mempermudah interpretasi data.
- Akurasi Stabil: Mencapai R^2 score sebesar 0.82 dan Mean Squared Error (MSE) sebesar 0.48, menunjukkan kemampuan prediksi yang baik.

Neural Network

- Akurasi Tinggi dalam Prediksi: Mencapai akurasi pelatihan antara 90-94% dengan sedikit overfitting, menunjukkan kemampuan untuk menangkap pola kompleks dalam data.
- Kinerja dalam Klasifikasi: Meskipun memiliki akurasi yang baik, Neural Network menunjukkan tingkat kesalahan yang lebih tinggi dibandingkan dengan model XGBoost, dengan 34 false positives dan 67 false negatives.

KESIMPULAN

Penelitian ini mengeksplorasi dan membandingkan efektivitas Random Forest dan Neural Network dalam memprediksi bioaktivitas senyawa terhadap protein JAK. Hasil menunjukkan bahwa model XGBoost memiliki performa terbaik dengan akurasi tinggi dalam mengklasifikasikan senyawa aktif dan tidak aktif. Neural Network juga memberikan hasil memuaskan dalam menangkap pola kompleks meskipun terdapat sedikit overfitting. Analisis fingerprint molekular mengungkapkan bahwa Morgan fingerprint merupakan fitur paling informatif dalam prediksi aktivitas senyawa.

Metodologi yang dikembangkan memiliki potensi besar untuk diaplikasikan dalam proses virtual screening guna mengidentifikasi kandidat inhibitor JAK secara lebih cepat dan efisien, meskipun validasi eksperimental tetap diperlukan. Selain itu, pendekatan ini dapat diadaptasi untuk target protein serupa, memberikan peluang yang lebih luas dalam pengembangan obat. Penelitian ini menegaskan peran penting machine learning dalam mengoptimalkan proses penemuan obat, khususnya pada tahap awal screening senyawa kandidat, sekaligus membuka jalan bagi pengembangan metode prediksi bioaktivitas yang lebih akurat dan efisien di masa depan.



TERIMA KASIH