Prediksi Aktivitas Obat Secara In Silico Menggunakan Deskriptor dan Fingerprint Molekuler

In Silico Prediction of Drug Activity Using Molecular Descriptors and Fingerprints

Della Septiani[1 *], Dara Cantika Dewi[2], Aisyah Tiara Pratiwi[3], Mahdia Nisrina Maharani M[4], Kholisaturrohmah[5]

[1] Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera, Lampung Selatan, Indonesia
**E-mail: della.121450109@student.itera.ac.id*

Abstrak

Proyek ini bertujuan untuk mengembangkan model prediksi aktivitas senyawa terhadap EGFR (Epidermal Growth Factor Receptor) menggunakan teknik pembelajaran mesin. Model ini mengandalkan fingerprint molekuler dan deskriptor seperti berat molekul (MW), logP, serta jumlah donor dan akseptor hidrogen untuk memprediksi nilai pIC50 senyawa terhadap EGFR. Dataset yang digunakan mencakup senyawa dengan aktivitas EGFR yang telah diketahui. Hasil evaluasi menunjukkan bahwa model memiliki Mean Squared Error (MSE) sebesar 1.41 dan R^2 sebesar 0.50 pada data aktual, yang menunjukkan kinerja model yang moderat. Sementara pada data prediksi, model menunjukkan MSE sebesar 0.43 dan R^2 sebesar 0.84, yang mengindikasikan peningkatan akurasi prediksi yang signifikan. Dengan demikian, model ini dapat menjadi alat yang efektif dalam menyaring senyawa potensial untuk pengembangan obat EGFR, mengurangi waktu dan biaya dalam penemuan obat. Hasil ini menunjukkan bahwa pendekatan komputasional berbasis pembelajaran mesin dapat berkontribusi pada efisiensi proses penemuan obat dengan target EGFR.

Kata kunci: EGFR; pIC50; *Canonical Smiles*; *Virtual Screening*; Kanker

Abstract

This project aims to develop a predictive model for compound activity against EGFR (Epidermal Growth Factor Receptor) using machine learning techniques. The model relies on molecular fingerprints and descriptors such as molecular weight (MW), logP, and the number of hydrogen donors and acceptors to predict the pIC50 values of compounds against EGFR. The dataset used includes compounds with known EGFR activity. Evaluation results show that the model has a Mean Squared Error (MSE) of 1.41 and an R^2 of 0.50 on actual data, indicating moderate model performance. However, on predicted data, the model achieves an MSE of 0.43 and an R^2 of 0.84, indicating a significant improvement in prediction accuracy. Thus, this model can serve as an effective tool for screening potential EGFR inhibitors, reducing time and costs in drug discovery. These results demonstrate that computational approaches based on machine learning can contribute to the efficiency of the drug discovery process targeting EGFR.

Keywords: EGFR; pIC50; Canonical Smiles; Virtual Screening; Cancer

PENDAHULUAN

Kanker merupakan salah satu penyebab utama kematian di dunia, dengan angka kejadian yang terus meningkat setiap tahun. Salah satu jenis kanker yang paling umum adalah kanker paru-paru, yang terdiri dari dua kategori utama: *small cell lung cancer* (SCLC) dan *non-small cell lung cancer* (NSCLC). Menurut American Cancer Society (2020), kanker paru-paru menyumbang 25% dari total kematian akibat kanker, dengan 228.280 kasus baru di Amerika Serikat pada tahun 2020, termasuk 116.300 laki-laki dan 112.520 perempuan. Di Indonesia, kanker paru-paru juga menjadi penyebab utama kematian terkait kanker, menyumbang 11,4% dari total kematian pada tahun 2020 (WHO, 2020). Angka ini diprediksi meningkat hingga 83% pada tahun 2040 dibandingkan tahun 2018 [1].

Target terapi kanker yang telah menjadi fokus penelitian adalah *Epidermal Growth Factor Receptor* atau disebut juga dengan EGFR, yang merupakan protein yang berperan penting dalam regulasi pertumbuhan dan proliferasi sel. Namun, mutasi atau aktivitas berlebih dari EGFR sering kali dikaitkan dengan perkembangan berbagai jenis kanker, seperti kanker paru-paru, payudara, dan kolorektal [2]. Oleh karena itu, penghambatan aktivitas EGFR menjadi salah satu strategi yang menjanjikan dalam pengembangan terapi kanker.

Pendekatan in silico berbasis pembelajaran mesin menjadi alternatif yang efisien dalam mengatasi keterbatasan tersebut. Virtual screening menggunakan data molekuler, seperti deskriptor (berat molekul, logP, jumlah donor dan akseptor hidrogen) dan fingerprint molekuler, memungkinkan prediksi aktivitas senyawa terhadap EGFR secara akurat. Penelitian ini dilakukan untuk menganalisis bioaktivitas senyawa terhadap target EGFR menggunakan data dari platform ChEMBL. Metode yang digunakan adalah Random Forest Regresi, sebuah algoritma pembelajaran mesin yang mampu menangani data non-linear secara efektif, untuk memprediksi nilai aktivitas biologis berdasarkan data molekuler yang tersedia. Dataset terdiri dari senyawa dengan nilai pIC50 yang telah diketahui, di mana pIC50 digunakan sebagai ukuran efektivitas senyawa dalam menghambat aktivitas EGFR [3].

Penelitian ini bertujuan untuk mengembangkan model prediksi aktivitas senyawa terhadap EGFR menggunakan teknik pembelajaran mesin. Model ini memanfaatkan fingerprint molekuler dan berbagai deskriptor kimia, seperti berat molekul (MW), logP, serta jumlah donor dan akseptor hidrogen. Dengan menggunakan dataset senyawa yang telah diketahui aktivitasnya terhadap EGFR, model ini diharapkan mampu memprediksi nilai pIC50, yaitu ukuran efektivitas senyawa dalam menghambat EGFR [4]. Pendekatan ini tidak hanya bertujuan untuk meningkatkan akurasi prediksi tetapi juga untuk mengurangi waktu dan biaya yang diperlukan dalam proses penemuan obat [5]. Dengan memanfaatkan model prediksi berbasis pembelajaran mesin, penelitian ini berkontribusi pada pengembangan teknologi komputasional dalam bidang farmasi, khususnya dalam desain obat yang ditargetkan untuk kanker.

METODE

Penelitian ini dilakukan untuk menganalisis bioaktivitas senyawa terhadap target EGFR menggunakan data dari platform ChEMBL. Metode yang digunakan adalah Random Forest Regresi untuk memprediksi nilai aktivitas biologis berdasarkan data yang tersedia. Berikut adalah penjelasan lengkap mengenai data yang digunakan dan pendekatan yang diterapkan.

Alat dan Bahan

Alat yang digunakan adalah Laptop untuk menjalankan perangkat lunak dan

pemrosesan data. Dalam penelitian ini, komputer yang digunakan adalah Lenovo ideapad 5 dengan spesifikasi *prosesor* Intel Core i7 dan RAM 12GB, yang mendukung kinerja pemrograman dan analisis data. Program dijalankan di bahasa pemrograman Python menggunakan Google Colab. Untuk Bahan, disertakan dataset yang diambil dari API ChEMBL yang memuat senyawa kimia dengan aktivitas terhadap EGFR.

Deskripsi Data

Dataset yang digunakan dalam penelitian ini adalah EGFR Bioactivity Dataset, yang terdiri dari 33.727 baris dan 9 kolom. Dataset ini diperoleh dari platform ChEMBL Dataset melalui tautan CHEMBL. Dataset mencakup informasi bioaktivitas senyawa, seperti pengukuran IC50, Ki, dan aktivitas lain terhadap target biologis EGFR. Berikut adalah kolom-kolom yang terdapat dalam dataset :

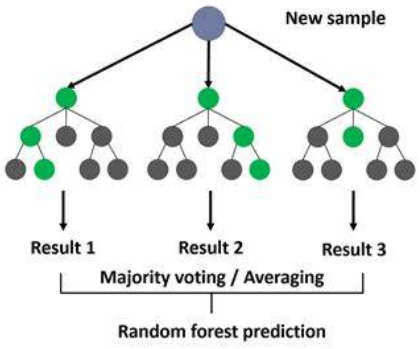
- *Molecule* ChEMBL ID : ID unik untuk molekul kecil. Berguna untuk pelacakan molekul.
- AssayChEMBLID : ID unik untuk percobaan (assay) bioaktivitas. Ini menghubungkan molekul ke data uji biologis.
- *Activity Type* : Jenis aktivitas yang diukur (misalnya, IC50, EC50, Ki, dll.). Menunjukkan seberapa efektif molekul berinteraksi dengan targetnya (EGFR).
- *Value* : Nilai aktivitas biologis yang diukur .
- *Units* : satuannya (misalnya, IC50 dalam nM).
- *Standard Type*: Versi standar dari *Activity Type*.
- Standard Value : Nilai aktivitas biologis dalam satuan standar
- *Standard Units* : Satuan standar dari aktivitas (misalnya, nM).
- SMILES : Representasi string dari struktur kimia molekul kecil. Berguna untuk analisis kimia, manipulasi molekul, dan docking.

Dataset pada penelitian ini dapat dilihat lebih rinci pada **Tabel 1.** yang dilampirkan dalam halaman tersendiri setelah badan naskah. Dataset ini digunakan sebagai input untuk model Random Forest Regressor, yang bertujuan memprediksi nilai aktivitas biologis senyawa berdasarkan karakteristik molekulnya. Untuk memahami lebih lanjut, metode yang digunakan dijelaskan pada subbab berikut.

Random Forest Regressor

Random Forest Regressor adalah algoritma pembelajaran mesin berbasis *supervised learning* yang populer untuk analisis data. Algoritma ini memanfaatkan data berlabel untuk melatih model, menggabungkan prediksi dari beberapa Pohon Keputusan (*Decision Trees*) untuk menghasilkan hasil yang lebih akurat dan andal [6].

Dalam penelitian ini, algoritma digunakan untuk memprediksi nilai bioaktivitas senyawa terhadap target EGFR berdasarkan dataset dari ChEMBL. Model dilatih menggunakan data seperti jenis aktivitas, nilai standar, satuan, dan struktur molekul dalam format SMILES. Dengan memanfaatkan kombinasi prediksi dari beberapa pohon, algoritma ini diharapkan mampu mengidentifikasi senyawa dengan potensi bioaktivitas tinggi terhadap EGFR. Berikut ilustrasi dari model Random Forest Regressor dapat dilihat pada **Gambar 1.**



Gambar 1. Ilustrasi Random Forest Regressor

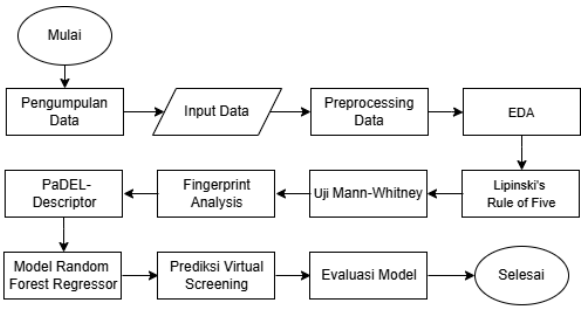
Tahapan Penelitian

1. Pengumpulan data
   Dataset diperoleh dari platform

ChEMBL, berisi informasi bioaktivitas senyawa seperti SMILES, nilai IC50, dan atribut lainnya.

2. *Preprocessing*
   Data dibersihkan dari duplikasi dan nilai kosong, kemudian distandarisasi ke dalam satuan nanomolar (nM) serta dinormalisasi untuk memastikan keseragaman.
3. EDA
   EDA dilakukan untuk memahami distribusi nilai bioaktivitas, korelasi antar fitur, dan mengidentifikasi senyawa aktif terhadap EGFR.
4. *Lipinski Rule of Fine*
   Senyawa dianalisis menggunakan *Lipinski Rule of Five* untuk memastikan sifat farmakokinetik yang baik, seperti berat molekul dan logP.
5. Uji Mann Whitney
   Statistik Mann-Whitney digunakan untuk membandingkan nilai bioaktivitas antara senyawa aktif dan tidak aktif.
6. *Fingerprint Analysis*
   Fingerprint molekul diekstraksi dari SMILES untuk menghasilkan representasi numerik struktur kimia senyawa
7. *PaDEL-Descriptor*
   Deskriptor molekul dihasilkan dengan PaDEL-Descriptor, mencakup sifat fisikokimia sebagai input model.
8. Model *Random Forest Regressor*
   Random Forest Regressor digunakan untuk memprediksi nilai bioaktivitas senyawa berdasarkan *deskriptor* molekul.
9. Prediksi *Virtual Screening*
   Model yang telah dilatih digunakan untuk memprediksi bioaktivitas senyawa baru yang belum diuji secara eksperimen.
10. Evaluasi Model
    Model dievaluasi menggunakan R^2 dan MSE untuk memastikan akurasi dan kemampuan generalisasi prediksi.

Flowchart



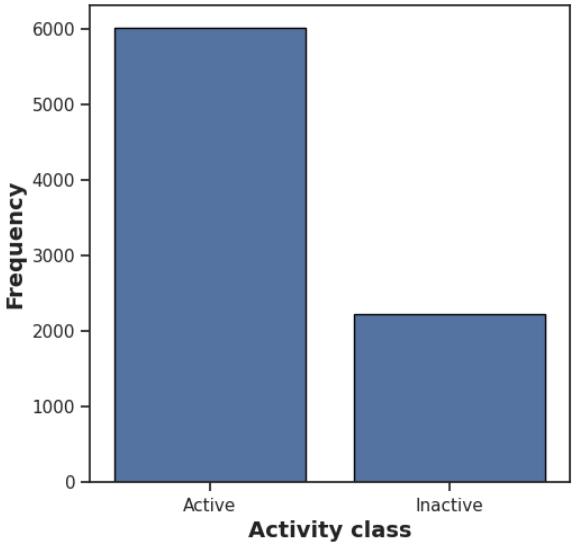
Gambar 2. Flowchart

HASIL DAN PEMBAHASAN

Penelitian ini menghadirkan analisis mendalam mengenai pengembangan model prediktif untuk aktivitas senyawa inhibitor EGFR, menggunakan dataset komprehensif yang diperoleh dari database ChEMBL. Proses *preprocessing* data yang dilakukan dengan sangat teliti menghasilkan dataset final yang terdiri dari 8,251 senyawa, yang kemudian diklasifikasikan secara sistematis berdasarkan nilai IC50. Klasifikasi ini mencakup tiga kategori utama yang memungkinkan diferensiasi yang jelas antara senyawa-senyawa dengan potensi inhibisi berbeda: kategori aktif (IC50 < 100 nM) yang menunjukkan potensi inhibisi kuat, intermediate (100 nM ≤ IC50 ≤ 10,000 nM) untuk potensi inhibisi moderat, dan inaktif (IC50 > 10,000 nM) untuk senyawa dengan aktivitas minimal atau tidak ada. Tahapan preprocessing melibatkan serangkaian langkah kritis termasuk penghapusan missing values, eliminasi duplikasi struktur SMILES, serta transformasi matematis nilai IC50 ke pIC50 untuk mengoptimalkan analisis kuantitatif.

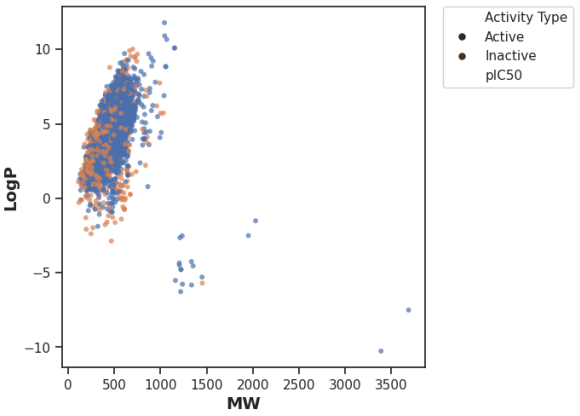
Gambar 3. mengungkapkan pola distribusi yang sangat informatif dan menarik. Grafik batang menampilkan dominasi yang signifikan dari senyawa aktif, dengan rasio proporsi sekitar 3:1 dibandingkan senyawa inaktif. Distribusi yang tidak seimbang ini memberikan wawasan berharga tentang arah penelitian sebelumnya dalam pengembangan inhibitor

EGFR, dimana fokus utama diberikan pada identifikasi dan optimasi senyawa-senyawa dengan aktivitas inhibisi yang kuat. Meskipun bias ini mencerminkan kecenderungan yang wajar dalam pengembangan obat, hal ini juga menimbulkan tantangan metodologis dalam pengembangan model prediktif yang dapat diandalkan. Untuk mengatasi ketidakseimbangan ini, diperlukan strategi validasi silang yang cermat dan interpretasi hasil yang mempertimbangkan karakteristik distribusi dataset.



Gambar 3. Kelas Bioactivity EGFR

Hubungan struktur-aktivitas yang terungkap melalui analisis scatter plot MW versus LogP **Gambar 4**. menjelaskan tentang karakteristik molekular yang berkontribusi terhadap aktivitas inhibitor EGFR. Senyawa-senyawa aktif menunjukkan kecenderungan yang jelas untuk terkonsentrasi pada range berat molekul 300-500 Da dan nilai LogP antara 0-5, yang sangat sesuai dengan prinsip-prinsip drug-likeness yang dikemukakan dalam aturan Lipinski. Korelasi positif yang teramati antara MW dan LogP tidak hanya mengindikasikan hubungan intrinsik antara ukuran molekul dan karakter lipofilik, tetapi juga mengungkapkan adanya "sweet spot" yang optimal untuk aktivitas biologis. Area optimal ini diidentifikasi melalui ukuran titik yang merepresentasikan nilai pIC50, dimana senyawa-senyawa dengan aktivitas tertinggi cenderung memiliki kombinasi MW dan LogP yang seimbang.



Gambar 4. Plot persebaran dengan MW dan LogP

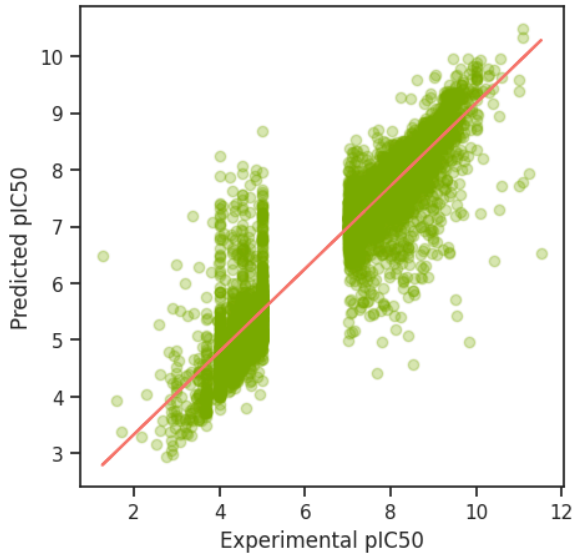
Pada **Tabel 2.** menunjukan hasil analisis statistik menggunakan uji Mann-Whitney menghasilkan temuan yang sangat signifikan untuk seluruh deskriptor Lipinski yang dievaluasi. Nilai p-value yang sangat kecil untuk parameter MW (2.386505e-120) mengindikasikan pentingnya kontrol ukuran molekul dalam menentukan aktivitas biologis. Similaritas, signifikansi yang tinggi untuk LogP (6.453325e-42) menegaskan peran kritis lipofilisitas dalam menentukan interaksi dengan target biologis. Donor ikatan hidrogen (4.655712e-09) dan akseptor ikatan hidrogen (2.922315e-107) juga menunjukkan perbedaan yang sangat signifikan antara senyawa aktif dan inaktif, menggarisbawahi pentingnya kapasitas pembentukan ikatan hidrogen dalam interaksi protein-ligan. Hasil-hasil ini secara kolektif memberikan landasan kuantitatif yang kuat untuk strategi optimasi properti fisikokimia dalam pengembangan inhibitor EGFR yang efektif.

Tabel 2. Uji statistik Mann-Whitney antar kelas

Deskriptor	P-value	Interpretasi

MW	2.386505e-120	Distribusi berbeda (Tolak H0)
LogP	6.453325e-42	Distribusi berbeda (Tolak H0)
NumHDonors	4.655712e-09	Distribusi berbeda (Tolak H0)
NumHAcceptors	2.922315e-107	Distribusi berbeda (Tolak H0)
plC50	0.0	Distribusi berbeda (Tolak H0)

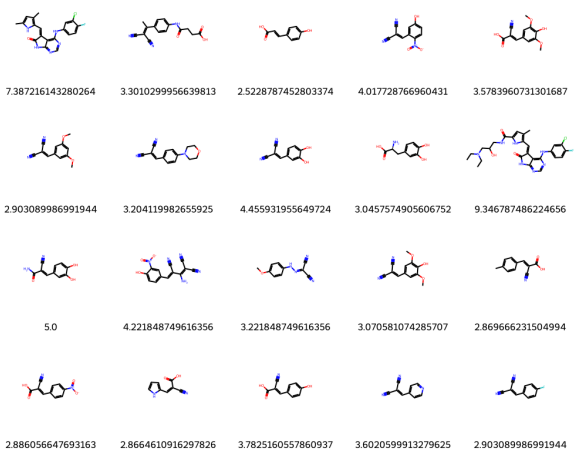
Model Random Forest yang dikembangkan menunjukkan performa yang cukup baik dengan Mean Squared Error (MSE) sebesar 1.41 pada test set dan koefisien determinasi (R^2) sebesar 0.50. Ini menunjukkan model mampu menjelaskan 50% variabilitas data pada test set, meskipun performanya pada keseluruhan data lebih tinggi (R^2 = 0.84). Perbedaan ini mengindikasikan bahwa model masih dapat dioptimalkan, terutama untuk menangani data baru.



Gambar 5. Scatter Plot nilai plC50

Pada **Gambar 5.** *scatter plot* antara nilai pIC50 prediksi dan eksperimental menunjukkan korelasi linear yang cukup baik. Mayoritas prediksi mendekati garis diagonal, menunjukkan prediksi yang akurat pada sebagian besar data. Namun, beberapa outlier terlihat pada senyawa dengan nilai pIC50 yang sangat rendah atau tinggi. Outlier ini kemungkinan disebabkan oleh keterbatasan model dalam menangani senyawa ekstrem atau adanya bias dalam dataset yang lebih banyak memuat senyawa aktif.

Kluster data utama pada plot mengindikasikan dua kelompok senyawa dengan aktivitas rendah (pIC50 2-5) dan aktivitas tinggi (pIC50 6-10). Model memprediksi kluster ini dengan baik, tetapi ada ruang untuk perbaikan, terutama pada data ekstrem. Dengan optimasi lebih lanjut, seperti menyeimbangkan dataset dan menambahkan fitur molekuler baru, model dapat lebih andal untuk digunakan dalam virtual screening inhibitor EGFR.



Gambar 6. Hasil virtual screening senyawa

Dari **Gambar 6.** menunjukkan hasil virtual screening senyawa, dengan struktur molekul yang divisualisasikan bersama nilai pIC50 prediksi. Molekul dengan struktur yang lebih kompleks, seperti yang memiliki beberapa cincin aromatik dan gugus donor hidrogen, cenderung memiliki nilai pIC50 lebih tinggi misalnya 7.38. Sebaliknya,

molekul dengan struktur yang lebih sederhana menunjukkan nilai pIC50 yang lebih rendah misalnya 2.52. Analisis ini menunjukkan bahwa fitur-fitur kimia seperti cincin aromatik dan gugus fungsional memainkan peran penting dalam interaksi dengan binding site EGFR, mendukung potensi senyawa sebagai kandidat inhibitor.

Implementasi virtual screening terhadap lima senyawa uji menghasilkan prediksi pIC50 yang bervariasi dari 3.88 hingga 4.69. Paracetamol menunjukkan nilai pIC50 tertinggi (4.69), diikuti oleh antibiotik (4.57), ibuprofen (4.42), kafein (4.23), dan aspirin (3.88). Hasil ini menunjukkan bahwa kelima senyawa memiliki potensi inhibisi terhadap EGFR, meskipun aktivitasnya relatif moderat.

Tabel 2. Prediksi pIC50 untuk Lima Senyawa Uji

Molekul	Predicted plC50
Paracetamol	4.685321
Antibotik	4.568879
Ibuprofen	4.419715
Kafein	4.231836
Aspirin	3.877368

Implementasi virtual screening terhadap lima senyawa uji menghasilkan prediksi pIC50 yang bervariasi dari 3.877368 hingga 4.685321. Paracetamol menunjukkan nilai pIC50 tertinggi (4.685321), diikuti oleh antibiotik (4.568879), ibuprofen (4.419715), kafein (4.231836), dan aspirin (3.877368). Hasil ini menunjukkan bahwa kelima senyawa memiliki potensi inhibisi terhadap EGFR, meskipun aktivitasnya relatif moderat.

Analisis struktur molekul mengungkapkan bahwa paracetamol, dengan nilai pIC50 tertinggi, memiliki kombinasi fitur optimal seperti cincin aromatik dan gugus hidroksil yang mendukung interaksi ikatan hidrogen dengan EGFR. Antibiotik menonjol karena memiliki sistem cincin aromatik yang kompleks dan donor hidrogen, yang juga berkontribusi pada afinitasnya. Sebaliknya, aspirin memiliki aktivitas terendah, meskipun memiliki cincin aromatik dan gugus karboksilat, yang kemungkinan disebabkan oleh keterbatasan fleksibilitas struktur dalam menyesuaikan binding site EGFR. Hasil ini menunjukkan bahwa paracetamol dan antibiotik adalah kandidat potensial untuk pengembangan lebih lanjut. Penyesuaian struktur, seperti penambahan gugus fungsional tertentu, dapat dilakukan untuk meningkatkan aktivitas biologis molekul ini.

KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi bioaktivitas senyawa terhadap EGFR menggunakan algoritma Random Forest Regressor dengan memanfaatkan dataset ChEMBL. Model yang dikembangkan menunjukkan performa prediksi yang cukup baik, dengan R^2 sebesar 0,50 pada data uji, meskipun masih terdapat ruang untuk optimasi, terutama pada senyawa ekstrem. Hasil virtual screening menunjukkan bahwa fitur molekul seperti cincin aromatik dan donor hidrogen berperan penting dalam meningkatkan nilai pIC50, dengan paracetamol dan antibiotik menonjol sebagai kandidat potensial untuk pengembangan lebih lanjut. Studi ini membuktikan efisiensi pendekatan pembelajaran mesin dalam mempercepat proses penemuan obat, khususnya dalam desain inhibitor EGFR untuk terapi kanker.

SARAN

penelitian ini dapat diperluas dengan menjalin kolaborasi bersama laboratorium eksperimental untuk menguji hasil virtual screening yang diperoleh. Pengujian secara in vitro atau in vivo terhadap senyawa yang telah diprediksi memiliki aktivitas tinggi terhadap EGFR dapat memberikan validasi eksperimental terhadap model yang dikembangkan. Kolaborasi semacam ini akan menjembatani pendekatan komputasional dengan aplikasi dunia nyata, sehingga hasil penelitian dapat memberikan kontribusi langsung dalam bidang farmasi dan kesehatan.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Tirta Setiawan, S.Pd., M.Si., sebagai dosen Bioinformatika, atas bimbingan, dukungan, dan arahan yang diberikan selama proses penelitian ini. Penjelasan beliau mengenai konsep pembelajaran mesin dan analisis bioinformatika sangat membantu penulis dalam memahami proses prediksi aktivitas senyawa terhadap EGFR serta menyelesaikan penelitian ini dengan baik

DAFTAR RUJUKAN

[1] Prasetyo SJ. Analisis Ketahanan Hidup Penderita Adenokarsinoma Paru dengan Mutasi Exon 19 Del dan 21 L858R yang Mendapatkan Pengobatan EGFR-TKI. Usuacid [Internet]. 2021 [cited 2024 Dec 25]; Available from: https://repositori.usu.ac.id/handle/123456789/49551

[2] Kalhan SC. One carbon metabolism in pregnancy: Impact on maternal, fetal and neonatal health. Molecular and Cellular Endocrinology. 2016 Nov;435:48–60.

[3] Sangande F, Uneputty JP. IDENTIFIKASI SENYAWA BAHAN ALAM SEBAGAI INHBITOR TIROSIN KINASE EGFR: SKRINING IN SILICO BERBASIS FARMAKOFOR DAN MOLECULAR DOCKING. Jurnal Fitofarmaka Indonesia. 2021 Jan 29;8(1):1–6.

[4] Ermawati N. KAJIAN NARATIF: DRUG TARGET THERAPY PADA PASIEN NON SMALL CELL LUNG CANCER (NSCLC) DENGAN MUTASI EGFR POSITIF. JURNAL FARMASI DAN KESEHATAN INDONESIA. 2023 Mar 31;3(1):14–25.

[5] Vlavcheski F, Tsiani E. Attenuation of Free Fatty Acid-Induced Muscle Insulin Resistance by Rosemary Extract. Nutrients. 2018 Nov 2;10(11):1623.

[6] M. A. Pratama, M. Munawaroh, and W. J. Pranoto, "Perbandingan performa algoritma linear regresi dan random forest untuk prediksi harga bawang merah di Kota Samarinda," TEKTONIK: Jurnal Ilmu Teknik, vol. 1, no. 2, pp. 172–182, 2024.

LAMPIRAN TABEL

Tabel 1. Dataset

No	Molecule ChEMBL ID	Assay ChEMBL ID	Activity Type	Value	Units	Standard Type	Standard Value	Standard Units	SMILES
1	CHEMBL68920	CHEMBL674637	Active	0.041	uM	IC50	41	nM	Cc1cc(C)c(/C=C2\C(=O)Nc3ncnc(Nc4ccc(F)c(Cl)c4)c32)[nH]1
2	CHEMBL68920	CHEMBL621151	Intermediate	0.3	uM	IC50	300	nM	Cc1cc(C)c(/C=C2\C(=O)Nc3ncnc(Nc4ccc(F)c(Cl)c4)c32)[nH]1
3	CHEMBL68920	CHEMBL615325	Intermediate	7.82	uM	IC50	7820	nM	Cc1cc(C)c(/C=C2\C(=O)Nc3ncnc(Nc4ccc(F)c(Cl)c4)c32)[nH]1
…..	…..	…..	……	…..	…..	…..	…..	…..	……
33.728	CHEMBL4755229	CHEMBL5304266	Intermediate	100	%	% Ctrl	100	%	CNC(=O)COc1cc2cc(Nc3nc(N4CCC(C(=O)N(C)C)CC4)nc3Cl)cc(OC)c2n(C)c1=O
33.729	CHEMBL5306577	CHEMBL5304478	Active	9	%	Inhibition	9	%	CCC[C@H]1CNCc2c1[nH]c(=O)c(C#N)c2N1CCC2(CC1)CC2