ABSTRAK

Inhibisi enzim Sitokrom P450 (CYP450) pada isoenzim CYP2C9, CYP2D6, dan CYP3A4 merupakan faktor risiko utama terjadinya interaksi obat-obat (*drug-drug interaction*) yang berpotensi memicu toksisitas serius. Pengujian laboratorium konvensional memerlukan biaya tinggi dan waktu yang lama, sehingga pendekatan komputasional berbasis *in silico* menjadi solusi yang mendesak. Namun, pemodelan *machine learning* menghadapi tantangan *class imbalance* pada *dataset* bioaktivitas publik. Selain itu, penggunaan *threshold* standar (0,5) sering menghasilkan nilai sensitivitas yang rendah, yang secara klinis berisiko karena gagal mendeteksi inhibitor berbahaya. Masalah ini diperumit oleh sifat kompleks algoritma *ensemble* dan vektor yang memiliki tingkat transparansi rendah (*black box*), sehingga menyulitkan identifikasi alasan di balik sebuah prediksi medis.

Penelitian ini menerapkan pendekatan kuantitatif dengan desain eksperimental yang bersifat komparatif. Sebanyak 54.364 data bioaktivitas diakuisisi dari database ChEMBL dan PubChem melalui proses *preprocessing* yang ketat, mencakup penghapusan fragmen garam, audit struktur, dan resolusi konflik aktivitas. Representasi molekul dikodekan menggunakan fitur hibrida dari *Morgan fingerprint* dan *MACCS keys* menjadi vektor 2215-bit, dengan penerapan teknik SMOTE untuk menangani ketidakseimbangan data. Sebagai bentuk perlakuan eksperimental, strategi optimasi mencakup pencarian *hyperparameter* melalui *framework* Optuna dan *decision threshold optimization* menggunakan Indeks Youden. Tiga algoritma disandingkan secara komparatif, yaitu *Random Forest*, *Support Vector Machine* (SVM), dan *Extreme Gradient Boosting* (XGBoost), guna menentukan arsitektur terbaik untuk data *sparse* berdimensi tinggi. Hasil penelitian kemudian direalisasikan melalui proses serialisasi model menjadi format biner terkompresi (`.pkl`) menggunakan pustaka joblib, diikuti dengan validasi fungsionalitas inferensi menggunakan kode Python pada Jupyter Notebook.

Hasil eksperimen menunjukkan keunggulan signifikan algoritma XGBoost dengan nilai *Matthews Correlation Coefficient* (MCC) tertinggi secara konsisten pada angka 0,5387 (CYP2C9), 0,5836 (CYP2D6), dan 0,6540 (CYP3A4). Implementasi Indeks Youden terbukti berhasil meningkatkan sensitivitas deteksi secara dramatis, seperti pada target CYP2C9 yang melonjak dari 0,6139 ke 0,7914, sehingga meminimalkan risiko *false negative* pada prediksi inhibitor. Melalui analisis transparansi model dengan metode SHAP (*SHapley Additive exPlanations*), fitur aromatik (`Bit_MACCS_125`) teridentifikasi sebagai kontributor utama aktivitas inhibisi. Validasi lokal pada studi kasus Amlodipine mengonfirmasi kemampuan model dalam menangkap fragmen struktural yang relevan secara farmakologi. Penelitian ini memberikan kontribusi pada pengembangan instrumen *virtual screening* yang akurat, sensitif, dan transparan dalam menjamin keamanan metabolisme obat.

Kata Kunci: Sitokrom P450, Interaksi Obat, *Machine Learning*, *Threshold Optimization*, *Explainable* AI

ABSTRACT

Cytochrome P450 (CYP450) enzyme inhibition, specifically in CYP2C9, CYP2D6, and CYP3A4 isoenzymes, is a primary risk factor for Drug-Drug Interactions (DDI) that can trigger serious toxicity. Conventional laboratory testing is costly and time-consuming, making in silico computational approaches an urgent necessity. However, Machine Learning modeling faces challenges with class imbalance in public bioactivity datasets. Furthermore, the use of standard decision thresholds (0.5) often results in low sensitivity, which is clinically risky as it fails to detect dangerous inhibitors. This issue is complicated by the inherent complexity of ensemble and vector algorithms, often characterized as "black box" models, which make it difficult to understand the logic behind a prediction, thereby hindering transparency in medical decision support systems.

This study applies a quantitative approach with an experimental and comparative design. A total of 54,364 bioactivity data points were acquired from the ChEMBL and PubChem databases through rigorous preprocessing, including salt stripping, structural auditing, and activity conflict resolution. Molecular representations were encoded using hybrid features of Morgan Fingerprints and MACCS Keys into 2215-bit vectors, with the application of the SMOTE technique to handle data imbalance. As a form of experimental treatment, optimization strategies included hyperparameter search via the Optuna framework and Decision Threshold Optimization using Youden's Index. Three algorithms were compared: Random Forest, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), to determine the best architecture for high-dimensional sparse data. The research findings were then realized through model serialization into a compressed binary format (.pkl) using the joblib library, followed by inference functionality validation using a Python script within a Jupyter Notebook environment.

Experimental results demonstrate the significant superiority of the XGBoost algorithm, consistently achieving the highest Matthews Correlation Coefficient (MCC) scores at 0.5387 (CYP2C9), 0.5836 (CYP2D6), and 0.6540 (CYP3A4). The implementation of Youden's Index successfully increased detection sensitivity dramatically; for instance, the CYP2C9 target jumped from 0.6139 to 0.7914, thereby minimizing false-negative risks in inhibitor prediction. Through model transparency analysis using the SHAP (Shapley Additive Explanations) method, aromatic features (Bit_MACCS_125) were identified as the primary contributors to inhibitory activity. Local validation on an Amlodipine case study confirmed the model's ability to capture pharmacologically relevant structural fragments. This research contributes to the development of accurate and transparent virtual screening instruments to ensure drug metabolic safety.

Keywords: *Cytochrome P450, Drug Interactions, Machine Learning, Threshold Optimization, Explainable AI*