



Pemanfaatan *Hybrid Molecular Fingerprints* dan *XGBoost* untuk

Identifikasi Senyawa Herbal Antimalaria

Mediliadita Dwi Dimastia¹, Easy Jovanie², Dwi Bagus Dermawan³, Mohammad Hamim Zajuli Al Faroby⁴

^{1,3} Afiliasi pertama (Jurusan/Program Studi, Universitas)

¹ mediliadita@student.telkomuniversity.ac.id

² easyjovanie@student.telkomuniversity.ac.id

³ dwibagud@student.telkomuniversity.ac.id

⁴ alfarobymhz@telkomuniversity.ac.id

⁴ Afiliasi kedua (Jurusan/Program Studi, Universitas)

Corresponding author email: alfarobymhz@telkomuniversity.ac.id

Abstract: Malaria remains a significant global health challenge, especially in tropical regions, with drug resistance and side effects driving the need for safer alternatives. Herbal compounds offer potential due to their biodiversity and eco-friendly properties. This study employs a supervised learning approach using Extreme Gradient Boosting (XGBoost) to predict antimalarial herbal compounds. Data on active compounds were sourced from PubChem, ChEMBL, DUDE, and OMIM, while herbal compound data came from HerbalDB and IJAH Analytics. The methodology included data collection, hybrid molecular fingerprint feature extraction (PubChem and Klekota-Roth), class labeling, XGBoost modeling, evaluation, and prediction. The XGBoost model achieved 0.75 accuracy, identifying key candidates like Anacardic Acid and Emetine with high probability. Other promising compounds, such as Cannabigerolic Acid and Ginkgolic Acid, require further toxicity and mechanism studies. This research demonstrates that combining hybrid molecular fingerprints with XGBoost accelerates the discovery of scientifically validated antimalarial herbal compounds, offering a faster, data-driven approach to drug development.

Keywords: Malaria, Molecular Fingerprint, Extreme Gradient Boosting

Abstrak: Malaria masih menjadi tantangan kesehatan global, terutama di daerah tropis, dengan resistensi obat dan efek samping yang mendorong pencarian alternatif lebih aman. Senyawa herbal berpotensi karena keanekaragaman hayati dan sifatnya yang ramah lingkungan. Penelitian ini menggunakan pendekatan supervised learning dengan Extreme Gradient Boosting (XGBoost) untuk memprediksi senyawa herbal antimalaria. Data senyawa aktif diperoleh dari PubChem, ChEMBL, DUDE, dan OMIM, sedangkan data senyawa herbal diambil dari HerbalDB dan IJAH Analytics. Metodologi meliputi pengumpulan data, ekstraksi fitur menggunakan hybrid molecular fingerprint (PubChem + Klekota-Roth), pelabelan kelas, pemodelan XGBoost, evaluasi, dan prediksi. Model XGBoost mencapai akurasi 0,75 dan mengidentifikasi kandidat potensial seperti Anacardic Acid dan Emetine dengan probabilitas tinggi. Senyawa lain seperti Cannabigerolic Acid dan Ginkgolic Acid juga menunjukkan potensi, tetapi memerlukan penelitian lebih lanjut terkait toksisitas dan mekanisme kerjanya. Penelitian ini membuktikan bahwa kombinasi hybrid molecular fingerprint dan XGBoost dapat mempercepat penemuan senyawa herbal antimalaria berbasis data ilmiah secara lebih efisien.

Kata kunci: Malaria, Molecular Fingerprint, Extreme Gradient Boosting

I. PENDAHULUAN

Malaria adalah penyakit menular yang disebabkan oleh parasit Plasmodium, yang tetap menjadi ancaman kesehatan global yang mengkhawatirkan. Pada tahun 2021, Organisasi Kesehatan Dunia (WHO) melaporkan 247 juta kasus dan 619.000 kematian akibat malaria, dengan sebagian besar korban berasal dari wilayah tropis dan subtropis, termasuk Afrika dan Asia Tenggara [1]. Di Indonesia, malaria masih endemik di wilayah timur seperti Papua, Papua Barat, dan Nusa Tenggara Timur. Kota Jayapura [2], misalnya, mencatat peningkatan 28% kasus pada tahun 2021 akibat keterbatasan akses terhadap pengobatan dan lingkungan yang mendukung perkembangbiakan nyamuk Anopheles [1]. Tantangan utama dalam eliminasi malaria adalah resistensi parasit terhadap obat antimalaria (seperti klorokuin dan artemisinin) serta resistensi nyamuk terhadap insektisida [3]. Terapi kombinasi berbasis artemisinin (ACT) yang saat ini menjadi andalan mulai kehilangan efektivitasnya



di Asia Tenggara akibat mutasi pada gen kelch13 pada *Plasmodium falciparum* [4]. Selain itu, obat sintetis seperti artesunat dan mefloquin sering menyebabkan efek samping seperti pusing, mual, dan gangguan kardiovaskular, serta memiliki biaya produksi yang tinggi [5].

Kondisi ini mendorong pencarian alternatif pengobatan yang lebih aman, efektif, dan terjangkau. Senyawa herbal dari tanaman lokal Indonesia memiliki potensi besar, mengingat keanekaragaman hayati negara ini menempati peringkat kedua di dunia setelah Brasil, dengan lebih dari 30.000 spesies tanaman, di mana 40% di antaranya endemik [6]. Contoh yang sukses adalah *Artemisia annua*, sumber senyawa artemisinin yang telah menyelamatkan jutaan nyawa dan diakui oleh WHO sebagai pengobatan utama untuk Malaria [7]. Namun, masih banyak tanaman lokal, seperti *Brucea javanica* (kwatol) dan *Arcangelisia flava* (yellowwood), yang belum dieksplorasi secara sistematis, meskipun secara empiris digunakan dalam pengobatan tradisional Malaria [8]. Kemudian, *Artemisia annua* (sumber artemisinin) telah terbukti efektif, tetapi masih banyak tanaman lokal yang belum dieksplorasi. Pendekatan komputasi seperti machine learning dan molecular docking dapat mempercepat identifikasi senyawa herbal antimalaria dengan biaya lebih rendah daripada metode eksperimental tradisional [9].

Hambatan utama dalam meneliti senyawa herbal adalah biaya tinggi dan waktu yang dibutuhkan untuk skrining eksperimental konvensional. Di sinilah pendekatan *in silico* seperti machine learning dan simulasi docking molekuler muncul sebagai solusi revolusioner. Teknik ini mampu memprediksi aktivitas antimalaria senyawa dengan menganalisis fitur kimia (misalnya, molecular fingerprints) dan interaksi dengan protein parasit target [10]. Beberapa studi sebelumnya yang dilakukan oleh Amiroch dkk. (2022) dalam artikel "Machine Learning for the Prediction of Antiviral Compounds Targeting Avian Influenza A/H9N2 Viral Proteins" [11] berhasil mengembangkan model machine learning berbasis algoritma XGBoost dan Support Vector Classification (SVC) untuk memprediksi senyawa antivirus yang menargetkan protein kunci H9N2, seperti hemagglutinin (HA) dan neuraminidase (NA). Studi ini menggunakan pendekatan hybrid molecular fingerprint (PubChem dan Klekota-Roth) untuk mengekstrak fitur kimia senyawa, menghasilkan akurasi prediksi hingga 85-90%, serta mengidentifikasi senyawa kandidat yang berpotensi menghambat replikasi virus. Sementara itu, dalam studi terkait di AIP Conference Proceedings [12], analisis interaksi protein-protein (PPI) pada virus H9N2 mengungkapkan peran kritis protein NS1 dan PB1 dalam proses infeksi dan replikasi virus. Hasil analisis jaringan interaksi ini tidak hanya memvalidasi target protein prioritas untuk pengembangan obat, tetapi juga memberikan dasar biologis untuk pemilihan fitur dalam pemodelan machine learning. Kedua studi ini saling melengkapi dalam menyediakan kerangka kerja komputasi terintegrasi, mulai dari identifikasi target biologis hingga prediksi senyawa antivirus, sehingga memperkuat pendekatan *in silico* dalam penemuan obat influenza.

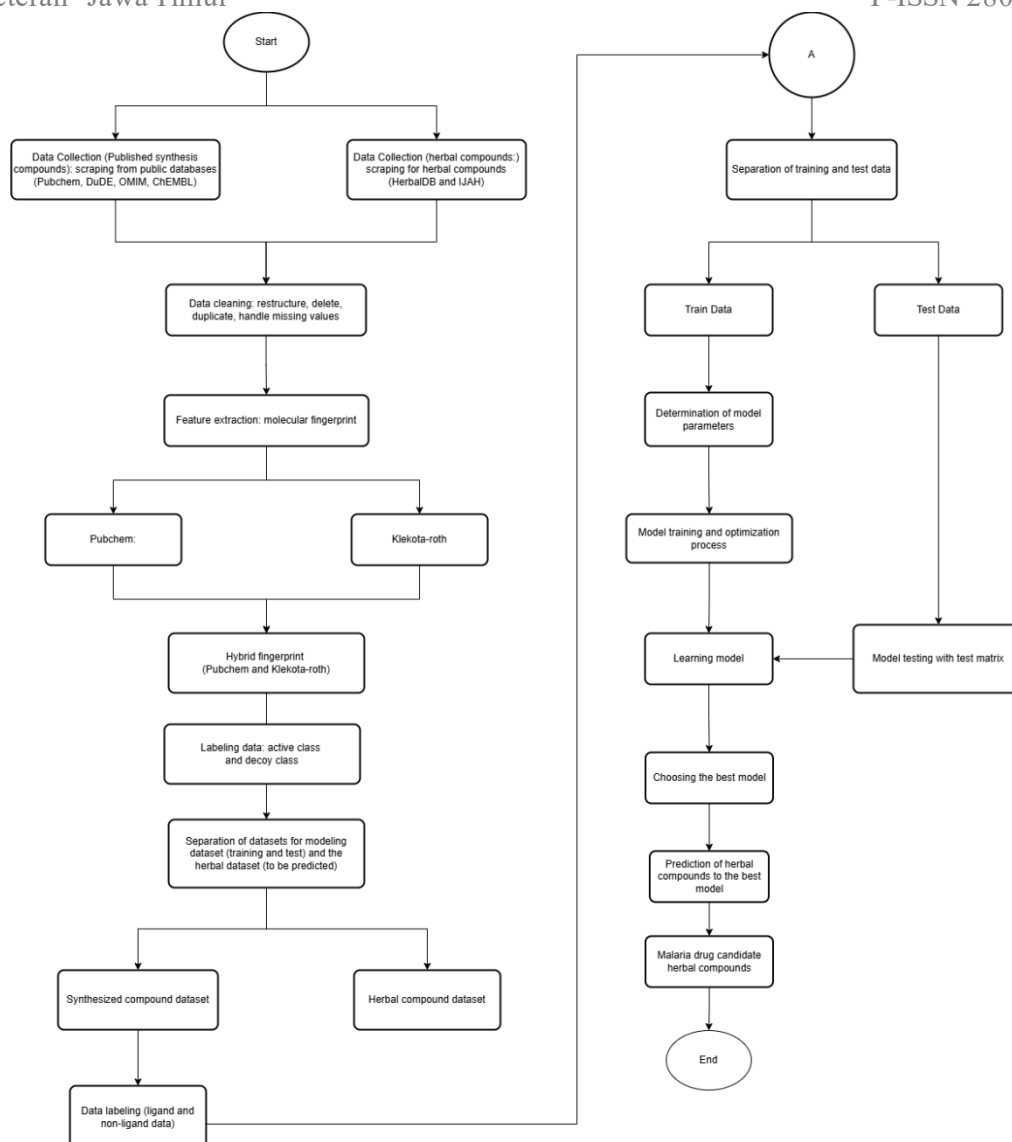
Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk membangun model klasifikasi berbasis supervised learning dengan memanfaatkan hybrid molecular fingerprint sebagai pendekatan efektif untuk memprediksi senyawa herbal yang berpotensi sebagai obat antimalaria. Untuk mencapai tujuan ini, penelitian ini akan dilaksanakan dalam empat tahap metodologis utama. Pertama, data tentang senyawa sintetis aktif dan tidak aktif dikumpulkan dari basis data OMIM [13], PubChem [14], dan DUDE [15], serta data senyawa herbal dari basis data HerbalDB dan IJAH (Indonesia Jamu-Herb) Analytics. Kedua, ekstraksi fitur dilakukan menggunakan hybrid molecular fingerprint (seperti PubChem dan Klekota-Roth) [16] untuk mewakili karakteristik kimia senyawa secara komprehensif. Ketiga, membangun model prediktif dengan pipeline Scikit-learn yang mencakup pelatihan algoritma



supervised learning [17] (seperti gradient boosting atau XGBoost) dan optimasi hiperparameter. Terakhir, validasi model dilakukan melalui metrik evaluasi (akurasi, AUC-ROC) dan simulasi docking untuk mengonfirmasi afinitas senyawa terpilih terhadap protein target. Melalui pendekatan terstruktur ini, penelitian diharapkan dapat memberikan kontribusi signifikan dalam mempercepat penemuan obat antimalaria yang lebih aman, terjangkau, dan berbasis sumber daya lokal.

II. METHOD

Penelitian ini bertujuan untuk memperoleh protein yang terkait dengan Malaria. Protein-protein ini berpotensi menjadi target protein patogen penyebab Malaria. Senyawa dari protein-protein ini akan diprediksi menggunakan senyawa herbal untuk mengatasi tantangan dalam pengobatan Malaria, seperti resistensi obat dan efek samping obat sintesis. Penelitian ini terdiri dari enam tahap utama, yaitu pengumpulan data, ekstraksi fitur senyawa, penandaan kelas dan pemisahan data, pemodelan dan pengujian, prediksi senyawa herbal, dan tahap akhir berupa publikasi ke jurnal nasional. Alur penelitian dapat dilihat pada **Gambar 1**. Setiap tahap ini mencakup proses yang lebih rinci terkait pelaksanaan penelitian. Berikut adalah penjelasan masing-masing tahap tersebut.



Gambar 1. Diagram Alir Prediksi Senyawa Herbal Antimalaria

Pengumpulan Data

Penelitian ini menggunakan data sekunder yang diperoleh dari basis data protein dan ligan melalui proses pengumpulan data yang terstruktur. Data senyawa aktif diperoleh dari basis data bioinformatika yang andal dan terbuka, termasuk PubChem [14], ChEMBL [19], DUDE [15], dan OMIM (Online Mendelian Inheritance in Man) [13], yang menyediakan informasi komprehensif tentang struktur kimia, aktivitas biologis, dan interaksi protein. Dari basis data OMIM, 272 gen locus unik awalnya diidentifikasi. Analisis selanjutnya menggunakan aplikasi STRING-DB [20] mengidentifikasi 12 kluster protein, di antaranya 12 protein target yang terkait dengan Malaria seperti TNF, IL10, TIRAP, dan CCR5, dipilih. Protein-protein ini dianalisis lebih lanjut untuk memperoleh data senyawa dalam format SMILES atau InChI, beserta Compound IDs (CIDs) dan nama IUPAC, untuk mewakili struktur kimia dan sifat ikatannya.

Selain senyawa sintetis, penelitian ini juga mengintegrasikan senyawa herbal sebagai objek studi. Senyawa herbal adalah senyawa alami yang diekstraksi dari tumbuhan asli Indonesia. Sebagai negara dengan keanekaragaman hayati tertinggi di dunia, terutama dalam kekayaan flora, Indonesia memiliki potensi besar untuk eksplorasi senyawa bioaktif. Setiap variasi flora memiliki profil senyawa herbal yang unik, yang secara empiris diketahui memiliki efektivitas tinggi dan efek samping minimal dibandingkan dengan senyawa sintetis [21]. Hal ini yang mendorong kami untuk mengumpulkan data senyawa herbal dari basis data khusus, seperti IJAH (Indonesia Jamu-Herb) Analytics dan HerbalDB 2.0 [22], menggunakan teknik web scraping. Basis data ini menyediakan informasi terstruktur tentang senyawa bioaktif dari tanaman obat tradisional Indonesia, yang berpotensi dikembangkan sebagai kandidat obat antimalaria. Dari total 6.756 data senyawa herbal yang berhasil dikumpulkan, proses penyaringan ketat dilakukan sehingga hanya 4.494 data yang memenuhi kriteria kelengkapan informasi, seperti memiliki Compound ID yang terverifikasi (identifikasi unik senyawa kimia). Contoh representatif data senyawa herbal yang telah disunting dapat dilihat pada **Tabel 1**.

Tabel 1. Contoh Hasil Scraping Senyawa Herbal

Senyawa	ID Knapsack	ID Metabolite	ID Pubchem
Monoterpene Myrcene	C00000853	M00000853	Myrcene
(+)-8(1)-Hydroxypinoresinol	C00007191	M00007191	3010930
(+)-alpha-Cyperone	C00012759	M00012759	273568

Dataset yang mencakup senyawa sintetis dan alami memungkinkan penyelidikan mendalam terhadap 4.494 senyawa herbal untuk diuji kompatibilitasnya dengan protein target malaria, seperti *Plasmodium falciparum* atau enzim kunci dalam siklus hidup parasit. Studi ini menggunakan pendekatan penandaan molekuler hibrid yang menggabungkan fitur kimia dari PubChem dan Klekota-Roth untuk mewakili karakteristik struktural senyawa secara multidimensi. Selain itu, teknik pembelajaran mesin berbasis Gradient Boosting diterapkan untuk membangun model klasifikasi yang mampu memprediksi potensi antimalaria senyawa herbal. Integrasi kedua jenis dataset (sintetis dan herbal) tidak hanya memperluas cakupan analisis, tetapi juga meningkatkan ketahanan model prediktif melalui variasi data yang lebih representatif. Dengan demikian, pendekatan ini diharapkan dapat mengidentifikasi senyawa kandidat antimalaria baru yang lebih aman, terjangkau, dan berbasis sumber daya lokal, sekaligus mengurangi ketergantungan pada obat sintetis yang berisiko resistensi.

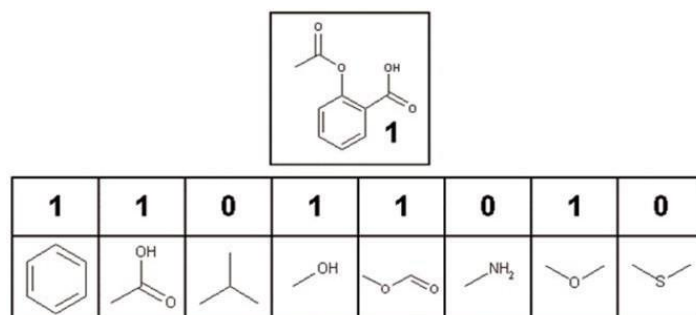
2.2 Praproses Data

Setelah data dikumpulkan, tahap prapemrosesan data atau pembersihan data dilakukan untuk memastikan kualitas data yang akan digunakan [23]. Proses ini meliputi restrukturisasi data untuk menyesuaikan format agar konsisten (misalnya, standarisasi representasi molekul seperti SMILES (Simplified Molecular Input Line Entry System) / InChI (International Chemical Identifier), menghapus data duplikat untuk menghindari redundansi, dan menangani data yang hilang dengan mengisi nilai kosong menggunakan metode statistik atau menghapus data jika tidak dapat diperbaiki. Langkah-langkah ini sangat penting untuk memastikan keandalan dataset sebelum analisis lebih lanjut menggunakan molecular fingerprinting dan machine learning.



2.3 Ekstraksi Fitur Komposit

Data yang diperlukan untuk diekstraksi adalah kode SMILES dari suatu senyawa, yang merupakan string yang mewakili ikatan dalam struktur senyawa tersebut. Ekstraksi fitur dilakukan menggunakan metode sidik jari molekuler, dengan dua jenis sidik jari yang digunakan, yaitu PubChem, yang menghasilkan 881 fitur, dan Klekota-Roth, yang menghasilkan 4860 fitur [18]. Proses ekstraksi fitur melibatkan pencatatan struktur molekul yang mengandung ikatan tertentu. Misalnya, jika suatu molekul memiliki ikatan cincin, maka substruktur tersebut dikodekan sebagai 1. Sebaliknya, jika suatu senyawa tidak memiliki ikatan kimia yang sesuai dengan fitur kunci, maka substruktur tersebut dikodekan sebagai 0.



Gambar 2. Fitur Ekstraksi menggunakan Molecular Fingerprint

Setelah senyawa diekstraksi menggunakan dua jenis sidik jari, langkah berikutnya adalah menggabungkan kedua fingerprints tersebut yang dikenal sebagai proses hybrid. Dalam situasi di mana terdapat banyak fitur, mungkin ada fitur di mana semua data memiliki nilai. Oleh karena itu, perlu memilih fitur yang relevan dari semua fitur yang ada. Jika ada fitur yang memiliki nilai 0 di semua data senyawa, fitur tersebut akan dihapus karena dianggap tidak signifikan. Setelah proses pemilihan fitur selesai, data akan dipisahkan antara senyawa yang terkait dengan Malaria dan senyawa herbal. Data yang akan dimodelkan menggunakan salah satu metode supervised learning, yaitu Gradient Boosting. Hasil pemodelan akan diproses lebih lanjut, sementara data herbal akan disimpan terlebih dahulu untuk diprediksi setelah model terbentuk.

2.4 Pelabelan Kelas and Pemisahan Senyawa

Sebelum digunakan dalam pemodelan, data terlebih dahulu diberi label untuk membedakan antara dua kategori utama, yaitu data ligan dan data non-ligan. Data ligan terdiri dari senyawa yang dapat berikatan dengan target biologis spesifik, seperti protein atau enzim, dan berperan dalam interaksi molekuler yang dapat memengaruhi aktivitas biologis. Sementara itu, data non-ligan terdiri dari senyawa yang tidak menunjukkan kemampuan signifikan untuk berinteraksi dengan target biologis.

Setelah proses pelabelan, data dibagi menjadi dua bagian utama: data latih dan data uji. Data latih digunakan untuk membangun dan melatih algoritma machine learning agar dapat mengenali pola yang membedakan ligan dari non-ligan. Model yang dihasilkan dari tahap pelatihan ini kemudian diuji menggunakan data uji, yang berfungsi untuk mengevaluasi kinerja model dalam mengklasifikasikan senyawa dengan benar. Proses ini memastikan bahwa model memiliki akurasi dan generalisasi yang baik sebelum diterapkan dalam pencarian senyawa obat potensial.



2.5 Pemodelan dan Pengujian

Algoritma ini mempelajari data dengan mengandalkan data latih sebagai sumber utama. Dalam pendekatan supervised learning, metode yang digunakan adalah klasifikasi, dengan algoritma Gradient Boosting atau XGBoost [24]. Gradient Boosting adalah teknik ensemble learning yang populer yang digunakan untuk tugas-tugas regresi dan klasifikasi [25]. Teknik ini membangun model prediktif yang kuat dengan menambahkan secara berurutan pembelajar yang lemah ke dalam ansambel, dengan setiap pembelajar baru mengoreksi kesalahan yang dibuat oleh pembelajar sebelumnya.

XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting) adalah algoritma machine learning yang termasuk dalam kategori pembelajaran ansambel berbasis boosting. Algoritma ini merupakan pengembangan dari Gradient Boosting Machine (GBM) dengan optimasi regularisasi yang ditingkatkan dan efisiensi komputasi [24]. Algoritma XGBoost beroperasi melalui proses berulang: secara berurutan membangun pohon keputusan di mana setiap model berikutnya menargetkan kesalahan residual dari pendahulunya. Secara khusus, algoritma ini menggunakan gradient descent untuk meminimalkan kerugian dan mengintegrasikan istilah regularisasi untuk mengontrol kompleksitas model dan mencegah overfitting. Persamaan 1 menyajikan model matematis XGBoost.

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in F \quad (1)$$

Dimana:

- $\hat{y}_i$ = nilai prediksi untuk sampel ke-i
- K = jumlah pohon
- f_k = fungsi prediksi pohon ke-k
- F = ruang dari kemungkinan semua fungsi pohon keputusan

Algoritma pengklasifikasi XGBoost menggunakan beberapa parameter kunci untuk meningkatkan kinerja prediktifnya dalam menganalisis senyawa herbal. Di antara parameter yang paling penting adalah `n_estimator`, yang menentukan jumlah pohon (boosting rounds) dalam model; `learning_rate` (η), mengontrol penyusutan ukuran langkah selama pelatihan untuk mencegah overfitting; `max_depth`, menentukan kedalaman maksimum dari setiap pohon keputusan; `subsample`, menentukan fraksi sampel latih yang digunakan untuk setiap pohon; dan `colsample_bytree`, yang mengatur fraksi fitur yang digunakan untuk membangun setiap pohon. Parameter-parameter ini bekerja secara sinergis untuk mengoptimalkan akurasi dan kemampuan generalisasi model dalam memprediksi sifat-sifat senyawa herbal.

Gradient Boosting

Gradient Boosting memulai prosesnya dengan membangun model ensemble menggunakan model dasar yang sederhana, seperti prediktor rata-rata. [25]

1. Inisialisasi model awal F_0 dengan nilai rata-rata target ($\text{mean}(y)$):
 $F_0 = \text{mean}(y) \quad (2)$
2. Perhitungan residual: residual (perbedaan antara nilai aktual dan prediksi) dihitung untuk setiap data latih. Residual ini mewakili kesalahan yang belum dapat ditangkap oleh model saat ini. Residual pertama adalah r_1 yang dihitung sebagai:
 $r_1 = y - \text{mean}(y) \quad (3)$

3. Pelatihan *weak learning* : suatu model yang lemah, biasanya berupa *decision stump* atau pohon regresi dengan kedalaman tertentu (biasanya 8-32 simpul terminal), dilatih untuk meminimalkan residu. Model ini memprediksi residual r_1 menggunakan fitur x .
4. Pembaruan prediksi ensemble, prediksi model yang lemah (γ_1) tidak langsung ditambahkan ke model awal, tetapi dikalikan dengan laju pembelajaran (v , nilai antara 0 dan 1) untuk menghindari overfitting. Prediksi ensemble diperbarui menjadi:
$$F_1 = F_0 + v \cdot \gamma_1 \quad (4)$$
5. Proses iterasi, residual baru (r_2) dihitung berdasarkan prediksi terbaru (F_1), dan proses diulangi dengan melatih model lemah baru pada residual. Langkah ini terus berlanjut hingga performa model tidak lagi mengalami peningkatan yang signifikan.

Pada tahap awal, algoritma diberikan parameter yang relevan. Selama proses pelatihan, optimasi parameter dilakukan melalui tuning berulang. Hasil tuning dievaluasi berdasarkan matriks uji dari data uji, yang mencakup metrik standar klasifikasi seperti akurasi, spesifisitas, sensitivitas, dan skor ROC. Setelah mendapatkan parameter terbaik untuk setiap model, hasil pengujian dari XGBoost dibandingkan dengan metode *supervised learning* lainnya seperti *random forest*, *support vector machine*, dan *logistic regression*. Model dengan kinerja terbaik kemudian dipilih sebagai model prediksi untuk senyawa herbal.

2.6 Analisis Prediksi Senyawa Herbal

Analisis prediksi senyawa herbal dilakukan menggunakan model prediksi optimal yang telah dikembangkan pada tahap sebelumnya. Pada tahap awal, senyawa herbal belum memiliki label, sehingga hubungan potensialnya dengan protein target Malaria belum teridentifikasi. Model yang dibangun dari data terkorelasi memungkinkan senyawa tak terklasifikasi untuk mencocokkan kemiripannya saat diprediksi. Hasil prediksi kemudian akan disaring untuk memilih 10 senyawa herbal teratas dengan skor probabilitas tertinggi. Nilai probabilitas tinggi menunjukkan bahwa senyawa-senyawa ini berpotensi menjadi kandidat efektif untuk terapi antimalaria.

III. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil penelitian yang menilai kandidat obat antimalaria yang potensial. Bagian berikut ini menguraikan metodologi untuk membuat set data, hasil dari upaya pemodelan, dan identifikasi senyawa yang menunjukkan nilai prediksi yang menjanjikan, yang menandainya sebagai kandidat obat antimalaria yang potensial.

3.1 Dataset

Sumber data dalam penelitian ini berasal dari situs resmi internasional untuk mendapatkan data senyawa yang akan diteliti. Senyawa-senyawa tersebut berasal dari database protein dan ligan, yaitu OMIM, PubChem, DUDE, dan ChEMBL. Selain menggunakan senyawa sintetik, penelitian ini menggunakan senyawa herbal sebagai prediksi. Senyawa herbal ini diperoleh dari HerbalDB dan IJAH (Indonesia Jamu-Herb) Analytics. Dari pengumpulan data senyawa tersebut, didapatkan data lokus dari penyakit Malaria. Dari data lokus tersebut, terdapat 12 protein yang berinteraksi dengan patogen penyebab malaria. Protein tersebut adalah TNF, IL10[26], TIRAP, ACKR1, CD36, TLR4[26], IFNGR1, CCR5, TLR2, IL4, IFNG, CCL2. Dari protein-protein ini, data senyawa yang terdiri dari CID, nama IUPAC, dan SMILE kemudian dicari. Terdapat tujuh protein yang memiliki data senyawa yang lengkap, termasuk CCL2, CCR5, CD36, IL4, TLR2, TLR4, dan TNF. Prosedur untuk memperoleh informasi ini menghasilkan identifikasi 109 senyawa yang terkait dengan tujuh protein target spesifik. Rincian senyawa-senyawa aktif dalam protein dapat dilihat pada **Tabel 2**.



Tabel 2. Jumlah Ligand Aktif untuk Protein Target Vital

Protein Code	Nama Protein	Jumlah Ligand
TNF	Tumor Necrosis Factor	47
CCL2	Chemokine (C-C motif) Ligand 2	5
TLR2	Toll-Like Receptor 2	2
CCR5	C-C Chemokine Receptor Type 5	31
CD36	Cluster of Differentiation 36	5
IL4	Interleukin-4	1
TLR4	Toll-Like Receptor 4	18

3.2 Ekstraksi Fitur

Berdasarkan proses ekstraksi fitur komposit menggunakan *fingerprint* molekuler hibrida, hasil dari ekstraksi fitur melalui teknik PubChem menghasilkan matriks data (baris) dengan fitur (kolom) yang mengukur 109×881 . Ekstraksi fitur menggunakan *fingerprint* Klekota-Roth menghasilkan dataset biner (ukuran: 109×4860). Setiap nilai biner mengkodekan keberadaan fitur (1) atau ketidakhadirannya (0), sesuai dengan status deteksi ikatan kimia atau unsur dalam senyawa.

3.3 Pemodelan dan Pengujian

Dalam penelitian ini, kami mengembangkan model klasifikasi senyawa antimalaria dengan menggunakan pendekatan *machine learning* berbasis Python. Data pelatihan diperoleh dari ekstraksi fitur molekuler *fingerprint* menggunakan metode fingerprint Klekota-Roth dan PubChem, sedangkan data pengujian menggunakan dataset senyawa herbal sebagai target prediksi. Sebelum pelatihan model, kami membagi data pelatihan secara proporsional dengan rasio 80:20, di mana 80% data digunakan untuk pelatihan model dan 20% untuk validasi internal, guna memastikan kualitas dan generalisasi model. Variabel independen (X) dalam penelitian ini terdiri dari fitur-fitur *fingerprint* molekuler, sedangkan variabel dependen (Y) merupakan label kelas yang terdiri dari ligan dan non-ligan, dengan distribusi kelas yang seimbang. Model dikembangkan menggunakan algoritma XGBoost Classifier dengan parameter awal yang dioptimalkan, termasuk learning rate sebesar 0.01 untuk mengontrol kontribusi setiap pohon, 100 estimator ($n_estimators$) sebagai jumlah pohon dalam ensemble, dan kedalaman maksimum (max_depth) 3 untuk mencegah overfitting. Implementasi keseluruhan proses dilakukan menggunakan library scikit-learn untuk pembagian dataset dan XGBoost untuk pembangunan model klasifikasi, dengan tujuan akhir menghasilkan model prediktif yang akurat untuk senyawa antimalaria, khususnya dari bahan herbal.

Perbandingan Model Klasifikasi Lainnya

Keberhasilan model klasifikasi secara kritis memengaruhi fase pengembangan akhir. Untuk mengoptimalkan kinerja, kami mengevaluasi beberapa model yang dibangun menggunakan dataset kombinasi awal dengan bantuan Python dan library-machine learning seperti scikit-learn dan XGBoost. Dataset kombinasi awal ini dipilih berdasarkan hasil sebelumnya, yang menunjukkan kinerja model yang lebih baik dibandingkan dengan dua dataset lainnya. Kami membandingkan model *XGB-Initial Combination* terhadap berbagai teknik klasifikasi, termasuk *Logistic Regression*, *K-Nearest Neighbors*, *Support Vector Machine*, dan *Multilayer Perceptron* Furthermore. Selanjutnya, kami membandingkan hasilnya dengan metode ensemble seperti Hutan Acak dan Penguatan Gradien.

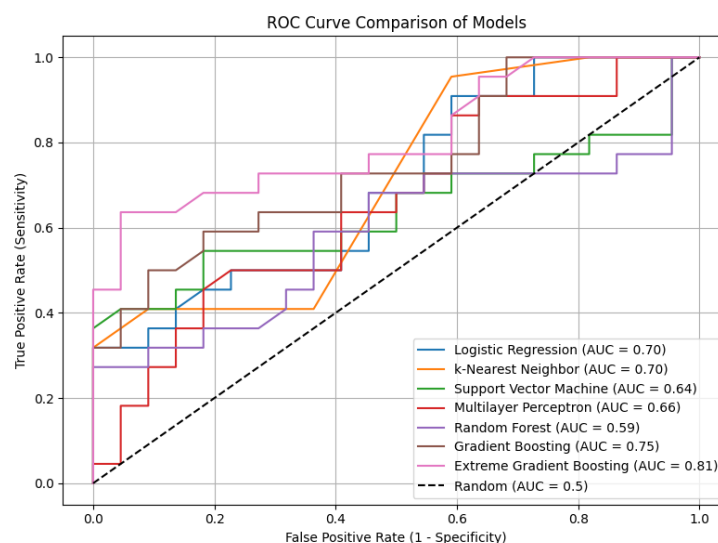


Semua metrik kinerja model komparatif akurasi, sensitivitas, recall, dan skor ROC dijelaskan secara rinci dalam **Tabel 3**.

Tabel 3. Evaluasi Kinerja Model Machine Learning: Akurasi, *Recall*, Sensitivitas, dan Kurva ROC

Classification Models	Accuracy	Recall	Sensitivity	ROC Score
k-Nearest Neighbors	0.52	0.52	0.41	0.70
Logistic Regression	0.55	0.55	0.50	0.70
Multilayer Perceptron	0.57	0.57	0.50	0.66
Random Forest	0.57	0.57	0.46	0.59
Support Vector Machine	0.68	0.68	0.56	0.64
Gradient Boosting	0.61	0.61	0.64	0.76
Extreme Gradient Boosting	0.75	0.75	0.68	0.81

Berdasarkan Tabel 3, *Extreme Gradient Boosting* menunjukkan kinerja terbaik dibandingkan dengan model klasifikasi lainnya dalam empat metrik evaluasi, yaitu akurasi, recall, sensitivitas, dan skor ROC. Model ini memperoleh akurasi sebesar 0,75, recall 0,75, sensitivitas 0,68, dan skor ROC 0,81, semuanya merupakan nilai tertinggi di antara model-model lainnya. Ini menunjukkan bahwa XGB mampu menangkap pola data dengan lebih akurat dan konsisten. Sebaliknya, model dengan kinerja terendah adalah *k-Nearest Neighbors*, yang hanya mencapai akurasi dan recall sebesar 0.52, sensitivitas sebesar 0.41, dan ROC sebesar 0.70. Model *Logistic Regression* juga menunjukkan kinerja yang buruk, dengan skor akurasi dan recall sebesar 0.55, sensitivitas sebesar 0.50, dan ROC sebesar 0.70, sehingga kurang cocok untuk digunakan pada data biner berdimensi tinggi. Sementara itu, *Support Vector Machine* (SVM) menunjukkan performa yang cukup baik, dengan akurasi 0.68, recall 0.68, sensitivitas 0.56, dan ROC 0.64, membuatnya menjadi model yang masih kompetitif, meskipun tidak lebih baik dari XGB. Secara keseluruhan, XGB adalah model yang paling unggul untuk klasifikasi dalam konteks data ini.



Gambar 1. Grafik ROC untuk Setiap Metode Klasifikasi

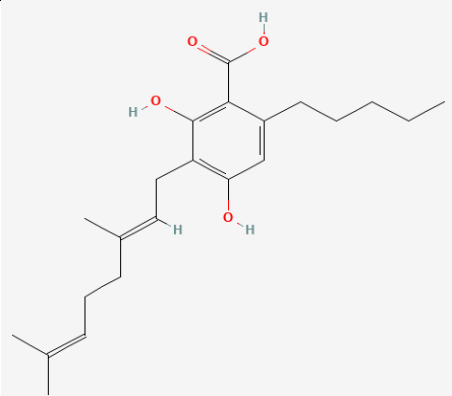
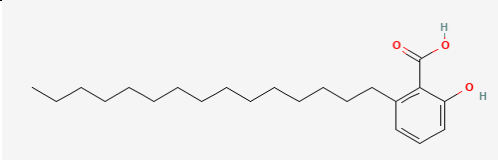
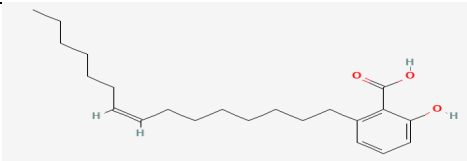
Gambar 3. menyajikan perbandingan kinerja berbagai algoritma klasifikasi yang digunakan untuk mengidentifikasi senyawa herbal yang berpotensi sebagai agen antimalaria berdasarkan sidik



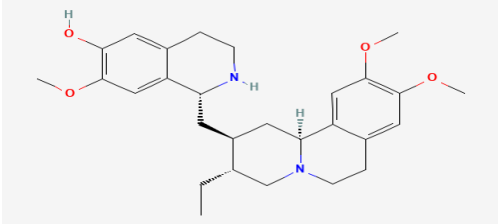
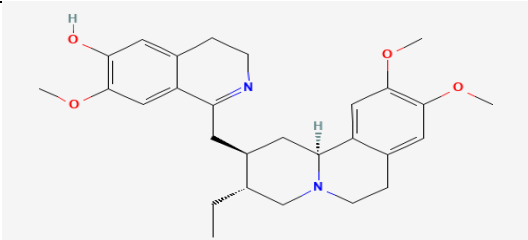
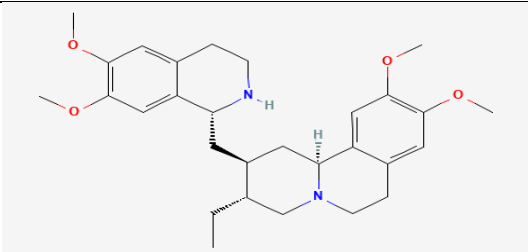
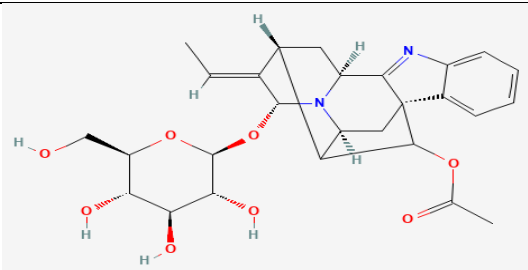
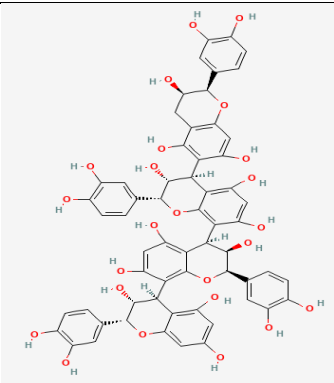
jari molekuler hibrida. Kurva ROC menunjukkan bagaimana tingkat positif sejati, juga dikenal sebagai sensitivitas, berkorelasi dengan tingkat positif palsu, yang dihitung sebagai satu dikurangi spesifisitas, pada berbagai batas klasifikasi, sambil menilai kapasitas masing-masing model untuk membedakan antara kelas ligan dan non-ligan.

Grafik tersebut menunjukkan bahwa model *Extreme Gradient Boosting* (XGBoost) memberikan kinerja yang unggul dengan nilai *Area Under the Curve* (AUC) sebesar 0,81, yang menunjukkan akurasi diskriminatif yang sangat baik. Model ini secara konsisten menunjukkan sensitivitas yang lebih tinggi dalam prediksi positif dibandingkan dengan model lainnya. Sementara itu, model *Gradient Boosting*, *Logistic Regression*, dan *k-Nearest Neighbors* juga menunjukkan kinerja yang kompetitif dengan nilai AUC masing-masing 0,75, 0,70, dan 0,70, yang mencerminkan kemampuan klasifikasi yang cukup baik. *Multilayer Perceptron* (AUC = 0.66) dan *Support Vector Machine* (AUC = 0.64) menunjukkan performa yang cukup baik, sementara *Random Forest* mencatat nilai AUC terendah yaitu 0.59 - mendekati performa model acak (AUC = 0.5) dan mengindikasikan efektivitas klasifikasi yang kurang baik pada set data ini. Secara keseluruhan, hasil ini menetapkan XGBoost sebagai model yang paling dapat diandalkan untuk klasifikasi senyawa antimalaria dalam penelitian ini, dan oleh karena itu dipilih sebagai model utama untuk proses seleksi dan prediksi selanjutnya.

Tabel 4. Hasil Prediksi Senyawa Herbal

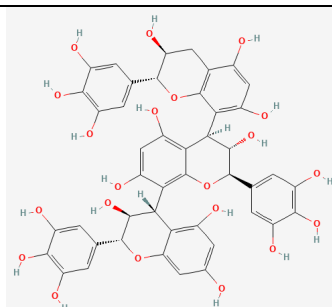
Senyawa	Compound 2D Structures	Probability Values
Cannabigerolic acid		0,786
Anacardic acid		0,786
Ginkgolic Acid		0,786



Cephaeline		0,786
Psychotrine		0,777
Emetine		0,777
Raucaffricine		0,776
[Epicatechin-(4beta->8)]2-epicatechin-(4beta->6)-catechin		0,765



[Gallocatechin-(4 α -
>8)]₂-gallocatechin



0,765

Berdasarkan **Tabel 4**, menunjukkan bahwa terdapat senyawa baru yang diindikasikan dapat menjadi kandidat obat herbal antimalaria, antara lain senyawa Cannabigerolic acid dengan probabilitas 0,786 yang terdapat pada tanaman *Cannabis sativa* (ganja), yaitu senyawa yang berpotensi sebagai antiinflamasi dan pelindung saraf. Namun, belum banyak penelitian langsung terhadap nyamuk, tetapi beberapa cannabinoid menunjukkan efek pengusir serangga. Senyawa kedua, Anacardic Acid, ditemukan pada kacang mete (*Anacardium occidentale*) dan kulitnya. Senyawa ini bermanfaat sebagai antibakteri, antioksidan, dan anti tumor. Anacardic acid menunjukkan efek pengusir dan larvasida terhadap nyamuk *Aedes aegypti* (penyebab demam berdarah) [27]. Senyawa ketiga adalah Ginkgolic Acid, yang ditemukan dalam daun *Ginkgo biloba*. Senyawa ini bermanfaat sebagai antibakteri dan antijamur; namun, dapat menjadi racun dalam dosis tinggi. Beberapa penelitian tentang senyawa ini menunjukkan potensi sebagai larvasida nyamuk[28], tetapi belum banyak dikembangkan. Selanjutnya, ada Cepheline & Emetine yang ditemukan pada akar *Psychotria ipecacuanha* (tanaman ipecac). Senyawa ini berguna sebagai antiprotozoa (misalnya amoebiasis) dan memiliki efek toksik pada larva nyamuk.

Senyawa kelima, Psychotrine dengan probabilitas 0,777, diindikasikan dapat menjadi kandidat obat antimalaria. Psychotrine ditemukan pada tanaman *Psychotria ipecacuanha* dan beberapa spesies *Psychotria* lainnya. Berikutnya ada senyawa Raucaffricine dengan probabilitas 0,776. Senyawa ini ditemukan pada *Rauvolfia serpentina* (ular tedung). Senyawa ini bertindak sebagai prekursor alkaloid antihipertensi (misalnya reserpin) dan ekstraknya dilaporkan memiliki efek insektisida [29]. Selanjutnya, senyawa [Epicatechin-(4 β $\rightarrow$ 8)]₂-Epicatechin-(4 β $\rightarrow$ 6)-Catechin dan [Gallocatechin-(4 α $\rightarrow$ 8)]₂-Gallocatechin, mengandung senyawa polifenol (proanthocyanidin) dengan probabilitas 0,765. Senyawa ini ditemukan dalam teh hijau, anggur, kakao, dan beberapa buah beri. Bermanfaat sebagai antioksidan, pelindung jantung, dan antikanker. Beberapa polifenol (seperti katekin) telah dilaporkan memiliki efek pengusir atau menghambat perkembangan jentik nyamuk, tetapi senyawa spesifik ini belum banyak diteliti[30].

Berdasarkan analisis struktur senyawa herbal, Anacardic Acid dari jambu mete dan Emetine dari *Psychotria ipecacuanha* merupakan kandidat yang paling potensial sebagai antimalaria karena terbukti memiliki efek larvasida dan antiprotozoa yang didukung oleh data probabilitas yang tinggi (0,786-0,777). Senyawa lain seperti Cannabigerolic Acid (ganja) dan Ginkgolic Acid (*Ginkgo biloba*) juga menjanjikan tetapi memerlukan penelitian lebih lanjut mengenai toksisitas dan mekanisme kerjanya terhadap nyamuk *Anopheles*. Sementara polifenol dari teh dan kakao berpotensi sebagai pengusir nyamuk, meskipun efektivitasnya masih memerlukan pengujian yang lebih mendalam.

IV. KESIMPULAN

Berdasarkan temuan penelitian, metode *Extreme Gradient Boosting* (XGBoost) merupakan model terbaik untuk mengembangkan model prediksi senyawa protein Malaria dengan kandidat senyawa obat herbal. Model ini memiliki akurasi tertinggi dibandingkan metode klasifikasi lainnya, yaitu 0.75, dalam mengidentifikasi senyawa herbal potensial sebagai kandidat obat antimalaria menggunakan pendekatan *hybrid molecular fingerprint*. Hasil analisis menunjukkan bahwa senyawa Anacardic Acid dari jambu mete (*Anacardium occidentale*) dan Emetine dari akar ipecac (*Psychotria ipecacuanha*) merupakan kandidat paling menjanjikan dengan skor probabilitas lebih dari 0.70. Kedua senyawa ini dipilih berdasarkan aktivitas biologis yang terdokumentasi, termasuk sifat larvasida terhadap nyamuk dan efek antiprotozoa yang relevan dengan mekanisme antimalaria. Penelitian ini juga mengidentifikasi beberapa senyawa lain seperti Cannabigerolic acid dari *Cannabis sativa* dan Ginkgolic acid dari *Ginkgo biloba* yang menunjukkan potensi, meskipun diperlukan penelitian lebih lanjut terkait profil toksisitas dan mekanisme kerjanya secara spesifik terhadap *Plasmodium*. Temuan ini memperkuat posisi XGBoost sebagai alat prediktif yang efektif dalam *screening* senyawa bioaktif dari bahan alam, khususnya untuk Malaria. Namun, penelitian lebih lanjut masih diperlukan untuk memvalidasi hasil ini, terutama uji *in vivo* terhadap parasit Malaria, optimasi formulasi, serta evaluasi toksisitas untuk memastikan keamanan penggunaannya. Implikasi penelitian ini tidak hanya penting untuk pengembangan obat herbal antimalaria, tetapi juga membuka peluang pemanfaatan kecerdasan buatan dalam eksplorasi obat berbasis alam.

UCAPAN TERIMA KASIH

Penulis menghaturkan terima kasih kepada Bapak Mohammad Hamim Zajuli Al Faroby, S.Si., M.Mat. selaku dosen pembimbing kami yang membantu dalam penulisan artikel dan pihak Telkom University Surabaya sebagai universitas yang mendukung penulisan artikel penelitian yang dilakukan oleh penulis, serta semua pihak lainnya yang membantu.

REFERENSI

- [1] WHO, “World Malaria Report 2021,” 2021.
- [2] S. Madayanti, M. Raharjo, and H. Purwanto, “Faktor Risiko Yang Mempengaruhi Kejadian Malaria di Wilayah Distrik Jayapura Selatan Kota Jayapura,” *Jurnal Kesehatan Lingkungan Indonesia*, vol. 21, no. 3, pp. 358–365, Oct. 2022, doi: 10.14710/jkli.21.3.358-365.
- [3] K. J. Wicht, S. Mok, and D. A. Fidock, “Molecular Mechanisms of Drug Resistance in *Plasmodium falciparum* Malaria,” *Annu Rev Microbiol*, vol. 74, no. 1, pp. 431–454, Sep. 2020, doi: 10.1146/annurev-micro-020518-115546.
- [4] L. Zhu *et al.*, “Artemisinin resistance in the malaria parasite, *Plasmodium falciparum*, originates from its initial transcriptional response,” *Commun Biol*, vol. 5, no. 1, p. 274, 2022.
- [5] Y. Y. S. Rahayu, T. Araki, and D. Rosleine, “Factors affecting the use of herbal medicines in the universal health coverage system in Indonesia,” *J Ethnopharmacol*, vol. 260, p. 112974, 2020.
- [6] D. W. Purnomo, M. Siregar, J. R. Witono, and D. Usmadi, “Rencana 10 Tahun (2020-2030) Pengembangan Kebun Raya di Indonesia,” *Warta Kebun Raya*, vol. 18, no. 2, pp. 1–16, 2020.
- [7] M. Budiarti *et al.*, “The use of antimalarial plants as traditional treatment in Papua Island, Indonesia,” *Heliyon*, vol. 6, no. 12, p. e05562, Dec. 2020, doi: 10.1016/J.HELIYON.2020.E05562/ATTACHMENT/7EABD3A3-FC82-4EB9-A5C0-43408FC8B99D/MMC1.DOCX.
- [8] A. S. Nugraha *et al.*, “Isolation of Antimalarial Agents From Indonesian Medicinal Plants: *Swietenia mahagoni* and *Pluchea indica*,” *Nat Prod Commun*, vol. 17, no. 1, Jan. 2022, doi: 10.1177/1934578X211068926/SUPPL_FILE/SJ-DOCX-1-NPX-10.1177_1934578X211068926.DOCX.



- [9] T. B. Kimber, Y. Chen, and A. Volkamer, "Deep learning in virtual screening: recent applications and developments," *Int J Mol Sci*, vol. 22, no. 9, p. 4435, 2021.
- [10] I. A. Guedes *et al.*, "New machine learning and physics-based scoring functions for drug discovery," *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/S41598-021-82410-1,.
- [11] S. Amiroch, M. I. Irawan, I. Mukhlash, M. H. Z. Al Faroby, and C. A. Nidom, "Machine Learning for the Prediction of Antiviral Compounds Targeting Avian Influenza A/H9N2 Viral Proteins," *Symmetry (Basel)*, vol. 14, no. 6, p. 1114, May 2022, doi: 10.3390/sym14061114.
- [12] S. Amiroch, M. H. Z. Al Faroby, M. I. Irawan, I. Mukhlash, and A. C. Nidhom, "Analysis of protein-protein interaction to obtain significant protein in influenza virus type A/H9N2," 2022, p. 020021. doi: 10.1063/5.0083336.
- [13] J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "OMIM.org: leveraging knowledge across phenotype-gene relationships," *Nucleic Acids Res*, vol. 47, no. D1, pp. D1038–D1043, 2019.
- [14] S. Kim *et al.*, "PubChem in 2021: new data content and improved web interfaces," *Nucleic Acids Res*, vol. 49, no. D1, pp. D1388–D1395, Jan. 2021, doi: 10.1093/nar/gkaa971.
- [15] H. Kuswanto, R. Y. Nurhidayah, and H. Ohwada, "Comparison of Feature Selection Methods to Classify Inhibitors in DUD-E Database," *Procedia Comput Sci*, vol. 144, pp. 194–202, 2018, doi: <https://doi.org/10.1016/j.procs.2018.10.519>.
- [16] J. Adamczyk and P. Ludynia, "Scikit-fingerprints: Easy and efficient computation of molecular fingerprints in Python," *SoftwareX*, vol. 28, p. 101944, Dec. 2024, doi: 10.1016/J.SOFTX.2024.101944.
- [17] M. N. Fahmi, "Implementasi Mechine Learning menggunakan Python Library : Scikit-Learn (Supervised dan Unsupervised Learning)," *Sains Data Jurnal Studi Matematika dan Teknologi*, vol. 1, no. 2, pp. 87–96, Dec. 2023, doi: 10.52620/sainsdata.v1i2.31.
- [18] M. H. Z. AlFaroby, H. N. Fadhillah, S. Amiroch, and R. S. Hidayat, "XGB-Hybrid Fingerprint Classification Model for Virtual Screening of Meningitis Drug Compounds Candidate," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Jun. 2022, doi: 10.22219/kinetik.v7i2.1424.
- [19] F. M. I. Hunter, A. P. Bento, N. Bosc, A. Gaulton, A. Hersey, and A. R. Leach, "Drug Safety Data Curation and Modeling in ChEMBL: Boxed Warnings and Withdrawn Drugs," *Chem Res Toxicol*, vol. 34, no. 2, pp. 385–395, Feb. 2021, doi: 10.1021/acs.chemrestox.0c00296.
- [20] D. Szklarczyk *et al.*, "The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Res*, vol. 49, no. D1, pp. D605–D612, Jan. 2021, doi: 10.1093/nar/gkaa1074.
- [21] Y. Y. S. Rahayu, T. Araki, and D. Rosleine, "Factors affecting the use of herbal medicines in the universal health coverage system in Indonesia," *J Ethnopharmacol*, vol. 260, p. 112974, Oct. 2020, doi: 10.1016/j.jep.2020.112974.
- [22] W. Ananta Kusuma, "Bioinformatics Approaches to Natural Product Discovery," *BIO Web Conf*, vol. 41, p. 02004, Dec. 2021, doi: 10.1051/bioconf/20214102004.
- [23] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/J.GLTP.2022.04.020.
- [24] J. Zheng *et al.*, "Metabolic syndrome prediction model using Bayesian optimization and XGBoost based on traditional Chinese medicine features," *Heliyon*, vol. 9, no. 12, 2023.
- [25] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, Feb. 2023, doi: 10.1016/J.JKSUCI.2023.01.014.
- [26] A. F. Alexander, I. Kelsey, H. Forbes, and K. Miller-Jensen, "Single-cell secretion analysis reveals a dual role for IL-10 in restraining and resolving the TLR4-induced inflammatory response," *Cell Rep*, vol. 36, no. 12, p. 109728, Sep. 2021, doi: 10.1016/J.CELREP.2021.109728.



- [27] W. F. Zuharah, A. Yousaf, K. L. Ooi, and S. F. Sulaiman, “Larvicidal activities of family Anacardiaceae on *Aedes* mosquitoes (Diptera: Culicidae) and identification of phenolic compounds,” *J King Saud Univ Sci*, vol. 33, no. 5, p. 101471, Jul. 2021, doi: 10.1016/J.JKSUS.2021.101471.
- [28] “Larvicidal Activity of Extracts of *Ginkgo biloba* Exocarp for Three Different Strains of *Culex pipiens pallens*.” Accessed: May 29, 2025. [Online]. Available: https://www.researchgate.net/publication/7159877_Larvicidal_Activity_of_Extracts_of_Ginkgo_biloba_Exocarp_for_Three_Different_Strains_of_Culex_pipiens_pallensI
- [29] D. Das and G. Chandra, “Mosquito larvicidal activity of *Rauvolfia serpentina* L. seeds against *Culex quinquefasciatus* Say,” *Asian Pac J Trop Med*, vol. 5, no. 1, pp. 42–45, Jan. 2012, doi: 10.1016/S1995-7645(11)60243-5.
- [30] S. Sharma, A. Verma, and N. Srivastava, “A review on medicinal plants having mosquito repellents activity,” ~ 82 ~ *Journal of Pharmacognosy and Phytochemistry*, vol. 13, no. 3, 2024, doi: 10.22271/phyto.2024.v13.i3b.14944.
- [1]