Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) p-ISSN: 2355-7699
Vol. 12, No. 4, Agustus 2025, hlm. 943- 950 e-ISSN: 2528-6579

PERBANDINGAN ANN, RANDOM FOREST, DAN XGBOOST DALAM KLASIFIKASI ANTIBIOTIK DENGAN PENERAPAN METODE SAMPLING

Edy Saputra Rusdi[*1], A. Muh. Amil Siddik[2], Naimah Aris[3], Muhammad Ardiansyah Asrifah[4], Nur Hilal A. Syahrir[5], Aidawayati Rangkuti[6], Wahyudi Rusdi[7]

[1,2,3,4,6]Universitas Hasanuddin, Makassar, [5]Universitas Sulawesi Barat, Majene,
[7]IAIN Sultan Amai Gorontalo, Gorontalo
Email: [1]edy_saputra@sci.unhas.ac.id, [2]amilsiddik@unhas.ac.id, [3]newima@gmail.com, [4]ancakiky@gmail.com, [5]nurhilal.asyahrir@unsulbar.ac.id, [6]aidarangkuti@yahoo.com [7]wahyudirusdi@iaingorontalo.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 03 Desember 2024, diterima untuk diterbitkan: 27 Agustus 2025)

Abstrak

Banyak obat potensial telah ditemukan dari produk alami laut (*Marine Natural Product*). Hal ini menunjukkan bahwa senyawa laut merupakan sumber penting dalam pengembangan dan penemuan obat. Meskipun banyak senyawa laut yang menunjukkan aktivitas biologis tertentu, hanya sedikit yang tercatat sebagai senyawa antibakteri. Oleh karena itu, menemukan senyawa yang berpotensi sebagai senyawa antibakteri dari organisme laut masih menjadi tantangan. Tujuan dari penelitian ini adalah untuk memanfaatkan pendekatan komputasi untuk menemukan senyawa antibakteri dari produk alami laut yang berpotensi menjadi obat. Penelitian ini berfokus pada penggunaan model *Artificial Neural Network* (ANN), *Random Forest*, dan *XGBoost* untuk melakukan klasifikasi berdasarkan kemiripan kimiawi antara senyawa produk alami laut di Indonesia dengan senyawa antibakteri. Untuk mengatasi ketidakseimbangan data, digunakan teknik *resampling* berupa SMOTE dan *undersampling (US)*. Hasil penelitian menunjukkan bahwa akurasi XGBoost + SMOTE memiliki nilai yang paling tinggi, yaitu 98.89%, mengungguli model ANN 97.57%, *Random Forest* (RF) 97.06%, serta model dengan resampling lain seperti ANN+SMOTE 98.67% dan RF + SMOTE 98.59%. Sementara itu, penerapan teknik undersampling menyebabkan penurunan akurasi secara signifikan, di mana XGBoost + US, RF + US, dan ANN + US masing-masing hanya mencapai 91.12%, 91.59%, dan 87.85%. Dari 73 senyawa biota laut, hanya senyawa yang memiliki CID 101767277 yang diprediksi sebagai senyawa yang potensial sebagai antibakteri.

Kata kunci: *Artificial Neural Network, Random Forest, XGBoost, antibakteri, SMOTE*

COMPARISON OF ANN, RANDOM FOREST, AND XGBOOST IN ANTIBIOTIC CLASSIFICATION WITH SAMPLING METHODS

Abstract

Many potential drugs have been discovered from marine natural products. This suggests that marine compounds are essential in drug development and discovery. Although many marine compounds exhibit certain biological activities, only a few have been recorded as antibacterial compounds. Therefore, finding compounds with potential as antibacterial compounds from marine organisms remains a challenge. This paper aims to utilize computational approaches to discover antibacterial compounds from marine natural products that have the potential to become drugs. This research focuses on the use of Artificial Neural Network (ANN), Random Forest (RF), and XGBoost models to perform classification based on chemical similarity between compounds of marine natural products in Indonesia and antibacterial compounds. To overcome data imbalance, resampling techniques such as SMOTE and undersampling (US) were used. The results showed that the accuracy of XGBoost + SMOTE has the highest value, which is 98.89%, outperforming the ANN model 97.57%, Random Forest (RF) 97.06%, as well as models with other resampling such as ANN+SMOTE 98.67% and RF + SMOTE 98.59%. Meanwhile, the application of undersampling techniques caused a significant decrease in accuracy, where XGBoost + US, RF + US, and ANN + US only reached 91.12%, 91.59%, and 87.85%, respectively. Of the 73 marine biota compounds, only compounds that have CID 101767277 *are predicted as potential antibacterial compounds.*

Keywords: *Artificial Neural Network, Random Forest, XGBoost, antibacterial, SMOTE*

1. PENDAHULUAN

Antibiotik adalah zat organik atau anorganik yang dapat membunuh atau membatasi perkembangan mikroorganisme, yang dapat berdampak pada kemampuan mereka untuk bertahan hidup (Rusdi et al., 2023). Perkembangan antibiotik diakui sebagai salah satu pencapaian terbesar dalam bidang ilmu pengetahuan dan kedokteran pada abad ke-20. Penggunaan antibiotik secara luas telah berhasil mengurangi tingkat kematian dan komplikasi yang berkaitan dengan penyakit infeksi serius, seperti tuberkulosis (TBC), sifilis, pneumonia, dan gonore, baik pada manusia maupun hewan (Li et al., 2020). Meskipun banyak antibiotik telah dikembangkan, saat ini resistensi terhadap antibiotik semakin meningkat. Hal ini menyebabkan banyak obat menjadi kurang efektif, karena bakteri dapat beradaptasi dan mentolerir pengobatan tersebut. Akibatnya, pengobatan penyakit menular menjadi semakin sulit dan kompleks (Sarvananda & D Premarathne, 2022).

Salah satu alternatif pencarian untuk mengatasi masalah resistensi antibiotik adalah menggunakan Produk alami laut (*Marine Natural Products*, MNPs). MNPs telah menjadi fokus utama penelitian sebagai sumber senyawa bioaktif yang unik. Lingkungan laut yang ekstrem mendorong evolusi metabolit sekunder dengan struktur kimia yang khas, yang sering menunjukkan aktivitas biologis kuat, termasuk aktivitas antibakteri. Oleh karena itu, MNPs muncul sebagai kandidat potensial dalam pengembangan obat baru, terutama di era resistensi antibiotik yang menjadi ancaman kesehatan global (Lv & Zeng, 2024). Mencari MNPs, mengekstrak senyawa kimia, dan melakukan eksperimen laboratorium untuk mengungkap aktivitas biologis dari senyawa tersebut merupakan beberapa tantangan yang perlu dihadapi. Proses ini memakan waktu dan memerlukan biaya yang cukup besar. Untuk itu, peneliti mengadopsi pendekatan komputasi dalam penelitian ini guna mempercepat waktu penelitian dan mengurangi pengeluaran (Diéguez-Santana & González-Díaz, 2023).

Para peneliti sebelumnya telah mengevaluasi kemampuan prediktif berbagai teknik pembelajaran mesin, seperti *random forest* (RF), regresi logistik (LR), *gradient boosted regression trees* (GBRT), *support vector machine* (SVM), dan *multilayer perceptron* (MLP), untuk membangun model prediktif yang berkaitan dengan senyawa antibakteri (Li et al., 2022). Terdapat beberapa model klasifikasi pada *machine learning* seperti RF, XGBoost dan model *Artificial Neural Network* (ANN). Beberapa penelitian sebelumnya mengungkapkan hasil bahwa XGBoost mengungguli RF, SVM, dan *Neural Networks* dalam *urban forest classification* (Ramdani & Furqon, 2022). RF berkinerja lebih baik daripada *support vector machine*, dan XGBoost, dengan hasil validasi silang rata-rata untuk 13 agen antimikroba semuanya ≥ 94.6% (Gao et al., 2024). Penelitian lain menggabungkan algoritma RF dan *Convolutional Neural Network* (CNN) untuk mengklasifikasikan infeksi bakteri pada spesies tanaman, yang menunjukkan ketepatan yang luar biasa dalam klasifikasi gambar, *recall* sebesar 81.54%, *F1-Score* sebesar 81.54%, dan akurasi sebesar 81.54% (Banerjee et al., 2023).

Resampling dataset adalah teknik yang digunakan untuk mengubah distribusi data dengan tujuan meningkatkan performa model atau menangani masalah khusus seperti data yang tidak seimbang. Penelitian ini menerapkan model *Artificial Neural Network* (ANN), *Random Forest* (RF), dan XGBoost dengan memakai teknik *resampling* SMOTE (Widodo, Setiawan & Indraswari, 2024) dan *undersampling* (Sun et al., 2024) untuk klasifikasi senyawa MNP dari perairan Indonesia. Ketiga model tersebut memungkinkan prediksi aktivitas antibakteri berdasarkan fitur kimia dan struktural senyawa.

2. METODE PENELITIAN

Pada penelitian ini, alur penelitian dibagi menjadi 4, yaitu: pengumpulan dataset, *preprocessing*, melatih model, mengaplikasikan model terbaik ke dalam data uji biota laut, yang disajikan pada Gambar 1.

2.1. Data Penelitian

Dalam penelitian ini, dataset yang dikumpulkan terdiri dari tiga set senyawa: senyawa antibiotik, senyawa obat antibiotik, dan senyawa produk alami laut. Sebanyak 1601 senyawa terkait dengan antibiotik, baik obat maupun non-obat, diperoleh dari database PubChem (Kim et al., 2021). Senyawa-senyawa tersebut didapatkan dengan menggunakan kata kunci "*antibiotic*" pada kolom pencarian di situs PubChem. Semua senyawa tersebut diunduh dalam format molekul SDF.

Dataset kedua diperoleh dari basis data DrugBank. Peneliti mengumpulkan data senyawa antibiotik yang tersedia di DrugBank (Wishart et al., 2018), dan dari 3908 senyawa, hanya 535 senyawa obat antibiotik yang memiliki file SDF yang sesuai di PubChem. Sisanya, sebanyak 3373 senyawa, merupakan senyawa obat non-antibiotik.

Dataset terakhir adalah kumpulan senyawa laut yang diperoleh dari perairan Sulawesi Selatan, Indonesia. Senyawa ini diidentifikasi berdasarkan literatur (Hanif et al., 2019). Peneliti mengumpulkan 73 senyawa dari 17 organisme laut yang dikumpulkan di area tersebut. Senyawa-senyawa ini kemudian diverifikasi di PubChem, sama seperti dataset senyawa lainnya.

Semua senyawa yang didapatkan dari ketiga dataset kemudian digenerate untuk menghasilkan *fingerprint* yang akan dianalisis lebih lanjut.

Gambar 1 Alur Penelitian

2.2. Koefisien Jaccard

Dalam penelitian ini, kemiripan kimiawi antar senyawa dihitung sebagai input untuk model. Peneliti menghasilkan dua jenis matriks kemiripan kimia. Matriks pertama menggambarkan kemiripan antara 73 senyawa laut dengan 1601 senyawa antibiotik, sedangkan matriks kedua menunjukkan kemiripan antara 3908 senyawa obat dengan 1601 senyawa antibakteri. Kemiripan kimiawi ini didasarkan pada kemiripan struktur yang diukur menggunakan substruktur dari masing-masing senyawa, yang direpresentasikan sebagai sidik jari (*fingerprint)* dengan bantuan perangkat lunak Python. Untuk memperoleh nilai kemiripan, peneliti menggunakan koefisien Jaccard (Rusdi et al., 2023), yang dirumuskan pada persamaan (1) :

$$S_{Ab} = \frac{z}{[x+y+z]} \qquad (1)$$

Di mana x adalah jumlah bit yang mirip dalam senyawa pertama, y adalah jumlah bit yang mirip dalam senyawa kedua, dan z adalah jumlah bit yang mirip dalam kedua senyawa.

2.2. Pembagian Data

Matriks *similarity* antara senyawa obat dan senyawa antibakteri berukuran 3908 (target) × 1601, yang kemudian dibagi menjadi 80% data *training* dan 20% data *validation*. Dimana untuk target 3908 terdiri dari 2 kelas yaitu 535 senyawa obat antibiotik dan 3373 senyawa obat non-antibiotik. Sementara itu, data *testing* merupakan matriks *similarity* antara senyawa produk laut dan senyawa antibakteri dengan ukuran 73 (target) × 1601.

2.3. *Resampling*

Pada penelitian ini memakai 3 kasus: Pertama dataset tanpa *resampling*, kedua adalah SMOTE dan ketiga adalah *Undersampling* (US). *Resampling* dataset adalah teknik yang digunakan untuk mengubah distribusi data dengan tujuan meningkatkan performa model atau menangani masalah khusus seperti data yang tidak seimbang. Untuk setiap skenario, dataset dibagi menjadi 80% data pelatihan dan 20% data validasi untuk memastikan evaluasi performa model yang konsisten.

Pada matriks *similarity* antara senyawa obat dengan senyawa anti bakteri memiliki distribusi kelas pada Tabel 1 terlihat dataset tidak seimbang yaitu 3373 senyawa obat non-antibiotik dan 535 senyawa obat antibiotik.

SMOTE bekerja dengan menambahkan sampel sintetis pada kelas minoritas berdasarkan karakteristik data yang sudah ada, sehingga distribusi kelas menjadi lebih seimbang. Adapun rumus SMOTE pada persamaan (2):

$$X_{baru} = X_i + \left(\widehat{X}_i - X_i\right) \times \delta \qquad (2)$$

Vektor fitur dari kelas minoritas direpresentasikan dengan X_i. K-nearest neighbors (tetangga paling dekat) dari X_i direpresentasikan dengan $\widehat{X}_i$. Untuk menghitung sampel sintetis, digunakan bilangan acak antara nol dan satu yang dilambangkan dengan δ, kemudian dikalikan dengan selisih antara X_i dan $\widehat{X}_i$. Nilai hasil perkalian tersebut kemudian ditambahkan dengan X_i untuk menghasilkan sampel sintetis baru (X_{baru}) (Pulungan, Purnomo and Kurniasih, 2024).

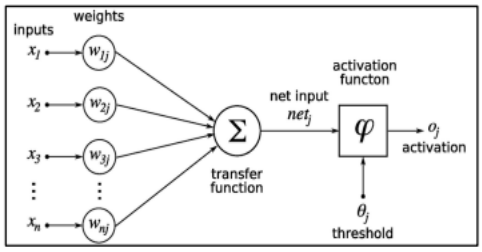
Undersampling adalah metode yang mengurangi ketidakseimbangan kelas dengan secara acak menghapus sebagian data dari kelas mayoritas agar setara atau lebih seimbang dengan kelas minoritas Teknik ini diterapkan ketika jumlah data pada kelas minoritas sudah mencukupi untuk pelatihan model, namun kelemahannya terletak pada kemungkinan terhapusnya data penting dari kelas mayoritas karena pemilihan dilakukan secara acak Hasil dari teknik ini dapat berupa jumlah data yang sama antara kelas mayoritas dan minoritas, atau pengurangan sebagian data mayoritas sesuai kebutuhan (Budi Utomo et al., 2024).

Tabel 1 Distribusi Kelas

Kasus	Non-Antibiotik	Antibiotik
Original	3373	535
SMOTE	3373	3373
Undersampling	535	535

2.4. Artificial Neural Networks (ANN)

Artificial Neural Network (ANN) yang dikenal dengan Jaringan Saraf Tiruan (JST) adalah rekayasa pengetahuan pada kecerdasan buatan atau teknik pengolahan informasi yang cara kerjanya seperti sistem saraf biologis khususnya pada sel otak manusia. Otak terdiri dari satu set sel-sel saraf yang saling berhubungan, atau unit informasi pengolahan dasar. Unit ini disebut dengan neuron, dimana saling terhubung dan bekerja sama menyelesaikan sebuah masalah seperti masalah klasifikasi atau prediksi dengan melakukan proses belajar melalui perubahan bobot sinapsisnya. Pada Gambar 3 terlihat dimana setiap input terhubung oleh bobot, semua input yang telah diberi bobot ini dikombinasikan dalam proses yang disebut fungsi penjumlahan (*summing function*), hasil penjumlahan ini disebut sebagai *net input*, fungsi aktivasi (*activation function*), dan keluaran (*output*) (Jalali & Etemadfard, 2024).



Gambar 2 Model Neuron ANN

Backpropagation merupakan algoritma yang melakukan pembelajaran terbimbing (*supervised learning*) dan digunakan pada jaringan *multi-layer* yang terdiri dari beberapa *hidden layer* serta bertujuan untuk meminimalkan error terhadap jaringan yang menghasilkan output. *Backpropagation* ini melakukan pelatihan jaringan dalam 3 tahap yaitu fase *feedforward* (tahap maju), fase *backpropagation*, dan fase perubahan atau penyesuaian bobot dengan bobot yang diatur secara terarah. Ketiga langkah tersebut akan dijalankan sampai memenuhi kondisi berhenti. Kondisi berhenti bergantung pada nilai iterasi maksimal yang telah ditentukan sebelumnya.

Tahap I (Fase *feedforward*) merupakan fase pengolahan masukan yang dihitung maju dengan menggunakan fungsi aktivasi yang telah ditetapkan dari input layer hingga respons hasil yang dihasilkan output layer.

$$Z_{net\ j} = v_0 + \sum_{i=1}^{n} x_i v_{ij} \qquad (3)$$

Pada persamaan (3), $Z_{net\ j}$ adalah Nilai input layer yang masuk menuju *hidden layer* untuk unit Z_j, v_{0j} adalah Nilai bobot bias pada *hidden layer* ke-j, x_i adalah nilai input pada unit ke-i, v_{ij} = nilai bobot antara input ke-i dan *hidden layer* ke-j.

Tahap II (Fase *Backpropagation*) merupakan fase di mana *error* dihitung dan disebarkan kembali dari *output* layer ke unit hidden layer.

$$\delta = \begin{cases} 0, & if\ Z_{netj} \le 0 \\ \delta_{netj}, & if\ Z_{netj} > 0 \end{cases} \qquad (4)$$

Persamaan (4) di mana δ_j adalah nilai aktivasi *error* pada unit *hidden layer* ke-j, δ_{netj} adalah nilai *error* pada unit *hidden layer* ke-j.

$$\delta_k = (t_k - Y_k) \qquad (5)$$

Persamaan (5) di mana δ_k adalah nilai aktivasi *error* pada output layer ke-k, t_k adalah nilai target *output* ke-k, Y_k adalah nilai *output* dari *output* layer-k setelah penerapan fungsi aktivasi.

Tahap III (Fase Perubahan Bobot dan Bias) merupakan fase dilakukannya perubahan (modifikasi) bobot agar dapat mengurangi *error* (kesalahan) yang terjadi.

$$\Delta w_{jk} = \alpha\ \delta_k\ Z_j \qquad (6)$$

Persamaan (6) di mana Δw_{jk} adalah perubahan bobot yang menghubungkan *hidden layer* j dengan layer k, Z_j adalah nilai output dari *hidden layer* j setelah penerapan fungsi aktivasi.

$$\Delta v_{ij} = \alpha\ \delta\ x_i \qquad (7)$$

Persamaan (7) di mana Δv_{ij} adalah perubahan bobot yang menghubungkan input layer i dengan *hidden layer* j, α adalah *learning rate*.

Model Arsitektur ANN yang dipakai dalam terdiri dari 3 layer yang dapat dilihat pada Tabel 2.

Tabel 2 Ringkasan Model Arsitektur ANN

Layer	Output	Parameter
Dense	(none,100)	160200
Dense1	(none,100)	10100
Dense2	(none,1)	101

Pada Tabel 2 Arsitektur yang terdiri dari 3 layer dengan total parameter adalah 1700401. Dengan menggunakan metode *grid search* didadapatkan *hyperparameter* yang digunakan adalah *learning_rate* = 0.005, *optimizer* = Adam, *epoch* = 200 dengan *earlystopping*, dan *batch_size* = 128.

2.5. XGBoost

Algoritma eXtreme Gradient Boosting (XGBoost), pertama kali diusulkan oleh Dr. Tianqi Chen dari University of Washington pada tahun 2014 (Zhang & Gong, 2020). XGBoost adalah versi perbaikan dari algoritma Gradient Boosting yang dapat secara efisien membangun pohon klasifikasi dan beroperasi secara paralel. XGBoost beroperasi berdasarkan prinsip *ensemble learning*, XGBoost juga mampu mengoptimalkan fungsi tujuan dengan efisien melalui pendekatan gradien stokastik (Nayan Kumar Sinha, 2020). Algoritma XGBoost Classifier sebagai berikut:

Tahap I Menghitung nilai objektif

$$ob = L(\theta) + \Omega(\theta) \qquad (8)$$

Persamaan (8) di mana adalah $L(\theta)$ fungsi loss dan $\Omega(\theta)$ adalah fungsi regularisasi.

Tahap II Menghitung Nilai fungsi loss

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) \quad (9)$$

Persamaan (9) di mana y_i adalah nilai sebenarnya dan $\hat{y}_i$ adalah nilai prediksi.

2.6. Random Forest (RF)

RF merupakan pengembangan dari metode *Decision Tree*, Proses membangun model RF dimulai dengan membuat beberapa salinan data menggunakan teknik acak yang disebut *bootstrap sampling*. Dari data asli, diambil beberapa sampel secara acak (dengan pengembalian), dan setiap sampel memiliki ukuran yang sama dengan data awal. Kemudian, dari setiap sampel ini dibangun sebuah pohon regresi menggunakan metode pemisahan berulang seperti pada algoritma CART. Yang membuat RF lebih acak adalah, pada setiap titik percabangan dalam pohon, hanya sebagian fitur yang dipilih secara acak untuk dipertimbangkan. Dari fitur-fitur tersebut, sistem memilih pemisahan terbaik yang menghasilkan kesalahan prediksi paling kecil. Proses ini terus dilakukan hingga tidak bisa dipisah lagi, dan pohonnya dibiarkan tumbuh sepenuhnya tanpa dipotong. Setelah semua pohon selesai dibuat, hasil akhirnya diperoleh dengan mengambil rata-rata prediksi dari semua pohon, yang itulah disebut sebagai prediksi dari model RF. (Makariou, Barrieu & Chen, 2021).

2.7. Evaluasi

Kriteria evaluasi yang dipakai pada dua kelas dengan *confusion matrix*. *Confusion matrix* memberikan informasi tentang jumlah prediksi yang ada di kolom dan jumlah kelas yang sebenarnya di baris, seperti pada Tabel 3.

Tabel 3. *Confusion Matrix*

	Kelas	**Prediksi**	
		Antibiotik	**Non-Antibiotik**
Aktual	Antibiotik	TP	FN
	Non-Antibiotik	FP	TN

Tabel 3 menyajikan *confusion matrix* yang menggambarkan performa klasifikasi dokumen ke dalam dua kategori, yaitu "Antibiotik" dan "Non-Antibiotik". *True Positive* (TP) menunjukkan jumlah senyawa yang benar diklasifikasikan sebagai "Antibiotik", sementara *True Negative* (TN) merupakan jumlah senyawa yang benar diklasifikasikan sebagai "Non-Antibiotik". *False Positive* (FP) mencerminkan jumlah senyawa "Non-Antibiotik" yang salah diklasifikasikan sebagai "Antibiotik", sedangkan *False Negative* (FN) menunjukkan jumlah senyawa "Antibiotik" yang salah diklasifikasikan sebagai "Non-Antibiotik". Matrix ini membantu dalam mengevaluasi akurasi dan efektivitas model klasifikasi yang digunakan. Dalam membandingkan algoritma ada beberapa metrik evaluasi (Chachoui et al., 2024) yang umumnya digunakan yaitu:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$Presisi = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

$$F1\ Score = \frac{2 \times presisi \times recall}{Presisi+Recall} \quad (12)$$

3. HASIL DAN PEMBAHASAN

3.1. Evaluasi Kinerja Model

Hasil Evaluasi Kinerja model dapat dilihat pada Tabel 4 berikut ini.

Tabel 4. Evaluasi Kinerja Model

Model	**Akurasi (%)**	**Presisi (%)**	***Recall* (%)**	**F1score (%)**
ANN	**97.57**	**97.49**	**93.52**	**95.37**
RF	97.06	97.52	91.64	94.28
XGBoost	97.44	97.41	93.13	95.11
ANN+SMOTE	98.67	98.67	98.66	98.66
RF+SMOTE	98.59	98.59	98.59	98.59
XGBoost+SMOTE	**98.89**	**98.89**	**98.88**	**98.88**
ANN+US	87.85	87.73	87.84	87.78
RF+US	**91.59**	**91.52**	**91.52**	**91.52**
XGBoost+US	91.12	91.08	91.01	91.05

Kasus I tanpa teknik resampling

Model ANN tanpa penyeimbangan data menunjukkan akurasi tertinggi sebesar 97.57%, dengan presisi 97.49%, *recall* 93.52%, dan *F1-score* 95.37%. Hal ini menunjukkan bahwa model cukup efektif dalam mengklasifikasikan data. Meskipun nilai *recall* lebih rendah dibandingkan presisi, model lebih sering mengidentifikasi senyawa yang antibakteri (positif) daripada mengabaikan yang seharusnya terklasifikasi positif.

Model RF tanpa penyeimbangan data memiliki akurasi sedikit lebih rendah dari ANN, yaitu 97.06%. Presisi yang dihasilkan sebesar 97.52% dan *recall* 91.64%, menghasilkan *F1-score* 94.28%. Hal ini menunjukkan bahwa model memiliki akurasi yang cukup tinggi. Namun, model masih kurang optimal dalam mengidentifikasi seluruh senyawa yang positif.

XGBoost menunjukkan performa dengan akurasi sebesar 97.44%, sedikit lebih rendah dari ANN namun lebih tinggi dari RF. Presisi dan *recall* masing-masing adalah 97.41% dan 93.13%, dengan *F1-score* 95.11%. Hal ini menunjukkan bahwa XGBoost memberikan keseimbangan yang cukup baik antara presisi dan *recall*.

Dari Tabel 4 Hasil evaluasi performa 3 model klasifikasi menunjukkan bahwa ANN memiliki kinerja terbaik dengan akurasi sebesar 97.57%, presisi 97.49%, *recall* 93.52%, dan *F1-score* 95.37%. Model RF memperoleh akurasi 97.06%, presisi 97.52%, *recall* 91.64%, dan *F1-score* 94.28%. Sementara itu, model XGBoost menunjukkan hasil yang kompetitif dengan akurasi 97.44%, presisi

97.41%, *recall* 93.13%, dan *F1-score* 95.11%. Secara umum, ANN unggul dalam keseluruhan metrik, meskipun XGBoost juga menunjukkan performa yang sangat baik dan mendekati ANN.

Kasus II dengan SMOTE

Pada saat SMOTE digunakan untuk menyeimbangkan dataset, performa model meningkat secara signifikan. ANN+SMOTE dengan akurasi yaitu 98.67%, dan presisi, *recall*, serta *F1-score* semuanya konsisten pada 98.66%. Ini menunjukkan peningkatan yang signifikan dalam kemampuan model untuk mengidentifikasi positif sebenarnya.

RF+SMOTE juga menunjukkan peningkatan performa dengan akurasi sebesar 98.59%, dan semua metrik (presisi, *recall*, dan *F1-score*) sebesar 98.59%.

XGBoost+SMOTE bahkan sedikit lebih baik dengan akurasi 98.89% dan konsistensi di semua metrik. Peningkatan ini menegaskan bahwa penggunaan SMOTE membantu model mengatasi ketidakseimbangan data, membuatnya lebih akurat dalam mengklasifikasikan sampel.

Penggunaan SMOTE: Teknik SMOTE secara konsisten meningkatkan akurasi dan metrik performa lainnya untuk semua model (ANN, RF, dan XGBoost), menunjukkan efektivitasnya dalam menangani ketidakseimbangan data.

Kasus III dengan US:

Saat model diterapkan dengan teknik *undersampling*, terjadi penurunan akurasi dibandingkan dengan model standar dan yang menggunakan SMOTE.

ANN+US menunjukkan akurasi sebesar 87.85%, dengan presisi dan *recall* sekitar 87.73% dan 87.84%, menghasilkan *F1-score* 87.78%.

RF+US sedikit lebih baik dengan akurasi 91.59% dan konsistensi presisi, *recall*, serta *F1-score* sebesar 91.52%.

XGBoost+US memiliki akurasi sebesar 91.12%, dengan sedikit perbedaan pada presisi, *recall*, dan *F1-score* sekitar 91.08%.

Teknik *undersampling* cenderung menurunkan akurasi karena berkurangnya data untuk pelatihan, meskipun hal ini dapat membantu mengurangi bias model pada kelas mayoritas.

Performa Model: ANN dengan SMOTE memberikan hasil terbaik, namun perbedaan antara model ANN, RF, dan XGBoost tidak terlalu signifikan saat menggunakan SMOTE, yang menunjukkan bahwa semua model dapat bekerja dengan baik jika dataset seimbang.

3.2. Pengetesan Model terbaik

Pemilihan 6 model terbaik, yaitu ANN, ANN+SMOTE, RF, RF+SMOTE, XGBoost, dan XGBoost+SMOTE, didasarkan pada kombinasi dari performa metrik evaluasi yang unggul serta representasi yang mencerminkan pengaruh teknik penyeimbangan data terhadap hasil klasifikasi. Selain itu, karena teknik *undersampling* terjadi penurunan akurasi yang cukup signifikan, sehingga model yang menerapkan teknik tersebut tidak terpilih. Setelah mendapatkan model yang paling baik performanya selanjutnya peneliti akan menguji dataset matriks *similarity* dari 73 senyawa biota laut dengan 1601 senyawa antibiotik kemudian memasukkan dataset tersebut ke dalam 6 model terbaik yang telah diuji performanya. Peneliti mendapatkan hasil dari 73 senyawa biota laut hanya 6 yang positif sebagai calon kandidat obat antibiotik, hasilnya ditampilkan pada Tabel 5.

CID 101767277 (Sarasinode K):
Senyawa ini berasal dari *M. sarassinorum*. Model ANN+SMOTE, RF+SMOTE, XGBoost, dan XGBoost+SMOTE menunjukkan hasil positif, dengan total empat model mendukung senyawa ini sebagai kandidat obat. Hasil ini menunjukkan bahwa Sarasinode K memiliki potensi yang lebih tinggi dibandingkan senyawa lain dalam Tabel 4.

CID 11441502 (Boneratamide B methyl ester):
Berasal dari *A. aplysinoides*, senyawa ini didukung oleh tiga model: RF+SMOTE, XGBoost, dan XGBoost+SMOTE. Hasil ini menunjukkan dukungan yang cukup baik, tetapi masih kurang dibandingkan dengan Sarasinode K.

CID 11384948 (Boneratamide A methyl ester):
Senyawa ini, juga dari *A. aplysinoides*, didukung oleh dua model, yaitu RF dan XGBoost+SMOTE. Meskipun masih terdapat potensi, dukungan dari model lebih rendah dibandingkan dengan senyawa lainnya.

CID 102010367 (-Theonellapeptolide IIe):
Senyawa ini ditemukan dalam *T. swinhoei* dan didukung oleh dua model, yaitu ANN dan ANN+SMOTE. Hasil ini menunjukkan bahwa meskipun ada beberapa indikasi potensi, senyawa ini masih memerlukan verifikasi lebih lanjut.

CID 11658230 (Sarasinode J):
Senyawa dari *M. sarassinorum* ini hanya didukung oleh satu model, yaitu XGBoost. Hal ini menandakan bahwa potensi senyawa ini masih sangat terbatas berdasarkan pengujian yang dilakukan.

CID 10781825 (Pandangolide 2):
Berasal dari jamur yang bersimbiosis dengan spons, senyawa ini didukung oleh satu model (ANN+SMOTE). Sama seperti Sarasinode J, senyawa ini masih membutuhkan lebih banyak penelitian untuk menentukan efektivitasnya.

Hasil pengujian menunjukkan bahwa **Sarasinode K** (CID 101767277) memiliki dukungan tertinggi, dengan empat model yang menunjukkan hasil positif. Sesuai dengan hasil penelitian senyawa ini bisa menjadi fokus utama untuk penelitian lebih lanjut sebagai kandidat obat antibakteri potensial. Sementara itu, senyawa lainnya memerlukan investigasi tambahan dan validasi lebih lanjut berdasarkan hasil model yang lebih terbatas.

Tabel 5. Pengujian calon kandidat obat

CID	Sources of Medicinal plant (Compound name)	1	2	3	4	5	**6**	Total
101767277	*M.sarassinorum (Sarasinode K)*	-	+	-	+	+	+	4
11441502	*A. aplysinoides(Boneratamide B methyl ester)*	-	-	-	+	+	+	3
11384948	*A. aplysinoides (Boneratamide A methyl ester)*	-	-	+	-	+	+	2
102010367	*T. swinhoei (-Theonellapeptolide IIe)*	+	+	-	-	-	-	**2**
11658230	*M.sarassinorum (Sarasinode J)*	-	-	-	-	+	-	1
10781825	*T. A fungus (symbiont) a sponge (host) (Pandangolide 2)*	-	+	-	-	-	-	1

Keterangan:
1 = ANN 2 = ANN+SMOTE 3 = RF 4 = RF+SMOTE
5 = XGBoost 6 = XGBoost+SMOTE

4. KESIMPULAN DAN SARAN

Penelitian mengidentifikasi senyawa antibakteri dari produk alami laut Indonesia menggunakan pendekatan pembelajaran mesin, yaitu *Artificial Neural Network* (ANN), *Random Forest* (RF), dan XGBoost. Untuk mengatasi ketidakseimbangan data, digunakan teknik *resampling* berupa SMOTE dan *undersampling* (US). Hasil evaluasi menunjukkan bahwa XGBoost+SMOTE memberikan performa terbaik dengan akurasi, presisi, *recall*, dan F1-score sebesar ≥98.88%. Model ANN+SMOTE dan RF+SMOTE juga menunjukkan kinerja tinggi dengan akurasi masing-masing 98.67% dan 98.59%. Sementara itu, penerapan *undersampling* menurunkan akurasi secara signifikan, di mana ANN+US hanya mencapai 87.85%, RF+US sebesar 91.59%, dan XGBoost+US sebesar 91.12%. Dari 73 senyawa, enam senyawa diidentifikasi sebagai kandidat antibiotik, dengan Sarasinode K (CID 101767277) sebagai kandidat utama. Penelitian lanjutan direkomendasikan menggunakan *uji in silico*, *in vitro*, dan *in vivo* untuk mengonfirmasi aktivitas antibakterinya.

ACKNOWLEDGMENTS

Penelitian ini didukung secara finansial oleh LPPM Universitas Hasanuddin melalui skema "Penelitian Dosen Pemula Unhas (PDPU)". Dengan Nomor: 00310/UN4.22/PT.01.03/2024.

DAFTAR PUSTAKA

BANERJEE, D., KUKREJA, V., HARIHARAN, S., JAIN, V. & JINDAL, V., 2023. Predicting Tulip Leaf Diseases: A Integrated CNN and Random Forest Approach. In: *2023 World Conference on Communication & Computing (WCONF)*. pp.1–6.

BUDI UTOMO, P., FARUQZIDDAN, M., HERDIKA SEPTA AULIA, E. & DINI AZZAHRA, S., 2024. Perbandingan Skenario Balancing Oversampling dan Undersampling dalam Klasifikasi Resiko Kambuh Kanker Tiroid menggunakan Algoritma SVM Linear. *JAMI: Jurnal Ahli Muda Indonesia*, [online] 5(2), pp.172–182.

CHACHOUI, Y., AZIZI, N., HOTTE, R. & BENSEBAA, T., 2024. Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning. *Computers and Education: Artificial Intelligence*, [online] 6, p.100222.

DIÉGUEZ-SANTANA, K. & GONZÁLEZ-DÍAZ, H., 2023. Machine learning in antibacterial discovery and development: A bibliometric and network analysis of research hotspots and trends. *Computers in Biology and Medicine*, [online] 155, p.106638.

GAO, Y., LI, H., ZHAO, C., LI, S., YIN, G. & WANG, H., 2024. Machine learning and feature extraction for rapid antimicrobial resistance prediction of Acinetobacter baumannii from whole-genome sequencing data. *Frontiers in Microbiology*, [online] 14.

HANIF, N., MURNI, A., TANAKA, C. & TANAKA, J., 2019. *Marine natural products from Indonesian waters*. *Marine Drugs*.

JALALI, R. & ETEMADFARD, H., 2024. Spatio-temporal analysis of COVID-19 lockdown effect to survive in the US counties using ANN. *Scientific Reports*, [online] 14(1), p.19608.

KIM, S., CHEN, J., CHENG, T., GINDULYTE, A., HE, J., HE, S., LI, Q., SHOEMAKER, B.A., THIESSEN, P.A., YU, B., ZASLAVSKY, L., ZHANG, J. & BOLTON, E.E., 2021. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research*, 49(D1), pp.D1388–D1395.

LI, W.-X., TONG, X., YANG, P.-P., ZHENG, Y., LIANG, J.-H., LI, G.-H., LIU, D., GUAN, D.-G. & DAI, S.-X., 2022. Screening of antibacterial compounds with novel structure from the FDA approved drugs using machine learning methods. *Aging*, 14(3), pp.1448–1472.

LI, Z., LI, M., ZHANG, Z., LI, P., ZANG, Y. & LIU, X., 2020. Antibiotics in aquatic environments of China: A review and meta-analysis. *Ecotoxicology and Environmental Safety*, [online] 199, p.110668.

LV, F. & ZENG, Y., 2024. *Novel Bioactive Natural Products from Marine-Derived Penicillium Fungi: A Review (2021–2023)*. *Marine Drugs*,

MAKARIOU, D., BARRIEU, P. & CHEN, Y., 2021. A random forest based approach for predicting spreads in the primary catastrophe bond market. *Insurance: Mathematics and Economics*, [online] 101, pp.140–162.

NAYAN K.S., 2020. Developing A Web based System for Breast Cancer Prediction using XGboost Classifier. *International Journal of Engineering Research and*, V9(06).

PULUNGAN, M.P., PURNOMO, A. & KURNIASIH, A., 2024. Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier. *Jurnal Teknologi Informasi dan Ilmu Komputer*, [online] 11(5), pp.1033–1042.

RAMDANI, F. & FURQON, M.T., 2022. The simplicity of XGBoost algorithm versus the complexity of Random Forest, Support Vector Machine, and Neural Networks algorithms in urban forest classification. *F1000Res*. [online]

RUSDI, E.S., SYAHRIR, N.H.A., SIDDIK, A.MUH.A., AMIR, S.B.H. & RUSDI, W., 2023. Graph Clustering Based on Chemical Similarity in Marine Compounds and Antibacterial Compounds. pp.329–338.

SARVANANDA, L. & D PREMARATHNE, A., 2022. The Growing Of Antibiotic Resistance: A Short Viewpoint. *Pharmaceutics and Pharmacology Research*, 5(3), pp.01–02.

SUN, Z., YING, W., ZHANG, W. & GONG, S., 2024. Undersampling method based on minority class density for imbalanced data. *Expert Systems with Applications*, [online] 249, p.123328.

WIDODO, A.O., SETIAWAN, B. & INDRASWARI, R., 2024. Machine Learning-Based Intrusion Detection on Multi-Class Imbalanced Dataset Using SMOTE. In: *Procedia Computer Science*. Elsevier B.V. pp.578–583.

WISHART, D.S., FEUNANG, Y.D., GUO, A.C., LO, E.J., MARCU, A., GRANT, J.R., SAJED, T., JOHNSON, D., LI, C., SAYEEDA, Z., ASSEMPOUR, N., IYNKKARAN, I., LIU, Y., MACIEJEWSKI, A., GALE, N., WILSON, A., CHIN, L., CUMMINGS, R., LE, DI., PON, A., KNOX, C. & WILSON, M., 2018. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1), pp.D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

ZHANG, D. & GONG, Y., 2020. The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure. *IEEE Access*, 8, pp.220990–221003.