Jurnal Ilmu Komputer & Agri-informatika

Tersedia secara *online* di:
http://journal.ipb.ac.id/index.php/jika

Volume 11 Nomor 2 halaman 111 – 121
e ISSN: 2654-9735, p ISSN: 2089-6026

SAE-DNN-GA: Sebuah Pendekatan Klasifikasi Multilabel dalam Prediksi Senyawa Herbal Potensial untuk Penyakit COVID-19

SAE-DNN-GA: A Multilabel Classification Approach in Predicting Potential Herbal Compounds for COVID-19 Diseases

EKO PRAJA HAMID WIJAYA[1]*, WISNU ANANTA KUSUMA[1], SONY HARTONO WIJAYA[1], AULIA FADLI[2]

Abstrak

COVID-19 adalah penyakit dengan laju penyebaran yang tinggi. Percepatan proses penemuan obat untuk penyakit tersebut sangat dibutuhkan. Penggunaan kembali obat (*drug repurposing*) merupakan salah satu alternatif dalam pengembangan dan penemuan obat dengan biaya murah serta waktu yang singkat. Tanaman herbal dapat digunakan sebagai obat dengan khasiat yang lebih baik, efek samping yang lebih sedikit, dan lebih murah. Prediksi interaksi obat-target dan penggunaan kembali obat dapat digunakan untuk mengeksplorasi senyawa herbal potensial. Penelitian ini mengatasi kelemahan klasifikasi biner dengan model DSSL-DTI (*Deep Semi Supervised Learning-Drug Target Interaction*) yang dioptimasi menggunakan algoritma genetika. Tujuan penelitian ini adalah mendeteksi kemungkinan adanya hubungan antar label menggunakan pendekatan klasifikasi multilabel dengan model yang dioptimasi. Data yang digunakan dalam penelitian ini antara lain: data protein, data interaksi senyawa-protein, dan data senyawa herbal. Data protein diperoleh dari situs GeneCards yang berisi kumpulan protein yang berasosiasi dengan COVID-19 dan ditemukan pada manusia. Data interaksi senyawa-protein diperoleh dari situs DrugBank dan SuperTarget. Adapun data senyawa herbal diperoleh dari HerbalDB. Hasil penelitian menunjukkan bahwa dengan menggunakan model SAE-DNN-GA yang diusulkan, prediksi senyawa herbal menghasilkan sepuluh senyawa yang berinteraksi dengan dua protein bernilai relevansi tertinggi, yaitu protein INS (7.094) dan ALB (3.178). Hasil ini diharapkan mampu meningkatkan hasil prediksi kandidat senyawa herbal sebagai obat penyakit COVID-19 menjadi lebih akurat.

Kata Kunci**:** Algoritma genetika, COVID-19, *deep semi supervised learning*, klasifikasi multilabel.

Abstract

COVID-19 is a disease with a high transmission rate, making accelerated drug discovery efforts critical. Drug repurposing offers a cost-effective and time-efficient alternative in drug development and discovery. Herbal plants can serve as effective medications with fewer side effects and reduced costs. Drug-target interaction prediction and drug repurposing can help identify potential herbal compounds. This study addresses limitations in binary classification by employing a DSSL-DTI (Deep Semi-Supervised Learning-Drug Target Interaction) model optimized through a genetic algorithm. The aim of this research is to detect possible relationships between labels using a multilabel classification approach with an optimized model. The data used in this study includes protein data, compound-protein interaction data, and herbal compound data. Protein data was sourced from GeneCards, a collection of proteins associated with COVID-19 found in humans. Compound-protein interaction data was obtained from DrugBank and SuperTarget, while herbal compound data was retrieved from HerbalDB. The study results indicate that, using the proposed SAE-DNN-GA model, the prediction of herbal compounds identified ten compounds interacting with two proteins with the highest relevance values: Insulin (7.094) and Albumin (3.178). These findings are expected to enhance the accuracy of predicting herbal compound candidates as potential treatments for COVID-19.

*Keywords***:** *COVID-19, deep semi-supervised learning, genetic algorithm, multi-label classification.*

[1] Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680;
[2] Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sumatera Utara, Deli Serdang, 20352;
* Penulis Korespondensi: Tel: 081223220222; Surel: ekoprajahamidwijaya@apps.ipb.ac.id

PENDAHULUAN

Kasus pneumonia misterius pertama kali ditemukan di Wuhan, Provinsi Hubei, Cina pada Desember 2019 (Lee dan Hsueh 2020). Penyakit ini kemudian dikenal dengan nama *Coronavirus Infectious Disease* 2019 (COVID-19) dan dinyatakan sebagai pandemi yang memiliki dampak besar terhadap kondisi sosial ekonomi global. Gejala utama penyakit ini meliputi demam, batuk kering, kesulitan bernapas, pneumonia, kegagalan multiorgan, hingga kematian, yang disebabkan oleh virus *Severe Acute Respiratory Syndrome Coronavirus-2* (SARS-CoV-2) (Gorbalenya *et al.* 2020).

Dalam waktu satu bulan, penyakit ini telah menyebar ke beberapa negara seperti Thailand, Jepang, dan Korea Selatan (Huang *et al.* 2020)China, was caused by a novel betacoronavirus, the 2019 novel coronavirus (2019-nCoV. Cepatnya laju penyebaran tersebut mendorong percepatan proses penemuan obat untuk penyakit ini. Pengembangan obat baru membutuhkan biaya dan waktu yang besar, rata-rata lebih dari 2 miliar USD dan sekitar 10-15 tahun (Berdigaliyev dan Aljofan 2020). Salah satu alternatif yang dapat dilakukan adalah penggunaan kembali obat (*drug repurposing*), yaitu proses menemukan dan mengidentifikasi kegunaan baru dari obat yang sudah lolos uji untuk berbagai penyakit lain (Pushpakom *et al.* 2019).

Indonesia memiliki potensi besar dalam mengembangkan obat herbal (jamu) sebagai produk kesehatan berkualitas tinggi, karena memiliki jumlah spesies tanaman herbal terbanyak di dunia (Salim dan Munadi 2017). Obat herbal umumnya memiliki khasiat yang lebih baik, efek samping yang lebih sedikit, dan harga yang lebih murah dibandingkan obat konvensional (Ekor 2014). Oleh karena itu, terdapat peluang besar bagi senyawa herbal yang terkandung dalam tanaman obat untuk digunakan dalam pengobatan COVID-19.

Eksplorasi kandidat senyawa herbal potensial dapat dilakukan dengan memprediksi interaksi obat-target, seperti dalam proses penggunaan kembali obat (Fadli *et al.* 2021). Penelitian oleh Fadli *et al.* (2021) menggunakan metode DSSL dengan model SAE-DNN (*Stacked Autoencoder-Deep Neural Network*) untuk memprediksi interaksi obat-target pada kasus COVID-19, menunjukkan bahwa penggunaan *circular fingerprint* untuk fitur senyawa memberikan hasil terbaik dengan akurasi 83%.

Meskipun demikian, penggunaan klasifikasi biner dalam prediksi interaksi obat-target memiliki beberapa kelemahan. Pendekatan ini cenderung menyederhanakan masalah dengan memodelkan dimensi senyawa dan protein yang tinggi serta asosiasinya yang kompleks ke dalam model klasifikasi biner, tanpa memperhatikan hubungan antar senyawa atau antar protein (Mei dan Zhang 2019). Selain itu, suatu senyawa dapat berinteraksi dengan lebih dari satu protein target, begitu pula sebaliknya (Chu *et al.* 2021).

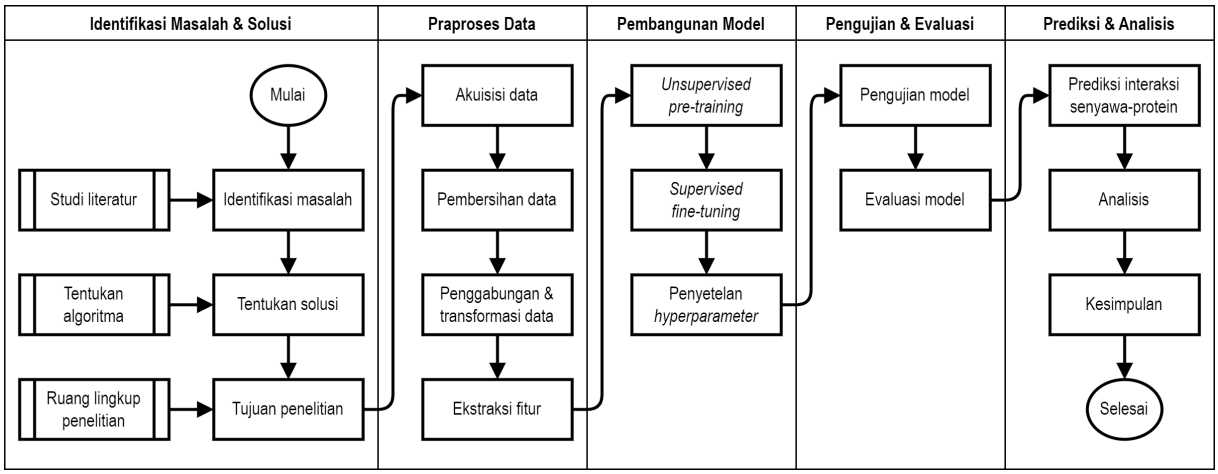
Modifikasi algoritma klasifikasi dengan metode DSSL merupakan salah satu pendekatan yang dapat digunakan untuk mengatasi kelemahan tersebut. Modifikasi ini memanfaatkan *unsupervised learning* seperti SAE sebagai prapelatihan (Erhan *et al.* 2010; Zhao dan Cao 2015; Ororbia II *et al.* 2016) untuk memperoleh bobot awal yang akan digunakan dalam pemodelan DNN, sedangkan DNN menggunakan pendekatan adaptasi algoritma untuk menyelesaikan permasalahan klasifikasi multilabel dengan performa yang sangat baik (Maxwell *et al.* 2017).

Optimasi *hyperparameter* DNN dapat meningkatkan akurasi prediksi model. *Grid search* dan *bayesian optimization* secara otomatis dapat menemukan *hyperparameter* optimal pada DNN. *Grid search* merupakan algoritma optimasi yang mencari semua kemungkinan kombinasi dalam ruang pencarian (Wicaksono dan Supianto 2018). *Bayesian optimization* membangun model probabilistik untuk memilih *hyperparameter* terbaik dari beberapa kemungkinan parameter dan mengikutsertakannya dalam pencarian *hyperparameter* terbaik pada iterasi selanjutnya (Shahriari *et al.* 2016). Algoritma genetika adalah salah satu algoritma populer untuk menyelesaikan permasalahan optimasi, dengan memberikan nilai optimal suatu fungsi yang mengadopsi proses evolusi pada populasi solusi (Rivera *et al.* 2020).

Pada penelitian ini dilakukan implementasi pendekatan klasifikasi multilabel dengan model yang di optimasi menggunakan algoritma genetika (SAE-DNN-GA) untuk mendeteksi kemungkinan adanya hubungan antar label dalam prediksi senyawa herbal potensial untuk penyakit COVID-19, sehingga akurasi prediksi model menjadi lebih optimal.

METODE

Tahapan penelitian ini terdiri atas identifikasi masalah dan solusi, praproses data, pembangunan model, pengujian dan evaluasi, serta prediksi dan analisis. Gambar 1 merupakan tahapan penelitian yang dilakukan pada penelitian ini.

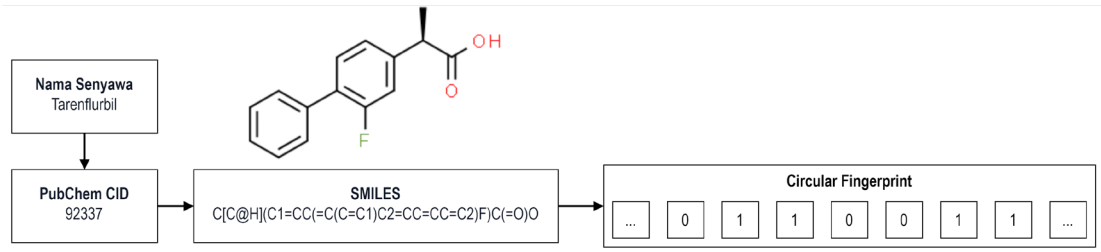


Gambar 1 Tahapan penelitian

Data dan Praproses Data

Data dalam penelitian ini terdiri atas tiga jenis, yaitu data protein, data interaksi senyawa-protein, dan data senyawa herbal. Data protein diperoleh dari situs GeneCards (Stelzer *et al.* 2016) pada Juli 2021,berisi kumpulan protein yang berasosiasi dengan COVID-19 dan terdapat pada manusia. Data tersebut digunakan sebagai input dalam pencarian senyawa yang berinteraksi dengan protein di situs SuperTarget (Hecker *et al.* 2012) dan DrugBank (Knox *et al.* 2024). Data senyawa herbal diperoleh dari situs HerbalDB (Syahdi *et al.* 2019).

Proses pembersihan dan penggabungan data dilakukan pada semua data yang diperoleh. Ekstraksi fitur senyawa menggunakan *circular fingerprint* dilakukan dengan mengidentifikasi ID PubChem setiap senyawa untuk memperoleh SMILES (*Simplified Molecular Input Line Entry System*) sebagai representasi struktur kimia senyawa. Gambar 2 merupakan proses ekstraksi fitur senyawa menggunakan *circular fingerprint*. Selanjutnya pembentukan *fingerprint* senyawa dilakukan, menghasilkan vektor fitur C ($C = [c_1, c_2, c_3, ..., c_n]$ dengan n adalah panjang *array fingerprint*) seperti terlihat pada Tabel 1. Proses transformasi pada atribut kelas data dilakukan dengan membuat *array* P ($P = [p_1, p_2, p_3, ..., p_m]$ dengan m adalah jumlah protein). Nilai p pada *array* P bernilai 1 jika senyawa pada baris tersebut berinteraksi dengan protein p, dan bernilai 0 jika tidak ada interaksi. Contoh data hasil praproses disajikan pada Tabel 2.



Gambar 2 Pengambilan deskriptor dan fitur senyawa

Tabel 1 Kumpulan data interaksi senyawa-protein

Senyawa	C_1	C_2	C_3	...	C_n	Protein Target
Dinoprostone	0	1	0	...	0	Q9Y5Y4, PTGDR2
Sulindac	0	0	0	...	0	Q9Y5Y4, PTGDR2
CHEMBL400233	0	0	0	...	0	O00142, TK2

Tabel 2 Kumpulan data hasil praproses

Senyawa					Protein				
C_1	C_2	C_3	...	C_n	P_1	P_2	P_3	...	P_m
0	0	0	...	0	1	0	0	...	0
0	1	0	...	0	1	0	0	...	0
0	1	0	...	0	0	1	0	...	0

Pembangunan Model

Pembangunan model dilakukan dengan pendekatan klasifikasi multilabel menggunakan metode *Deep Semi-Supervised Learning* (DSSL). Metode ini terdiri atas dua tahap utama, yaitu tahap *unsupervised learning* sebagai prapelatihan, dan tahap *supervised learning* sebagai pemodelan *Deep Neural Network* (DNN) untuk proses prediksi.

Pada tahap prapelatihan (*pre-training*), dilakukan proses *unsupervised learning* dengan melatih model *Stacked Autoencoder* (SAE) untuk inisialisasi bobot awal yang akan digunakan pada pemodelan DNN. Inisialisasi bobot menggunakan SAE ini bertujuan untuk menghasilkan model yang lebih optimal dibandingkan dengan inisialisasi bobot secara acak (Bahi dan Batouche 2018). Pada penelitian sebelumnya, inisialisasi bobot dengan SAE terbukti meningkatkan konvergensi dan performa model DNN dalam berbagai aplikasi (Erhan *et al.* 2010).

Setelah proses prapelatihan, pembangunan model prediksi DNN untuk menyelesaikan masalah klasifikasi multilabel dilakukan. Pembangunan model dilakukan berdasarkan hasil inisialisasi bobot menggunakan SAE. Lapisan masukan pada pemodelan SAE-DNN diambil dari data latih yang telah ditransformasi menjadi vektor fitur pada tahap praproses data. Proses ini memastikan bahwa setiap vektor fitur yang digunakan dalam model mencerminkan karakteristik penting dari data yang relevan untuk tugas prediksi.

Dalam tahap *supervised learning*, model DNN dilatih menggunakan data yang telah diberi label untuk memprediksi interaksi antara senyawa dan protein. Penggunaan klasifikasi multilabel memungkinkan model untuk menangani situasi di mana satu senyawa dapat berinteraksi dengan beberapa protein sekaligus. Hal ini sangat penting dalam konteks penelitian interaksi senyawa-protein, di mana banyak senyawa memiliki target protein yang beragam.

Penyetelan *Hyperparameter*

Penyetelan *hyperparameter* dilakukan untuk mencari kombinasi *hyperparameter* optimal yang diharapkan dapat meningkatkan performa model SAE-DNN (James *et al.* 2013). Penyetelan *hyperparameter* dalam penelitian ini menggunakan algoritma genetika yang pertama kali diperkenalkan oleh John Holland pada tahun 1975 (Holland 1992), dan didasarkan pada teori evolusi dengan prinsip seleksi alam yang dikembangkan oleh Darwin. Berbeda dengan *grid search* dan *bayesian optimization*, algoritma genetika menggunakan kriteria kinerja (*fitness*) untuk mendapatkan solusi optimal. Solusi optimal diperoleh melalui proses seleksi, mutasi, dan persilangan yang dilakukan secara berulang-ulang. Algoritma genetika mampu menangani ruang solusi yang kompleks dan tidak teratur serta masalah optimasi non-linear yang berdimensi tinggi (Roubos dan Setnes 2000).

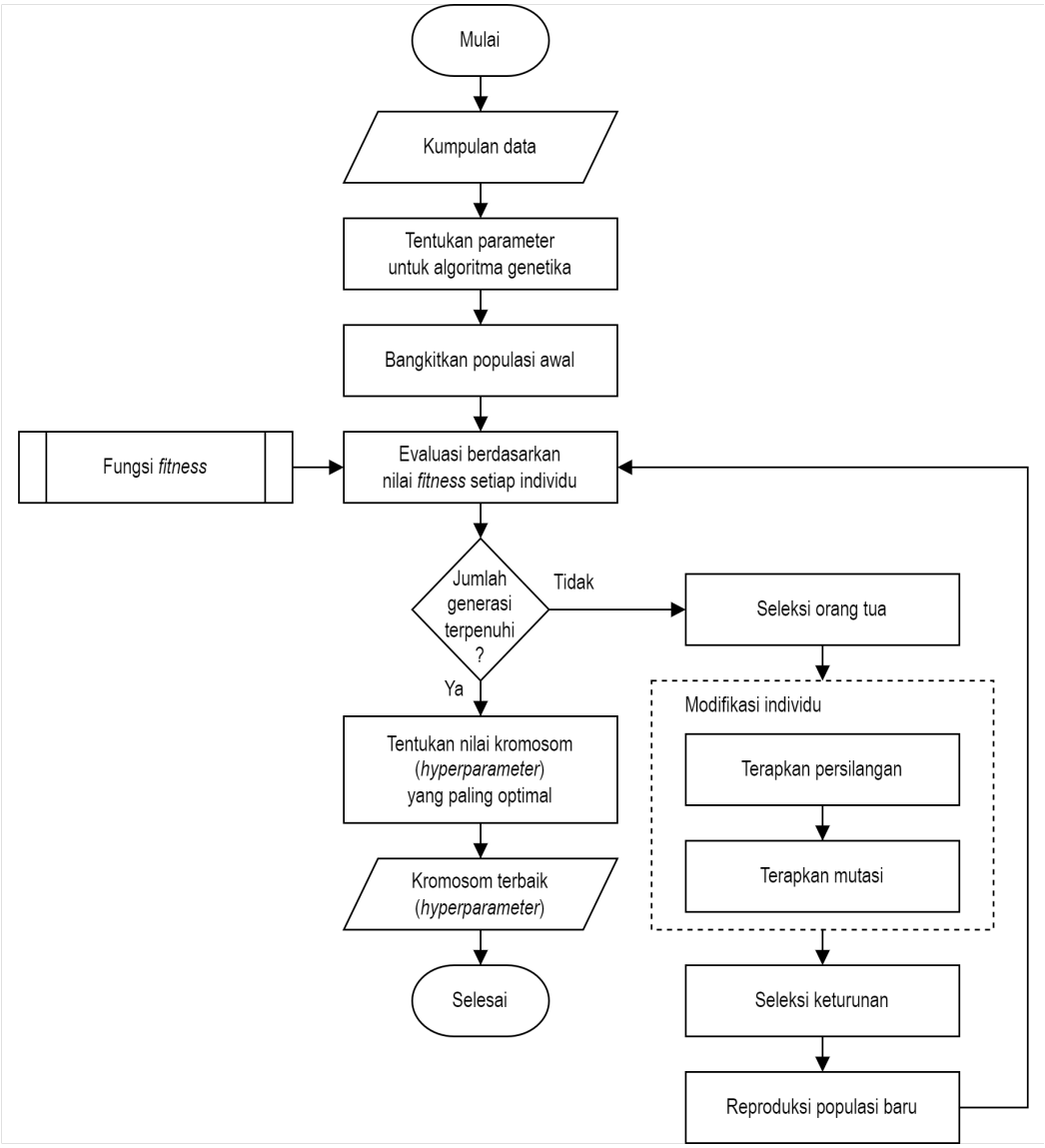
Ruang pencarian yang digunakan dalam proses penyetelan *hyperparameter* disajikan pada Tabel 3. Ruang pencarian untuk *hidden layer node* berada pada rentang nilai 1024 sampai 2048, *hidden layer* berada pada rentang nilai 1 sampai 5, *activation function* berada pada tiga pilihan fungsi aktivasi (ReLu, Sigmoid, dan TanH), *learning rate* berada pada rentang nilai 0.0001 sampai 0.01, dan *dropout rate* berada pada rentang nilai 0.1 sampai 0.5.

Tabel 3 Ruang pencarian pada penyetelan *hyperparameter*

Hyperparameter	Ruang Pencarian
Hidden layer node	1024-2048
Hidden layer	1-5
Activation function	ReLu, Sigmoid, TanH
Learning rate	0.0001-0.01
Dropout rate	0.1-0.5

Alur kerja penyetelan *hyperparameter* SAE-DNN menggunakan algoritma genetika disajikan pada Gambar 3. Proses awal dimulai dengan menentukan parameter operasi yang meliputi ukuran populasi, jumlah generasi, probabilitas persilangan, probabilitas mutasi, dan fungsi kriteria kinerja seperti terlihat pada Tabel 4. Setelah menyusun kromosom dengan membangkitkan 15 individu yang terdiri atas kombinasi masing-masing *hyperparameter*, dilakukan evaluasi individu berdasarkan fungsi kriteria kinerja. Seleksi dilakukan terhadap 465 kromosom dari 465 induk yang berasal dari populasi.

Setelah mendapatkan hasil seleksi populasi awal, reproduksi persilangan dan mutasi dilakukan berdasarkan bilangan acak yang dibangkitkan kurang dari nilai probabilitas yang telah ditentukan. Proses seleksi, persilangan, dan mutasi diulangi hingga salah satu kriteria penghentian (*stopping criteria*) terpenuhi. Pemilihan kromosom terbaik dilakukan dengan membandingkan nilai kriteria kinerja di setiap generasi.



Gambar 3 Alur kerja penyetelan *hyperparameter* menggunakan algoritma genetika

Tabel 4 Parameter yang digunakan dalam algoritma genetika

Parameter	Nilai
Ukuran populasi	15
Jumlah generasi	30
Probabilitas persilangan	0.7
Probabilitas mutasi	0.05

Pengujian dan Evaluasi

Model diuji menggunakan data senyawa herbal yang berasal dari situs HerbalDB. Data ini berisi nama dan *fingerprint* dari setiap senyawa. Kombinasi pasangan senyawa herbal dan protein dibuat dengan menggabungkan *fingerprint* senyawa herbal dan fitur protein yang terpilih sehingga terbentuk vektor fitur yang akan digunakan sebagai data uji.

Hasil pengujian dievaluasi menggunakan *iterative stratification*, yang merupakan modifikasi dari *k-fold cross-validation* dengan tujuan untuk menyeimbangkan jumlah kombinasi label dari data multilabel pada setiap *fold* (Sechidis *et al.* 2011; Szymański dan Kajdanowicz 2017). Model dengan nilai metrik tertinggi dipilih sebagai model terbaik dalam memprediksi permasalahan multilabel.

Pengukuran evaluasi yang digunakan meliputi:

1. *Accuracy* menyatakan persentase data uji yang diprediksi dengan benar (Sokolova dan Lapalme 2009), yang dapat dihitung dengan menggunakan Persamaan 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

2. *Recall*, disebut juga *True Positive Rate* (TPR) atau *sensitivity*, merupakan ketepatan model dalam memprediksi kelas positif. *Recall* menyatakan rasio kelas positif yang diprediksi benar dari kelas positif (Powers 2020), yang dihitung menggunakan Persamaan 2.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

3. *Precision* merupakan ketepatan model dalam prediksi positifnya. *Precision* menyatakan rasio kelas positif yang diprediksi benar dari semua prediksi positif (Powers 2020), yang dihitung menggunakan Persamaan 3.

$$Precision = \frac{TP}{TP + FN} \quad (3)$$

4. *F-Measure*, disebut juga *F-Score* atau *F1-Score*, menyatakan performa kelas minoritas secara menyeluruh yang mengatasi masalah perbandingan kualitas saat TP dan FP meningkat secara bersamaan. *F-Measure* memperhatikan nilai *precision* dan *recall* untuk mengukur performa kelas minoritas secara keseluruhan (Lin *et al.* 2014), yang dihitung menggunakan Persamaan 4.

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Prediksi dan Analisis

Prediksi interaksi senyawa herbal terhadap himpunan protein target dilakukan menggunakan model SAE-DNN dengan *hyperparameter* optimal yang diperoleh dari hasil evaluasi model. Data senyawa herbal diperoleh dari situs HerbalDB, yang berisi nama senyawa serta *circular fingerprint* dari setiap senyawa. *Circular fingerprint* ini digunakan sebagai representasi struktur kimia dari senyawa herbal dalam proses prediksi.

Setelah model memprediksi interaksi antara senyawa herbal dan protein target, langkah selanjutnya adalah mencari tanaman herbal yang mengandung senyawa-senyawa hasil prediksi tersebut. Proses pencarian ini dilakukan menggunakan situs KnapSack, yang menyediakan informasi tentang tanaman herbal dan kandungan kimiawinya (Nakamura *et al.* 2013). Dengan demikian, dapat diidentifikasi tanaman herbal yang memiliki potensi sebagai sumber senyawa aktif yang berinteraksi dengan protein target terkait COVID-19 (Chen *et al.* 2006).

HASIL DAN PEMBAHASAN

Pemodelan SAE-DNN-GA

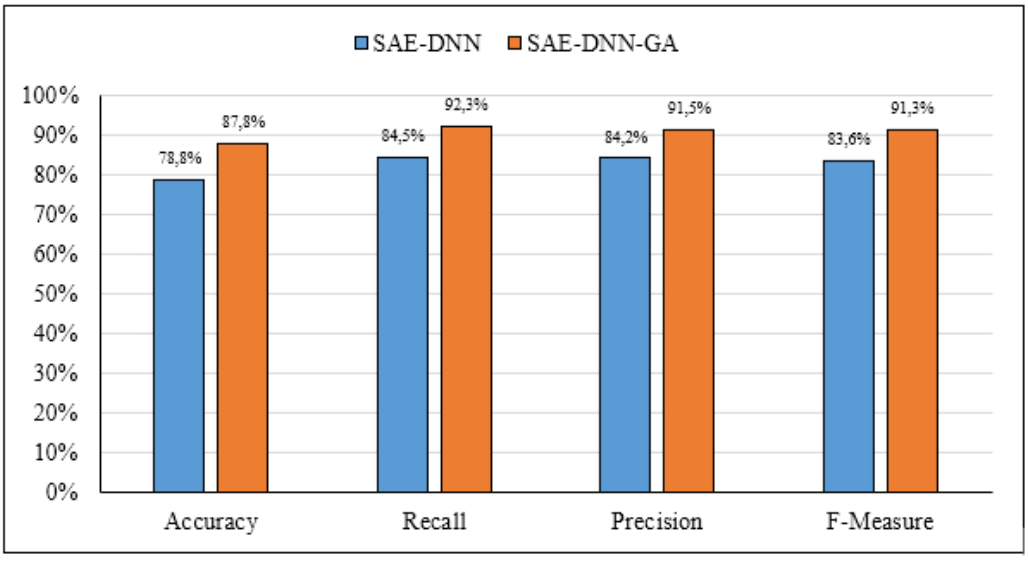
Hyperparameter bawaan seperti *hidden layer node* = 1024, *hidden layer* = 3, *activation function* = ReLu, *learning rate* = 0.01, dan *dropout rate* = 0.5 digunakan untuk membangun dan melatih model SAE-DNN. Penyetelan *hyperparameter* dilakukan menggunakan algoritma genetika, kemudian performa model dengan penyetelan *hyperparameter* (SAE-DNN-GA) dibandingkan dengan model tanpa penyetelan *hyperparameter* (SAE-DNN). Hasil penyetelan *hyperparameter* dapat dilihat pada Tabel 5.

Tabel 5 Hasil penyetelan *hyperparameter*

Hyperparameter	Algoritma Genetika
Hidden layer node	1700
Hidden layer	2
Activation function	ReLu
Learning rate	0.0001
Dropout rate	0.2

Evaluasi Model

Nilai rata-rata setiap metrik disajikan untuk melihat performa model. Gambar 4 menunjukkan perbandingan performa antara model SAE-DNN dan model SAE-DNN-GA. Model SAE-DNN-GA memperlihatkan performa yang paling tinggi dengan *accuracy* sebesar 87.8%, *recall* sebesar 92.3%, *precision* sebesar 91.5%, dan *f-measure* sebesar 91.3%. Sebaliknya, model SAE-DNN menunjukkan performa yang lebih rendah dengan *accuracy* sebesar 78.8%, *recall* sebesar 84.4%, *precision* sebesar 84.2%, dan *f-measure* sebesar 83.6%. Hal ini menunjukkan bahwa model SAE-DNN-GA lebih unggul dalam memprediksi kelas multilabel (dengan *accuracy* yang lebih tinggi), lebih baik dalam memprediksi kelas positif dari setiap label (dengan *recall* yang lebih tinggi), lebih akurat dalam prediksi positif tiap label (dengan *precision* yang lebih tinggi), serta lebih efektif dalam mengenali kelas minoritas (dengan *f-measure* yang lebih tinggi).



Gambar 4 Perbandingan performa SAE-DNN dan SAE-DNN-GA

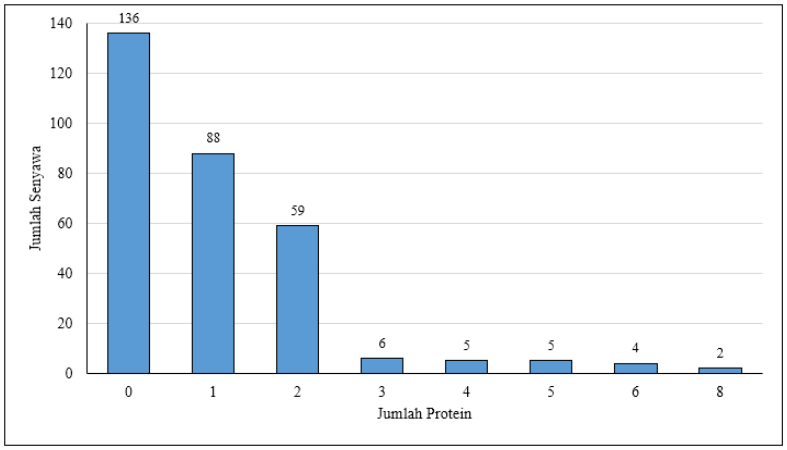
Performa SAE-DNN-GA yang lebih baik dibanding SAE-DNN dipengaruhi oleh beberapa faktor di bawah ini:

1. Algoritma genetika cenderung lebih eksploratif karena adanya operasi mutasi dan persilangan. Hal ini mengurangi peluang terjebak pada optima lokal dan meningkatkan kemungkinan menemukan solusi terbaik.
2. Algoritma genetika bekerja dengan baik pada ruang pencarian berdimensi tinggi karena sifatnya yang berbasis populasi dan operasinya yang luas.
3. Pada masalah yang sangat kompleks, algoritma genetika lebih efektif karena mampu melakukan lebih banyak evaluasi dan eksplorasi.

Prediksi Senyawa Herbal

Prediksi senyawa herbal dilakukan menggunakan model SAE-DNN-GA dengan kumpulan data senyawa herbal sebanyak 305 senyawa. Hasil prediksi menunjukkan 169 senyawa memiliki interaksi dengan protein COVID-19. Senyawa yang berinteraksi dengan satu protein berjumlah 88 dan 81 sisanya berinteraksi dengan lebih dari satu protein. Gambar 5 menyajikan secara lengkap jumlah senyawa herbal yang berinteraksi dengan protein COVID-19 hasil prediksi SAE-DNN-GA.

Nilai relevansi dari protein yang berinteraksi dengan senyawa menjadi indikator dalam menentukan senyawa potensial untuk COVID-19. Semakin tinggi nilai relevansi, maka semakin relevan protein tersebut dengan suatu penyakit. Tabel 6 menyajikan sepuluh interaksi senyawa-protein hasil prediksi SAE-DNN-GA dengan nilai relevansi tertinggi. Hasil pada Tabel 6 menunjukkan bahwa dua protein dengan nilai relevansi tertinggi, yaitu INS (7.094) dan ALB (3.178) berinteraksi dengan sepuluh senyawa. Insulin sangat penting untuk mengelola kadar glukosa darah pada pasien COVID-19, terutama mereka yang menderita diabetes (Yuan *et al.* 2023). Penelitian Mohamed *et al.* (2023) dan Kheir *et al.* (2021) menunjukkan bahwa protein INS dan ALB dapat berfungsi sebagai penanda prognosis untuk tingkat keparahan COVID-19.



Gambar 5 Jumlah senyawa herbal yang berinteraksi dengan himpunan protein

Tabel 6 Sepuluh interaksi senyawa-protein dengan nilai relevansi tertinggi

Senyawa	Protein	Probabilitas	Nilai Relevansi	Tanaman Herbal
Palmitic acid	INS	53.9%	7.094	*Mangifera indica* (Mangga) Annona squamosa (Srikaya)
Stearic acid	INS	53.9%	7.094	
Myristic acid	INS	53.9%	7.094	
Octanoic acid	INS	53.9%	7.094	*Mangifera indica* (Mangga) *Carica papaya* (Pepaya)
Nonanoic acid	INS	53.9%	7.094	
alpha-Terpinene	ALB	88.1%	3.178	
Caproic acid	ALB	78.5%	3.178	
Heptadecane	ALB	86.8%	3.178	*Mangifera indica* (Mangga)
Estradiol	ALB	99.9%	3.178	*Punica granatum* (Delima)
Hesperidin	ALB	72.6%	3.178	*Carica papaya* (Pepaya)

SIMPULAN

Model SAE-DNN-GA yang diusulkan mampu memberikan performa yang baik dalam prediksi senyawa herbal potensial untuk penyakit COVID-19. Performa ini lebih tinggi dibandingkan dengan model SAE-DNN, dengan *accuracy* sebesar 87.8%, *recall* sebesar 92.3%, *precision* sebesar 91.5%, dan *f-measure* sebesar 91.3%. Prediksi senyawa herbal menggunakan SAE-DNN-GA menghasilkan sepuluh senyawa yang berinteraksi dengan dua protein bernilai relevansi tinggi, yaitu protein INS (7.094) dan ALB (3.178).

DAFTAR PUSTAKA

Bahi M, Batouche M. 2018. Drug-Target Interaction Prediction in Drug Repositioning Based on Deep Semi-Supervised Learning. Di dalam: Amine A, Mouhoub M, Ait Mohamed O, Djebbar B, editor. *6th Computational Intelligence and Its Applications (CIIA) 2018*;. Cham: Springer International Publishing. (IFIP Advances in Information and Communication Technology). hlm. 302–313.

Berdigaliyev N, Aljofan M. 2020. An Overview of Drug Discovery and Development. *Future Med. Chem.* 12(10):939–947.doi:10.4155/fmc-2019-0307.

Chen X, Zhou H, Liu YB, Wang JF, Li H, Ung CY, Han LY, Cao ZW, Chen YZ. 2006. Database of Traditional Chinese Medicine And Its Application to Studies of Mechanism and to Prescription Validation. *Br. J. Pharmacol.* 149(8):1092–1103.doi:10.1038/sj.bjp.0706945.

Chu Y, Shan X, Chen T, Jiang M, Wang Y, Wang Q, Salahub DR, Xiong Y, Wei D-Q. 2021. DTI-MLCD: Predicting Drug-Target Interactions Using Multi-Label Learning with Community Detection Method. *Brief. Bioinform.* 22(3):bbaa205.doi:10.1093/bib/bbaa205.

Ekor M. 2014. The Growing Use of Herbal Medicines: Issues Relating to Adverse Reactions and Challenges In Monitoring Safety. *Front. Pharmacol.* 4(177):1–10.doi:10.3389/fphar.2013.00177.

Erhan D, Bengio Y, Courville A, Manzagol P-A, Vincent P, Bengio S. 2010. Why Does Unsupervised Pre-Training Help Deep Learning? *J. Mach. Learn. Res.* 11(19):625–660.

Fadli A, Kusuma WA, Annisa, Batubara I, Heryanto R. 2021. Screening of Potential Indonesia Herbal Compounds Based on Multi-Label Classification for 2019 Coronavirus Disease. *Big Data Cogn. Comput.* 5(4):75.doi:10.3390/bdcc5040075.

Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Leontovich AM, Neuman BW, *et al.* 2020. The Species Severe Acute Respiratory Syndrome-related Coronavirus: Classifying 2019-nCoV and Naming It SARS-CoV-2. *Nat. Microbiol.* 5(4):536–544.doi:10.1038/s41564-020-0695-z.

Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, Gilson MK, Bourne PE, Preissner R. 2012. Supertarget Goes Quantitative: Update on Drug-Target Interactions. *Nucleic Acids Res.* 40(Database issue):D1113-1117.doi:10.1093/nar/gkr912.

Holland JH. 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. Reprint. Massachussetts (MA): The MIT Press.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, *et al.* 2020. Clinical Features of Patients Infected with 2019 Novel Coronavirus In Wuhan, China. *Lancet Lond. Engl.* 395(10223):497–506.doi:10.1016/S0140-6736(20)30183-5.

James G, Witten D, Hastie T, Tibshirani R. 2013. An Introduction to Statictical Learning with Application in R. Ed ke-1. New York (NY): Springer. (Springer Text in Statictic).

Kheir M, Saleem F, Wang C, Man A, Chua J. 2021. Higher Albumin Levels on Admission Predict Better Prognosis in Patients with Confirmed COVID-19. *PLOS ONE*. 16(3):e0248358. doi:10.1371/journal.pone.0248358.

Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, Pon A, Cox J, Chin NEL, Strawbridge SA, *et al.* 2024. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res.* 52(D1):D1265–D1275.doi:10.1093/nar/gkad976.

Lee P-I, Hsueh P-R. 2020. Emerging Threats from Zoonotic Coronaviruses-from SARS and MERS to 2019-nCoV. *J. Microbiol. Immunol. Infect. Wei Mian Yu Gan Ran Za Zhi*. 53(3):365–367.doi:10.1016/j.jmii.2020.02.001.

Lin K-B, Weng W, Lai RK, Lu P. 2014. Imbalance Data Classification Algorithm Based on SVM and Clustering Function. *2014 9th International Conference on Computer Science & Education*;. New Jersey, United States. hlm. 544–548.

Maxwell A, Li R, Yang B, Weng H, Ou A, Hong H, Zhou Z, Gong P, Zhang C. 2017. Deep Learning Architectures for Multi-label Classification of Intelligent Health Risk Prediction. *BMC Bioinformatics*. 18(14):523.doi:10.1186/s12859-017-1898-z.

Mei S, Zhang K. 2019. A multi-label learning framework for drug repurposing. *Pharmaceutics*. 11(9):466.doi:10.3390/pharmaceutics11090466.

Mohamed AA, Nour AA, Mosbah NM, Wahba ASM, Esmail OE, Eysa B, Heiba A, Samir HH, El-Kassas AA, Adroase AS, *et al.* 2023. Evaluation of Circulating Insulin-Like Growth Factor-1, Heart-Type Fatty Acid-binding Protein, and Endotrophin Levels as Prognostic Markers of COVID-19 Infection Severity. *Virol. J.* 20(1):94.doi:10.1186/s12985-023-02057-4.

Nakamura K, Shimura N, Otabe Y, Hirai-Morita A, Nakamura Y, Ono N, Ul-Amin MA, Kanaya S. 2013. KNApSAcK-3D: A Three-Dimensional Structure Database of Plant Metabolites. *Plant Cell Physiol.* 54(2):e4.doi:10.1093/pcp/pcs186.

Ororbia II AG, Giles CL, Reitter D. 2016. Online Semi-Supervised Learning with Deep Hybrid Boltzmann Machines and Denoising Autoencoders. .doi:10.48550/arXiv.1511.06964.

Powers DMW. 2020. Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, Markedness, and Correlation. .doi:10.48550/arXiv.2010.16061. [diunduh 2024 Jul 11]. Tersedia pada: http://arxiv.org/abs/2010.16061

Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, *et al.* 2019. Drug Repurposing: Progress, Challenges and Recommendations. *Nat. Rev. Drug Discov.* 18(1):41–58.doi:10.1038/nrd.2018.168.

Rivera G, Cisneros L, Sánchez-Solís P, Rangel-Valdez N, Rodas-Osollo J. 2020. Genetic Algorithm for Scheduling Optimization Considering Heterogeneous Containers: a Real-World Case Study. *Axioms*. 9(1):27.doi:10.3390/axioms9010027.

Roubos H, Setnes M. 2000. Compact Fuzzy Models Through Complexity Reduction and Evolutionary Optimization. Vol. 2. *Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000 (Cat. No.00CH37063)*;. New Jersey (NJ): Institute of Electrical and Electronics Engineers Inc. hlm. 762–767.

Salim Z, Munadi E. 2017. Info Komoditi Tanaman Obat. Pertama. Jakarta: BPPP Kementerian Perdagangan Republik Indonesia.

Sechidis K, Tsoumakas G, Vlahavas I. 2011. On The Stratification of Multi-Label Data. Di dalam: Gunopulos D, Hofmann T, Malerba D, Vazirgiannis M, editor. *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2011*;. Berlin, Germany: Springer Berlin Heidelberg. (ECML PKDD 2011: Machine Learning and Knowledge Discovery in Databases). hlm. 145–158.

Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N. 2016. Taking The Human Out of The Loop: a Review of Bayesian Optimization. *Proc. IEEE*. 104(1):148–175.doi:10.1109/JPROC.2015.2494218.

Sokolova M, Lapalme G. 2009. A Systematic nalysis of pPerformance Measures for Classification Tasks. *Inf. Process. Manag.* 45(4):427–437.doi:10.1016/j.ipm.2009.03.002.

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, *et al.* 2016. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinforma.* 54(1):1.30.1-1.30.33.doi:10.1002/cpbi.5.

Syahdi R, Iqbal J, Munim A, Yanuar A. 2019. HerbalDB 2.0: Optimization of Construction of Three-Dimensional Chemical Compound Structures to Update Indonesian Medicinal Plant Database. *Pharmacogn. J.* 11(6):1189–1194.doi:10.5530/pj.2019.11.184.

Szymański P, Kajdanowicz T. 2017. A Network Perspective on Stratification of Multi-Label Data. .doi:10.48550/arXiv.1704.08756.

Wicaksono AS, Supianto AA. 2018. Hyperparameter Optimization Using Genetic Algorithm on Machine Learning Methods for Online News Popularity Prediction. *Int. J. Adv. Comput. Sci. Appl. IJACSA*. 9(12):263–267.doi:10.14569/IJACSA.2018.091238.

Yuan Y, Jiao B, Qu L, Yang D, Liu R. 2023. The Development of COVID-19 Treatment. *Front. Immunol.* 14:1–13.doi:10.3389/fimmu.2023.1125246.

Zhao J, Cao Z. 2015. A Label Extended Semi-Supervised Learning Method for Drug-Target Interaction Prediction. *2015 International Conference on Automation, Mechanical Control and Computational Engineering*;. Dordrecht: Atlantis Press. hlm. 1635–1640.