Optimalisasi Seleksi Fitur pada Dataset Penyakit Jantung Menggunakan ACO-PSO dan Machine Learning

Zaima Firoos Likan
Department of Computer Science
IPB University
Bogor, Indonesia
firooslikan@gmail.com

Aleeka Kiana S
Department of Computer Science
IPB University
Bogor, Indonesia
aleeka31kiana@apps.ipb.ac.id

Sulthan Farras Razin
Department of Computer Science
IPB University
Bogor, Indonesia
sulthanrazin@apps.ipb.ac.id

Allyvia Adzhani Saputra

Department of Computer Science

IPB University

Bogor, Indonesia

adzhaniallyvia@apps.ipb.ac.id

Muh Farid FB
Department of Computer Science
IPB University
Bogor, Indonesia
muhfaridfb23@gmail.com

Abstract

Penyakit kardiovaskular merupakan penyebab utama kematian secara global, dengan kompleksitas diagnosis yang tinggi akibat interaksi berbagai faktor risiko seperti tekanan darah, kolesterol, dan gaya hidup. Tugas akhir ini mengembangkan sistem prediksi penyakit jantung menggunakan pendekatan hybrid swarm intelligence yang menggabungkan algoritma Ant Colony Optimization (ACO) dan Particle Swarm Optimization (PSO). Dataset Heart Disease dari UCI digunakan dengan total 303 sampel dan 14 atribut. Proses seleksi fitur dilakukan dengan pendekatan ACO-PSO untuk meningkatkan efisiensi klasifikasi. Model klasifikasi dibangun menggunakan algoritma Random Forest dan XGBoost, serta dioptimasi menggunakan GridSearchCV. Seluruh implementasi dilakukan dengan Python di Google Colaboratory. Evaluasi performa dilakukan dengan metrik akurasi, presisi, recall, dan F1-score. Hasil menunjukkan bahwa model Random Forest dengan tuning menghasilkan akurasi 86,68%, lebih unggul dibandingkan XGBoost dengan akurasi 86,3%. Temuan ini menunjukkan bahwa integrasi ACO-PSO efektif dalam meningkatkan performa klasifikasi penyakit jantung.

Keywords: Heart Disease, Swarm Intelligence, ACO-PSO, Random Forest, XGBoost

I. LATAR BELAKANG

Penyakit kardiovaskular masih menjadi penyebab kematian utama secara global, dengan menyumbang 17,9 juta kasus kematian setiap tahun atau setara 31% dari seluruh kematian di dunia (UK Government 2019). Jumlah ini meningkat drastis dari 12,1 juta kasus pada tahun 1990 menjadi 20,5 juta pada tahun 2021 (World Heart Federation 2023). Di Indonesia, angka kematian akibat penyakit kardiovaskular mencapai 765.660 kasus pada tahun 2021 dan menempatkan Indonesia di peringkat 20% teratas negara dengan tingkat kematian tertinggi akibat penyakit jantung (World Heart Federation 2023). Berdasarkan data Survei Kesehatan Indonesia (2024), prevalensi penyakit jantung nasional mencapai 0,85%, dengan provinsi Yogyakarta memiliki prevalensi tertinggi sebesar 1,67%.

Diagnosis penyakit jantung terbilang cukup kompleks karena melibatkan banyak faktor seperti tekanan darah tinggi, kadar kolesterol, usia, serta pola hidup sehingga memerlukan sistem yang mampu memproses dan menganalisis data secara akurat dan efisien (American Heart Association 2022). Namun, metode diagnosis tradisional masih memiliki keterbatasan dalam hal efisiensi dan akurasi, serta rentan terhadap human error karena prosesnya bergantung pada interpretasi manual (Springer 2022).

Untuk mengatasi tantangan tersebut, pendekatan berbasis *Swarm Intelligence* (SI) mulai banyak digunakan. SI merupakan teknik komputasi yang meniru perilaku kolektif makhluk sosial seperti semut, lebah, dan burung untuk menyelesaikan masalah optimasi yang kompleks. Dua algoritma SI paling populer adalah *Particle Swarm Optimization* (PSO) dan *Ant Colony Optimization* (ACO). PSO meniru pola gerakan burung dalam kawanan dan telah berhasil digunakan untuk prediksi penyakit jantung dengan akurasi mencapai 84,61% dalam model MLP-PSO hybrid (Bataineh & Manacek 2022). Sedangkan ACO, yang meniru mekanisme pencarian makanan semut melalui jejak feromon, sangat efektif dalam menyelesaikan permasalahan optimasi kombinatorial dan seleksi fitur.

Keduanya memiliki keunggulan tersendiri: PSO unggul dalam kecepatan konvergensi namun rawan terjebak di local optima, sementara ACO unggul dalam eksplorasi solusi namun lambat konvergen. Oleh karena itu, dilakukan pendekatan gabungan atau fusion ACO-PSO untuk mengatasi kelemahan masing-masing dan mengoptimalkan hasil pencarian solusi (Jiang & Ma 2020). Menurut Khourdifi dan Bahaj (2019), model PA-RF (PSO-ACO + Random Forest) yang mereka usulkan berhasil mencapai akurasi 99,6%, jauh lebih tinggi dibandingkan SVM dan *Naive Bayes* yang hanya mencapai 83,6%.

Lebih lanjut, pendekatan *hybrid swarm* ini juga mulai dikombinasikan dengan teknik *deep learning* seperti LSTM dan Bi-LSTM untuk prediksi penyakit jantung. Hasil penelitian Avni *et al.* (2024) menunjukkan bahwa kombinasi *swarm intelligence* dan *deep learning* dapat mencapai akurasi hingga 97,86%. Hal ini menunjukkan bahwa integrasi teknik optimasi berbasis SI dalam diagnosis penyakit jantung memiliki potensi besar untuk meningkatkan efisiensi, akurasi, serta mendukung pengembangan *healthcare 4.0* yang lebih cerdas, preventif, dan personal (Venkatesh *et al.* 2024).

Dengan mempertimbangkan tingginya tingkat kematian akibat penyakit jantung dan kompleksitas proses diagnosisnya, dibutuhkan pendekatan cerdas yang mampu mengoptimalkan analisis data medis secara efisien dan akurat. Penggabungan Ant Colony Optimization (ACO) dan Particle Swarm Optimization (PSO) dalam sistem hybrid swarm intelligence menjadi solusi potensial dalam meningkatkan performa klasifikasi dan seleksi fitur yang relevan untuk diagnosis penyakit jantung. Oleh karena itu, tugas akhir ini difokuskan pada pengembangan metode ACO-PSO hybrid untuk mengatasi keterbatasan model konvensional, mengevaluasi kinerja model klasifikasi seperti Random Forest, serta menyediakan dasar teknologi diagnosis berbasis kecerdasan buatan yang dapat diandalkan untuk mendukung pengambilan keputusan klinis yang lebih cepat dan tepat.

II. TINJAUAN PUSTAKA

A. Penyakit Jantung Kardiovaskular

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia dan Indonesia, dengan prevalensi yang terus meningkat setiap tahunnya. Kompleksitas penyakit ini disebabkan oleh berbagai faktor risiko seperti hipertensi, diabetes, kolesterol tinggi, dan gaya hidup tidak sehat yang saling berinteraksi (*American Heart Association* 2022). Diagnosis dini dan akurat sangat penting untuk menurunkan angka kematian dan meningkatkan kualitas hidup pasien. Namun, metode tradisional diagnosis penyakit jantung seringkali menghadapi keterbatasan dalam hal kecepatan, akurasi, dan kemampuan mengolah data medis yang besar dan kompleks (Venkatesh *et al.* 2024).

B. Metode Pemodelan

B.1 Particle Swarm Optimization (PSO)

PSO adalah algoritma optimasi berbasis swarm intelligence yang terinspirasi dari perilaku sosial kawanan burung atau ikan dalam mencari makanan. PSO menggunakan partikel-partikel yang bergerak dalam ruang solusi dengan memanfaatkan pengalaman terbaik individu dan kelompok untuk menemukan solusi optimal. Dalam konteks diagnosis penyakit jantung, PSO banyak digunakan untuk seleksi fitur dan optimasi parameter model machine learning, sehingga meningkatkan akurasi prediksi dan efisiensi komputasi. Keunggulan **PSO** adalah kemampuannya dalam konvergensi cepat dan adaptasi terhadap masalah non-linear serta multidimensi (Bataineh & Manacek 2022)

B.2 Ant Colony Optimization (ACO)

Ant Colony Optimization (ACO) meniru perilaku koloni semut yang menggunakan jejak feromon untuk menemukan jalur terpendek menuju sumber makanan. ACO efektif dalam eksplorasi ruang solusi yang luas dan telah diaplikasikan dalam berbagai masalah optimasi kombinatorial, termasuk seleksi fitur dalam diagnosis penyakit jantung. ACO mampu menghindari jebakan local optima dengan mekanisme jejak feromon yang dinamis, namun memiliki kelemahan dalam hal kecepatan konvergensi jika digunakan secara mandiri (Khourdifi & Bahaj 2019)

B.3 Machine Learning

Machine learning merupakan pendekatan yang sangat efektif dalam prediksi penyakit jantung dengan kemampuan mengolah data medis yang besar dan kompleks untuk menghasilkan model klasifikasi yang akurat dan dapat diandalkan. Berbagai algoritma machine learning seperti Support Vector Machine (SVM), Decision Tree, Random Forest, dan Deep Neural Network telah digunakan untuk menganalisis data kardiovaskular dan menunjukkan tingkat akurasi tinggi dalam mendeteksi risiko penyakit jantung. Studi terkini menunjukkan bahwa metode machine learning, terutama yang menggunakan teknik ensemble dan hybrid, mampu memberikan prediksi yang lebih baik dibandingkan model statistik tradisional, dengan kemampuan mengidentifikasi biomarker dan pola risiko secara lebih mendalam (Azka Mir et al., 2024).

B.4 XGboost

XGboost adalah metode *Gradient Boosting Decision Tree* yang dikembangkan untuk efisiensi dan performa tinggi. XGBoost menggabungkan teknik *boosting* untuk menyusun model dari beberapa pohon keputusan secara berurutan, dengan optimasi yang membuatnya 10 kali lebih cepat dibandingkan metode *boosting* lain. XGBoost sering menunjukkan performa terbaik dalam klasifikasi dan regresi, serta mampu menangani data tidak seimbang dengan baik (Givari 2022).

C. Integrasi Hybrid ACO-PSO

Penggabungan ACO dan PSO sebagai metode hybrid bertujuan menggabungkan keunggulan eksplorasi ACO dan kecepatan konvergensi PSO. Model hybrid ini memanfaatkan mekanisme jejak feromon ACO untuk eksplorasi solusi yang lebih luas dan kemampuan PSO dalam mempercepat konvergensi ke solusi optimal. Studi oleh Khourdifi dan Bahaj (2019) menunjukkan bahwa model hybrid PSO-ACO yang dioptimasi dengan Random Forest menghasilkan akurasi prediksi penyakit jantung sebesar 99,6%, jauh melampaui performa algoritma tunggal seperti SVM dan Naive Bayes. Pendekatan ini juga efektif dalam mengatasi masalah overfitting dan meningkatkan generalisasi model (Khourdifi & Bahaj 2019).

D. Penerapan *Swarm Intelligence* dalam Prediksi Penyakit Jantung Kardiovaskular

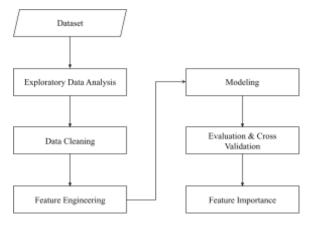
Swarm intelligence, khususnya gabungan ACO dan PSO, telah terbukti meningkatkan performa model machine learning dalam prediksi penyakit jantung dengan cara melakukan seleksi fitur yang optimal dan optimasi parameter. Pendekatan ini mampu mengelola data medis yang kompleks dan berukuran besar, sekaligus mengurangi risiko kesalahan diagnosis. Integrasi teknik swarm intelligence dengan metode deep learning seperti LSTM juga menunjukkan hasil yang menjanjikan dengan akurasi hingga 97,86%, membuka peluang pengembangan sistem diagnosis otomatis yang lebih canggih dan adaptif (Bataineh & Manacek 2022).

III. METODE

Tugas akhir ini menggunakan Heart Disease Dataset yang tersedia secara publik dan dapat diakses melalui situs

UCI Machine Learning Repository. Dataset ini merupakan salah satu benchmark dataset paling populer dalam penelitian diagnosis penyakit jantung berbasis machine learning. Dataset ini dilisensikan untuk penggunaan bebas dalam penelitian dan pendidikan, selama menyebutkan sumbernya.

Dataset terdiri dari 303 sampel dengan 14 kolom fitur, termasuk satu variabel target yang menunjukkan keberadaan penyakit jantung. Seluruh fitur terdiri dari kombinasi fitur numerik dan kategorikal, dengan sebagian nilai yang perlu diproses melalui normalisasi dan encoding agar sesuai dengan kebutuhan algoritma. Dataset ini mencakup indikator klinis penting yang sering digunakan dalam proses diagnosis penyakit jantung oleh tenaga medis. Tabel 1 berikut menjelaskan atribut-atribut utama dalam dataset yang digunakan dalam tahap preprocessing dan pemodelan klasifikasi. Proses implementasi ada pada Gambar 1.

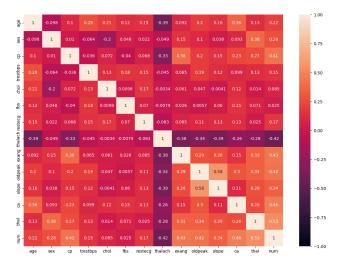


Gambar 1. Alur kerja implementasi tugas akhir

IV. HASIL DAN PEMBAHASAN

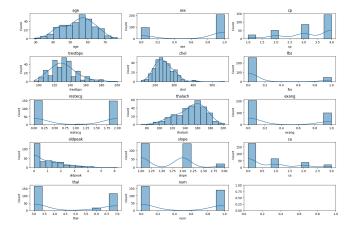
Pada tahap persiapan data, dilakukan pemeriksaan terhadap fitur target yakni 'num' dan diketahui bahwa kelas dari 'num' tidak mempresentasikan kondisi sebenarnya yang dijelaskan pada metadata. Untuk itu, row dengan fitur target bernilai >1 diubah menjadi 1. Hal ini bertujuan untuk menyesuaikan nilai target.

Setelahnya, dilakukan *exploratory data analysis* menggunakan *correlation diagram*, histogram, dan boxplot untuk memeriksa korelasi antar fitur, distribusi, dan identifikasi pencilan. Gambar 2 menunjukkan *correlation diagram* yang menunjukkan korelasi dan multikolinearitas antar fitur. Hasil *correlation diagram* menunjukkan bahwa tidak ada multikolinearitas dan ada beberapa fitur yang memiliki relasi paling tinggi dibanding fitur lainnya sehingga perlu dilakukan *feature selection* pada tahap selanjutnya untuk memvalidasi hasil diagram.



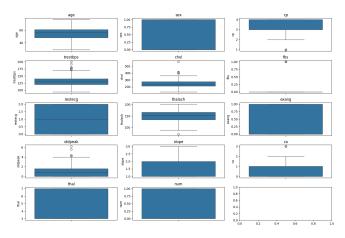
Gambar 2. Correlation diagram

Selain itu, Gambar 3 menunjukkan histogram setiap fitur. Hasil histogram menunjukkan bahwa distribusi dari setiap fitur memiliki distribusi yang baik, baik itu untuk fitur numerikal, maupun kategorikal. Namun, fitur kategorikal telah diproses dengan OneHotEncoding sehingga menjadi fitur numerikal int64 dan float64.



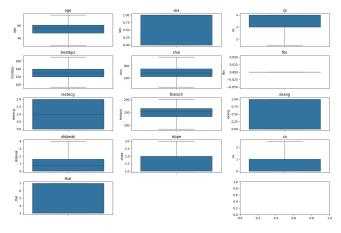
Gambar 3. Histogram fitur

Gambar 4 menunjukkan diagram boxplot untuk mengidentifikasi pencilan. Melalui diagram boxplot ini, dapat diidentifikasi beberapa fitur dengan beberapa nilai pencilan, yakni 'trestbps', 'chol', 'fbs', dan 'oldpeak'. Untuk itu, perlu dilakukan penanganan pencilan. Selanjutnya dilakukan penanganan nilai null values dengan menggunakan nilai modus karena jumlah row null values sangat sedikit.



Gambar 4. Boxplot diagram

Tahap selanjutnya setelah *data cleaning* adalah *feature engineering*. Pada tahap ini, dilakukan penanganan terhadap *outlier capping*. *Outlier capping* adalah teknik penanganan *outlier* dengan cara membatasi nilai ekstrem agar tetap berada dalam rentang yang wajar, tanpa menghapus data. Metode ini menggunakan batas atas dan bawah yang dihitung dari Interquartile Range (IQR), yaitu selisih antara kuartil ketiga (Q3) dan kuartil pertama (Q1). Nilai yang lebih kecil dari Q1 - 1.5×IQR akan diganti dengan nilai batas bawah tersebut, dan nilai yang lebih besar dari Q3 + 1.5×IQR akan diganti dengan batas atas. Dengan demikian, data ekstrem tidak dihapus melainkan "dicap" agar tidak terlalu memengaruhi analisis atau model.



Gambar 5. Boxplot diagram setelah outlier capping

Langkah terakhir sebelum pemodelan adalah *feature selection* sebagai proses memilih fitur yang paling relevan terhadap target. Metode yang digunakan untuk *feature selection* adalah uji Chi-Square, yang mengukur hubungan antara fitur kategorikal dan target. Dalam konteks ini, setiap fitur diuji apakah memiliki hubungan yang signifikan dengan target menggunakan nilai p-value. Jika p-value dari suatu fitur lebih kecil dari 0.05, maka fitur tersebut dianggap signifikan dan dipertahankan. Sebaliknya, fitur dengan p-value ≥ 0.05 dianggap tidak relevan dan dihapus dari data. Pada data ini, fitur yang signifikan adalah ['thal', 'exang', 'ca', 'oldpeak', 'slope', 'cp', 'sex', 'restecg'].

Selanjutnya, dilakukan splitting data training dan testing. Dataset hasil seleksi tersebut kemudian dibagi menjadi data latih (70%) dan data uji (30%) menggunakan train_test_split dengan random_state=42 untuk memastikan hasil yang reprodusibel. Dalam meningkatkan efisiensi model,

dilakukan *feature selection* lanjutan menggunakan metode kombinasi metaheuristik ACO-PSO (disimulasikan via *random subset selection* dengan *cross-validation* menggunakan *Random Forest*). Hasil seleksi terbaik, yaitu ['sex', 'exang', 'thal', 'restecg', 'slope', 'ca'] dengan *cross-validation accuracy*: 0.8679.

Model *random forest* dibangun dengan fitur hasil sebelumnya, model dievaluasi dengan data uji. Berikut merupakan hasil evaluasi *random forest* dengan akurasi 0.7692 dan *Precision, Recall, F1-Score* masing-masing kategori dapat dilihat pada tabel berikut:

Tabel 1. Precision, Recall, F1-Score Random Forest

Kelas	Precision	Recall	F1-Score
0	0.76	0.81	0.79
1	0.78	0.72	0.75
Accuracy	-	-	0.77

Selanjutnya, dilakukan *modelling* menggunakan XGBoost dengan parameter *default*. Model ini juga dibangun dengan fitur hasil ACO-PSO dan dievaluasi menggunakan data uji. Berikut merupakan hasil evaluasi XGBoost dengan akurasi 0.7363 dan *Precision, Recall, F1-Score* masing-masing kategori dapat dilihat pada tabel berikut:

Tabel 2. Precision, Recall, F1-Score Random Forest

Kelas	Precision	Recall	F1-Score
0	0.74	0.77	0.76
1	0.73	0.70	0.71
Accuracy	-	-	0.74

Hyperparameter tuning dilakukan untuk meningkatkan nilai accuracy pada data Heart Disease dengan menggunakan metode Random Forest dan XGBoost. Proses ini dilakukan dengan pendekatan GridSearchCV, dimana pendekatan ini mengevaluasi berbagai kombinasi dari parameter berdasarkan skor akurasi validasi silang atau cross-validation. Berikut merupakan hasil perbandingan nilai accuracy kedua model saat sebelum dan sesduah dilakukannya hyperparameter tuning.

Tabel 2. Hasil perbandingan nilai *accuracy*

Model	Accuracy		
	Sebelum tuning	Sesudah tuning	
Random Forest	0.77	0.8668	
XGBoost	0.74	0.8630	

Hasil tuning menunjukkan bahwa Metode *Random Forest* memiliki beberapa kombinasi parameter optimal dengan skor validasi silang sebesar 86.68%. Sedangkan pada Metode XGBoost, dengan menggunakan beberapa parameter optimal dihasilkan skor sebesar 86.3%. Peningkatan nilai *accuracy* ini menunjukkan bahwa hasil dari *hyperparameter* tuning dapat meningkatkan kemampuan generalisasi model.

Berdasarkan hasil evaluasi performa model klasifikasi pada kasus ini, *Random Forest* menunjukkan adanya performa yang lebih stabil, bahkan sebelum dilakukannya tuning. Dengan mempertimbangkan nilai *accuracy*, efisiensi dan kemudahan interpretasi, *Random Forest* merupakan pilihan yang lebih baik untuk diterapkan pada sistem prediksi penyakit jantung berdasarkan data historis.

V. Kesimpulan

Penyakit jantung yang kompleks memerlukan sistem prediktif yang mampu mengelola data medis secara akurat dan efisien. Penelitian ini membuktikan bahwa metode hybrid ACO-PSO efektif dalam proses seleksi fitur serta pengoptimalan parameter model klasifikasi. Dengan mengintegrasikan ACO dan PSO, sistem mampu menghindari jebakan local optima dan mempercepat konvergensi ke solusi terbaik. Model Random Forest menunjukkan performa terbaik dengan akurasi 86,68%, mengungguli model XGBoost yang hanya mencapai 86,3%. Proses ini dilakukan menggunakan Python dan library pendukung seperti Scikit-learn, Pandas, dan NumPy. menunjukkan bahwa pendekatan intelligence memberikan hasil yang kompetitif dan layak untuk dikembangkan lebih lanjut sebagai dasar sistem diagnosis penyakit jantung berbasis kecerdasan buatan yang lebih akurat, efisien, dan dapat diandalkan untuk mendukung pengambilan keputusan klinis.

REFERENCES

- Al Bataineh, A., & Manacek, S. (2022). MLP-PSO Hybrid Algorithm for Heart Disease Prediction. Journal of Personalized Medicine, 12(8), 1208. doi:10.3390/jpm12081208.
- [2] American Heart Association. (2022). 2022 Heart Disease & Stroke Statistical Update Fact Sheet Global Burden of Disease. Retrieved fromhttps://www.heart.org/-/media/PHD-Files-2/Science-News/2/202 2-Heart-and-Stroke-Stat-Update/2022-Stat-Update-factsheet-Global-Burden-of-Disease.pdf
- [3] Avni, Jaiswal, A., Ray, A., & Gola, S. (2024). Detection of Heart Failure Using Swarm Intelligence. CEUR Workshop Proceedings, 3682. Retrieved from https://ceur-ws.org/Vol-3682/Paper15.pdf
- [4] Azka, M., Rehman, A. U., Ali, T. M., Javaid, S., Almufareh, M. F., Humayun, M., & Shaheen, M. (2024). A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques. ESC Heart Failure. https://doi.org/10.1002/ehf2.14942.
- [5] Givari, M. R., Sulaeman, M. R., & Umaidah, Y. (2022). Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit. Nuansa Informatika, 16(1), 141-149. doi: 10.25134/nuansa.v16i1.5406.
- [6] Jiang, Y., & Ma, Y. (2020). Application of hybrid particle swarm and ant colony optimization algorithms to obtain the optimum homomorphic wavelet image fusion. Annals of Translational Medicine, 8(22), 1482. doi:10.21037/atm-20-7077
- [7] Springer. (2022). A systematic review on metaheuristic optimization techniques for feature selections in disease diagnosis: Open issues and

- challenges. Archives of Computational Methods in Engineering. doi:10.1007/s11831-022-09823-4
- [8] Survei Kesehatan Indonesia. (2024). 10 Provinsi dengan Prevalensi Penyakit Jantung Tertinggi. Retrieved from https://goodstats.id/article/10-provinsi-dengan-prevalensi-penyakit-ja ntung-tertinggi-r0vvq
- [9] UK Government. (2019). Health matters: Preventing cardiovascular disease. Retrieved from https://www.gov.uk/government/publications/health-matters-preventing-cardiovascular-disease/health-matters-preventing-cardiovascular-disease
- [10] Venkatesh, C., Prasad, V. S., Khan, M., Babu, J. C., & Dasu, M. V. (2024a). Exploration and comparison of the effectiveness of swarm intelligence algorithms for feature selection in cardiovascular disease prediction. Nature Scientific Reports. doi:10.1038/s41598-025-87598-0
- [11] Venkatesh, C., Prasad, V. S., Khan, M., Babu, J. C., & Dasu, M. V. (2024b). An automatic diagnostic model for the detection and classification of cardiovascular diseases based on swarm intelligence technique. Heliyon. doi:10.1016/j.heliyon.2024.e28567
- [12] World Heart Federation. (2023). Cardiovascular insights for Indonesia. World Heart Observatory. Retrieved from https://world-heart-federation.org/world-heart-observatory/countries/indonesia/