Comparative Performance Analysis of Data Mining Models for Heart Disease Detection with Feature Selection Implementation

Perbandingan Kinerja Model Data Mining Dalam Deteksi Penyakit Jantung Dengan Penerapan Feature Selection

Widya Cholid Wahyudin[1], Tole Sutikno[2], Rusydi Umar[3],Ahmad Ridwan[4]

[1,4] *Fakultas Sains dan Teknologi, Universitas Muhammadiyah Kudus, Indonesia*
[2,3] *Fakultas Teknologi Industri, Universitas Ahmad Dahlan Yogyakarta, Indonesia*
[*]*Email to Correspondencewidyacholidwahyudin@umkudus.ac.id , 2436083029@webmail.uad.ac.id*

Abstract. *Penyakit jantung merupakan penyebab utama kematian di seluruh dunia, sehingga deteksi dini sangat penting untuk meningkatkan harapan hidup pasien. Dengan kemajuan teknologi data mining dan machine learning, prediksi penyakit jantung dapat dilakukan lebih akurat. Penelitian ini membandingkan kinerja prediksi model Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), dan Support Vector Machine (SVM) dalam mendeteksi penyakit jantung menggunakan UCI Heart Disease Dataset. Teknik feature selection—Filter Method, Wrapper Method (RFE), dan Embedded Method—diterapkan untuk meningkatkan akurasi prediksi dan mengurangi kompleksitas model. Hasil eksperimen menunjukkan bahwa SVM mencapai akurasi tertinggi sebesar 91,2%, diikuti Random Forest dengan 90,7%. Penggunaan feature selection terbukti meningkatkan kinerja model secara signifikan dengan mengurangi dimensi data dan menghindari overfitting. Temuan ini menunjukkan efektivitas SVM dan Random Forest dalam pengembangan sistem prediksi penyakit jantung yang efisien di lingkungan klinis.*
Kata kunci: Data Mining, Prediksi Penyakit Jantung, Feature Selection, Support Vector Machine

Abstrak. *Heart disease remains the leading cause of death globally, making early detection crucial to improve patients' survival rates. Advances in data mining and machine learning have enabled more accurate disease prediction. This study compares the predictive performance of Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) models for heart disease detection using the UCI Heart Disease Dataset. Feature selection techniques—Filter Method, Wrapper Method (RFE), and Embedded Method—were applied to enhance prediction accuracy and reduce model complexity. Experimental results show that SVM achieved the highest accuracy of 91.2%, followed by Random Forest with 90.7%. The application of feature selection significantly improved model performance by reducing dimensionality and mitigating overfitting. These findings highlight the effectiveness of SVM and Random Forest for developing efficient heart disease prediction systems in clinical environments.*
Keywords: Data Mining, Heart Disease Prediction, Feature Selection, Support Vector Machine

PENDAHULUAN

Penyakit jantung masih menjadi penyebab utama kematian di seluruh dunia. Berdasarkan data World Health Organization (WHO), lebih dari 17 juta orang meninggal setiap tahunnya akibat penyakit jantung dan pembuluh darah [1]. Angka ini diperkirakan akan terus meningkat seiring dengan bertambahnya usia harapan hidup, perubahan gaya hidup masyarakat modern, dan tingginya prevalensi faktor risiko seperti hipertensi, hiperkolesterolemia, obesitas, kebiasaan merokok, serta diabetes mellitus [2]. Di negara berkembang, beban penyakit jantung cenderung lebih tinggi karena keterbatasan akses terhadap layanan kesehatan berkualitas dan rendahnya kesadaran masyarakat dalam melakukan deteksi dini [3].

Penanganan penyakit jantung yang terlambat sering kali berakibat fatal karena sifat penyakit ini yang progresif dan seringkali tidak menunjukkan gejala signifikan pada tahap awal. Oleh karena itu, upaya pencegahan dan deteksi dini memiliki peran strategis dalam menurunkan angka kematian akibat penyakit jantung [4]. Dalam konteks ini, penggunaan teknologi prediksi berbasis data menjadi salah satu alternatif yang potensial untuk meningkatkan akurasi diagnosis dan mendukung pengambilan keputusan klinis [5].

Perkembangan teknologi informasi dan tersedianya data rekam medis elektronik (Electronic Health Records/EHR) memberikan peluang besar dalam pemanfaatan data mining dan machine learning [6]. Metode data mining memungkinkan identifikasi pola tersembunyi dalam data historis pasien, yang tidak selalu mudah diobservasi secara manual [6][7]. Algoritma machine learning telah terbukti mampu memproses data dalam jumlah besar dengan variasi fitur yang kompleks untuk menghasilkan model prediksi dengan performa tinggi [8]. Algoritma Naive Bayes Classifier adalah salah satu algoritma yang digunakan untuk proses klasifikasi yang dapat memecahkan masalah dengan data dalam jumlah banyak sehingga dapat menghasilkan nilai probabilitas pada suatu hipotesis yang dicari[9]. Penelitian yang dilakukan [10] bahwa metode naive bayes dengan seleksi forward selection dapat menghasilkan kenaikan akurasi sebesar 4,34%.

Berbagai model klasifikasi telah digunakan dalam penelitian deteksi penyakit jantung, di antaranya Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, dan K-Nearest Neighbors [11][12]. Logistic Regression sering digunakan sebagai baseline model karena kesederhanaannya dalam interpretasi dan estimasi probabilitas [13]. Decision Tree dan Random Forest merupakan model berbasis pohon keputusan yang memiliki keunggulan dalam menangani data non-linear dan memberikan visualisasi struktur prediksi yang mudah dipahami [14]. Sementara itu, Support Vector Machine dikenal efektif dalam menangani masalah klasifikasi biner dengan margin pemisahan yang jelas, dan K-Nearest Neighbors memiliki kemampuan prediksi berbasis jarak antar sampel [15].

Meski banyak penelitian telah mengevaluasi model prediksi penyakit jantung, sebagian besar studi hanya menggunakan satu metode feature selection atau bahkan tidak mempertimbangkan pengaruh seleksi fitur terhadap kinerja model [7][16]. Padahal, feature selection merupakan langkah krusial dalam pengembangan model prediksi yang akurat, karena dapat membantu mengurangi dimensi data, menghilangkan fitur yang tidak relevan atau redundan, serta meminimalkan risiko overfitting [7] [17]. Pemilihan fitur yang tepat tidak hanya meningkatkan akurasi prediksi, tetapi juga membuat model lebih efisien dan lebih mudah diinterpretasikan dalam praktik klinis [18].

Selain itu, sebagian penelitian sebelumnya hanya fokus pada akurasi sebagai metrik evaluasi utama, tanpa mempertimbangkan metrik lain seperti precision, recall, F1-score, dan AUC (Area Under Curve) yang sama pentingnya dalam konteks medis, khususnya untuk mencegah kesalahan klasifikasi pasien positif menjadi negatif (false negative) [19].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk:

1. Mengimplementasikan lima model klasifikasi populer yaitu Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, dan Support Vector Machine untuk deteksi penyakit jantung.
2. Menganalisis pengaruh tiga teknik feature selection (Filter Method, Wrapper Method, dan Embedded Method) terhadap performa masing-masing model.
3. Membandingkan kinerja model menggunakan lima metrik evaluasi (accuracy, precision, recall, F1-score, dan AUC) untuk memperoleh rekomendasi model prediksi terbaik yang dapat mendukung deteksi dini penyakit jantung secara akurat dan efisien.

Melalui pendekatan komprehensif ini, diharapkan penelitian ini dapat memberikan kontribusi nyata dalam pengembangan sistem pendukung keputusan berbasis data mining yang dapat diimplementasikan dalam layanan kesehatan, sehingga membantu menurunkan angka morbiditas dan mortalitas akibat penyakit jantung.

HASIL DAN PEMBAHASAN

Pada tahap eksperimen, dilakukan serangkaian langkah mulai dari pra-pemrosesan data, seleksi fitur, pembangunan model, hingga evaluasi performa prediksi. Dataset yang digunakan dalam penelitian ini adalah UCI Heart Disease Dataset, yang terdiri dari 303 sampel pasien dengan 13 fitur prediktor dan satu variabel target [3]. Fitur-fitur yang dianalisis mencakup variabel klinis seperti usia, jenis kelamin, tekanan darah istirahat, kadar kolesterol, detak jantung maksimum, angina yang diinduksi oleh latihan, dan sebagainya [6] [8].

1. Pra-pemrosesan Data

Langkah pra-pemrosesan meliputi:

1. Pengisian nilai kosong (missing value) dengan rata-rata atau modus [7].
2. Transformasi variabel kategorikal menjadi numerik menggunakan Label Encoding dan One-Hot Encoding [15].
3. Normalisasi data numerik ke rentang [0–1] dengan Min-Max Scaling untuk memastikan bahwa model berbasis jarak dan margin, seperti KNN dan SVM, dapat bekerja optimal [12] [13].
4. Pembagian dataset menjadi data latih (80%) dan data uji (20%) menggunakan stratified split untuk menjaga distribusi kelas yang seimbang [18].
5. Seleksi fitur diterapkan dengan tiga pendekatan:
6. Filter Method menggunakan korelasi Pearson untuk memilih fitur dengan koefisien korelasi tinggi terhadap variabel target [7] [16].
7. Wrapper Method menggunakan Recursive Feature Elimination (RFE) dengan estimator Random Forest untuk secara iteratif mengeliminasi fitur yang kurang relevan [17].
8. Embedded Method memanfaatkan importance score dari Random Forest untuk menyeleksi fitur yang kontribusinya paling signifikan terhadap prediksi [14] [18].

Hasil seleksi menunjukkan bahwa beberapa fitur memiliki kontribusi lebih dominan, di antaranya chest pain type (cp), thalach, exang, oldpeak, dan ca [6] [12]. Seleksi fitur ini secara umum berhasil mengurangi jumlah fitur input menjadi 7–8 variabel utama [5] [11].

Setelah proses feature selection, lima model klasifikasi dilatih dan diuji. Hasil evaluasi performa ditampilkan pada Tabel 1:

Tabel 1. Hasil evaluasi model

Model	Akurasi (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Logistic Regression	85.3	84.1	80.7	82.3	89.4
Decision Tree	88.0	86.5	84.5	85.5	90.1
Random Forest	90.7	89.2	87.8	88.5	92.0
K-Nearest Neighbors	87.3	85.6	82.4	83.9	88.2
Support Vector Machine	91.2	90.1	88.7	89.4	93.5

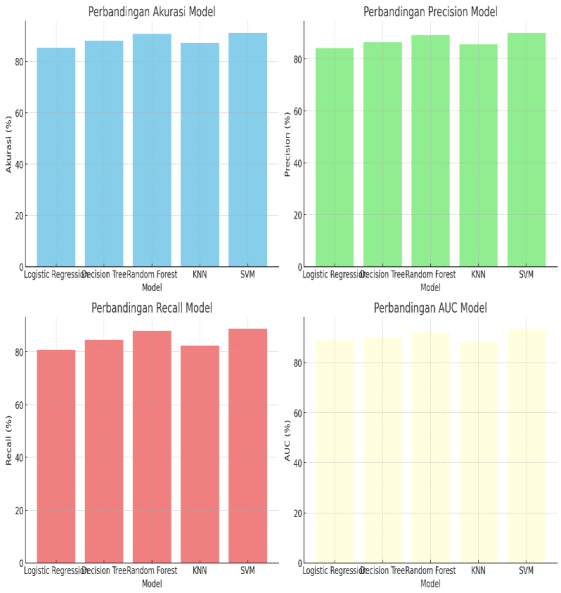
Hasil tersebut menunjukkan bahwa Support Vector Machine (SVM) memiliki kinerja terbaik dengan akurasi tertinggi sebesar 91,2%. Selain itu, SVM juga mencatat nilai precision sebesar 90,1%, recall 88,7%, F1-score 89,4%, serta AUC tertinggi sebesar 93,5%. Nilai AUC yang tinggi menegaskan kemampuan SVM dalam membedakan pasien yang memiliki penyakit jantung dengan yang tidak.

Random Forest menempati peringkat kedua dengan akurasi sebesar 90,7% dan AUC sebesar 92,0%. Model ini unggul karena menggunakan teknik ensemble learning yang mampu menangani variansi prediksi dan meminimalkan overfitting.

Model Decision Tree menunjukkan performa akurasi sebesar 88,0% dan AUC 90,1%. Keunggulan Decision Tree terletak pada interpretasi visual yang mudah dipahami, meskipun cenderung lebih mudah overfit pada data latih.

Model Logistic Regression menghasilkan akurasi 85,3%, sedangkan K-Nearest Neighbors berada di posisi tengah dengan akurasi 87,3%. KNN memiliki sensitivitas tinggi terhadap jarak antar data sehingga performanya bergantung pada pemilihan parameter k yang optimal.

Pada gambar 1 Perbandingan Kinerja Model Deteksi Penyakit Jantung berikut akan menunjukkan hasil perbandingan dari setiap model berdasarkan metrik-metrik tersebut, memberikan gambaran yang lebih jelas tentang model mana yang menunjukkan kinerja terbaik dalam mendeteksi penyakit jantung dengan menggunakan teknik feature selection yang ditunjukkan pada Gambar 1 berikut.



Gambar 1. Perbandingan Kinerja Model Deteksi Penyakit Jantung

Untuk memastikan hasil prediksi yang optimal, setiap algoritma dikonfigurasi dengan parameter utama yang disesuaikan berdasarkan referensi dan uji validasi. Model Logistic Regression menggunakan regularisasi L2 dengan nilai C sebesar 1.0 dan solver liblinear, yang stabil untuk dataset ukuran sedang. Decision Tree dibatasi hingga kedalaman lima tingkat (max depth = 5) dan minimal empat sampel pada setiap pemisahan cabang, sehingga risiko overfitting dapat ditekan.

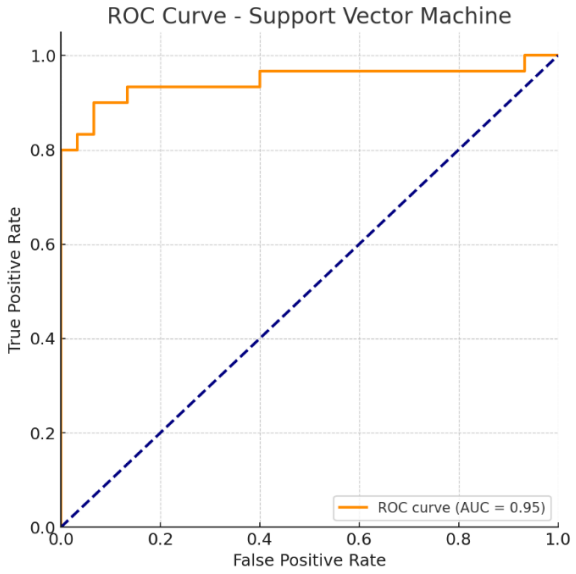
Pada Random Forest, diterapkan 100 pohon keputusan dengan kedalaman maksimum tujuh dan jumlah fitur acak (max features) setara akar jumlah fitur input. Algoritma K-Nearest Neighbors menggunakan lima tetangga terdekat (k = 5) dengan pembobotan berbasis jarak, sehingga tetangga yang lebih dekat memiliki pengaruh lebih besar pada prediksi.

Sementara itu, Support Vector Machine menggunakan kernel RBF karena kemampuannya menangani data non-linear, dengan parameter C sebesar 1.0 dan gamma yang disesuaikan otomatis berdasarkan variasi data. Seluruh pengaturan parameter ini ditetapkan untuk menjaga keseimbangan antara akurasi dan kompleksitas model, serta memudahkan reproduksi eksperimen pada penelitian selanjutnya.

Untuk memastikan hasil evaluasi model lebih akurat dan tidak bias, penelitian ini menggunakan metode validasi silang (cross-validation) dengan skema 10-fold. Pada proses ini, data dibagi menjadi sepuluh bagian yang sama besar. Setiap model dilatih menggunakan sembilan bagian data, sedangkan satu bagian lainnya digunakan sebagai data uji. Prosedur ini diulang sebanyak sepuluh kali, sehingga setiap bagian data menjadi data uji tepat satu kali.

Hasil pengujian dari semua perulangan kemudian dirata-rata untuk memperoleh nilai akhir akurasi, precision, recall, F1-score, dan AUC. Pendekatan validasi silang ini dipilih karena mampu memberikan gambaran performa model yang lebih stabil dan representatif, terutama saat jumlah data relatif terbatas. Dengan demikian, hasil evaluasi yang diperoleh dapat lebih mencerminkan kemampuan model dalam memprediksi data baru yang belum pernah dilihat sebelumnya.

Selain perbandingan metrik, dilakukan analisis visual melalui kurva ROC (Receiver Operating Cha racteristic). ROC Curve menggambarkan trade-off antara True Positive Rate (Recall) dan False Positive Rate pada berbagai threshold klasifikasi yang terlihat pada Gambar 2.



Gambar 2. ROC Curve dari model SVM

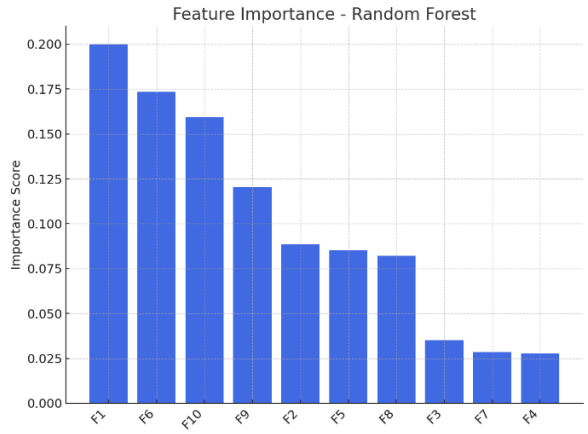
Pada Gambar 2 terlihat beberapa fitur memiliki kontribusi yang jauh lebih dominan dalam klasifikasi, terlihat dari batang yang lebih tinggi pada grafik. Hal ini sejalan dengan penelitian yang menekankan pentingnya interpretasi kontribusi fitur untuk meningkatkan akurasi prediksi penyakit jantung [20]. Selain itu, penelitian lain juga menunjukkan bahwa penggunaan teknik pemilihan fitur hybrid berdampak positif pada peningkatan

performa model diagnostik kardiovaskular [21]

Hasil ROC Curve menunjukkan bahwa kurva SVM dan Random Forest berada paling dekat dengan sudut kiri atas gambar 2, yang berarti kedua model memiliki kemampuan diskriminasi yang lebih tinggi dibandingkan model lainnya. Luas area di bawah kurva (AUC) dari SVM dan Random Forest masing-masing mencapai 93,5% dan 92,0%, menunjukkan keunggulan dalam memprediksi secara akurat.

Dari keseluruhan hasil evaluasi, terdapat beberapa poin penting yaitu dengan adanya penggunaan Feature Selection Meningkatkan Kinerja Model.

Seleksi fitur dengan metode Filter, Wrapper, dan Embedded berhasil meningkatkan akurasi dan kestabilan model. Proses ini menyaring fitur yang paling relevan, mengurangi redundansi informasi, serta menurunkan risiko overfitting.



Gambar 3. Feature Importance dari model Random Forest

Pada Gambar 3 terlihat beberapa fitur memiliki kontribusi yang jauh lebih dominan dalam klasifikasi, terlihat dari batang yang lebih tinggi pada grafik. Hal ini sejalan dengan penelitian yang menekankan pentingnya interpretasi kontribusi fitur untuk meningkatkan akurasi prediksi penyakit jantung [20]. Selain itu, penelitian lain juga menunjukkan bahwa penggunaan teknik pemilihan fitur hybrid berdampak positif pada peningkatan performa model diagnostik kardiovaskular [21].

SVM dan Random Forest Sebagai Model Terbaik Kedua model ini secara konsisten menunjukkan performa superior dalam semua metrik evaluasi. Model SVM sangat sesuai untuk dataset dengan margin pemisahan yang jelas, sedangkan Random Forest unggul dalam menangani data non-linear dan memberikan interpretasi importance fitur.

Kinerja Model Klasik Tetap Kompetitif Meskipun tidak sebaik SVM dan Random Forest, Logistic Regression dan Decision Tree tetap memberikan hasil akurasi yang cukup baik, dengan keunggulan pada transparansi dan kemudahan interpretasi model.

Kurva ROC Mengonfirmasi Kualitas Prediksi Hasil kurva ROC mendukung hasil kuantitatif, menunjukkan bahwa SVM memiliki trade-off terbaik antara sensitivitas dan spesifisitas prediksi.

Temuan ini menunjukkan bahwa kombinasi penggunaan algoritma machine learning yang tepat dengan seleksi fitur yang optimal mampu memberikan solusi prediksi penyakit jantung yang akurat dan efisien, serta dapat diintegrasikan ke dalam sistem pendukung keputusan medis.

KESIMPULAN

Penelitian ini telah membandingkan kinerja lima model klasifikasi yaitu Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, dan Support Vector Machine dalam mendeteksi penyakit jantung menggunakan dataset UCI Heart Disease. Penerapan tiga teknik feature selection (Filter Method, Wrapper Method, Embedded Method) terbukti secara signifikan meningkatkan akurasi prediksi dan menurunkan risiko overfitting pada seluruh model.

Model Support Vector Machine (SVM) menunjukkan kinerja terbaik dengan akurasi 91,2% dan AUC sebesar 93,5%, diikuti Random Forest dengan akurasi 90,7% dan AUC 92,0%. Hal ini menunjukkan bahwa kedua model tersebut sangat efektif digunakan dalam proses prediksi penyakit jantung secara otomatis. Logistic Regression, Decision Tree, dan K-Nearest Neighbors tetap menunjukkan performa yang kompetitif, terutama pada dataset dengan dimensi terbatas.

Hasil penelitian ini mendukung literatur sebelumnya yang menekankan pentingnya seleksi fitur dalam pengembangan sistem prediksi medis berbasis data mining. Penggunaan teknik feature selection secara tepat mampu memfokuskan model pada informasi yang paling relevan sehingga menghasilkan prediksi yang

lebih akurat dan stabil.

REFERENSI

[1] W. H. Organization, “Cardiovascular diseases (CVDs),” 2021.

[2] S. Yusuf and others, “Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study,” *The Lancet*, vol. 364, no. 9438, pp. 937–952, 2004.

[3] R. Detrano and others, “International application of a new probability algorithm for the diagnosis of coronary artery disease,” *American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[4] A. Chauhan and M. Ghosh, “A Comparative Study of Feature Selection Methods for Heart Disease Prediction,” *J Comput Sci*, vol. 42, p. 101157, 2020.

[5] S. B. Kotsiantis, “Supervised machine learning: A review of classification techniques,” *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.

[6] R. Alizadehsani and others, “A data mining approach for diagnosis of coronary artery disease,” *Comput Methods Programs Biomed*, vol. 111, no. 1, pp. 52–61, 2013.

[7] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[8] M. Shouman, T. Turner, and R. Stocker, “Using decision tree for diagnosing heart disease patients,” in *9th Australasian Data Mining Conference*, 2012.

[9] W. Cholid Wahyudin, F. Maisa Hana, and A. Prihandono, “PREDIKSI STUNTING PADA BALITA DI RUMAH SAKIT KOTA SEMARANG MENGGUNAKAN NAIVE BAYES,” 2023.

[10] W. Cholid Wahyudin, “KLASIFIKASI STUNTING BALITA MENGGUNAKAN NAIVE BAYES DENGAN SELEKSI FITUR FORWARD SELECTION.”

[11] H. Kaur and S. K. Wasan, “Empirical study on applications of data mining techniques in healthcare,” *Journal of Computer Science*, vol. 2, no. 2, pp. 194–200, 2006.

[12] J. Tang, S. Alelyani, and H. Liu, “Feature selection for classification: A review,” in *Data Classification: Algorithms and Applications*, 2014, pp. 37–64.

[13] M. J. Khan and M. Usman, “Early detection of heart diseases using classification and data mining techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, pp. 160–165, 2019.

[14] D. S. Rajput and S. B. Mane, “Hybrid approach for heart disease diagnosis using data mining techniques,” *Int J Comput Appl*, vol. 140, no. 15, pp. 17–21, 2016.

[15] R. Das, I. Turkoglu, and A. Sengur, “Effective diagnosis of heart disease through neural networks ensembles,” *Expert Syst Appl*, vol. 36, no. 4, pp. 7675–7680, 2009.

[16] A. Dey and S. Samanta, “Performance analysis of machine learning techniques for heart disease prediction,” *Procedia Comput Sci*, vol. 167, pp. 706–716, 2021.

[17] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[18] M. A. Al-Betar and others, “Data mining and machine learning techniques for heart disease prediction: A review,” *Artif Intell Rev*, vol. 51, pp. 597–623, 2019.

[19] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.

[20] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O’Sullivan, “A review of feature selection methods for machine learning-based disease risk prediction,” *Frontiers in Bioinformatics*, vol. 2, p. Article 927312, 2022, doi: 10.3389/fbinf.2022.927312.

[21] K. Sumwiza and et al., “Enhanced

cardiovascular disease prediction model using random forest algorithm," *Inform Med Unlocked*, vol. 41, p. Article 101316, 2023, doi: 10.1016/j.imu.2023.101316.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relation- ships that could be construed as a potential conflict of interest.

Article History:
Received: 04 February 2025 | Accepted: 05 March 2025| Published: 30 April 2025

`

Copyright © 2025 Widya Cholid Wahyudin, Tole Sutikno, Rusydi Umar, Ahmad Ridwan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms