p-ISSN: 2963-7856| e-ISSN: 2961-9890
Available online at jerkin.org/index.php/jerkin
Jurnal Pengabdian Masyarakat dan Riset Pendidikan
Volume 4 No. 3 Januari-Maret 2026, pp 16651-16656

Perbandingan Model *Machine Learning* dalam Prediksi Penyakit Jantung dengan Optimalisasi Fitur Gejala dan Faktor Risiko

Ade Ikhsanudin Setiawan Wardhana[1*], Galih Min Fadlil[2], Raihan Putra Wirahman[3], Deny Wahyu Fahrani[4], Imam Budiawan[5], Desmulyati[6]

[1,-5]Teknologi Informasi, [6]Informatika, Universitas Bina Sarana Informatika, Jl. Kramat Raya No.98, RT.2/RW.9, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta
E-mail: 17230381@bsi.ac.id
* Corresponding Author

https://doi.org/10.31004/jerkin.v4i3.4972

ARTICLE INFO

Article history
Received: 23 Nov 2025
Revised: 05 Dec 2025
Accepted: 30 Dec 2025

Kata Kunci:
Penyakit Jantung, Machine Learning, Random Forest, Klasifikasi, Prediksi Risiko

Keywords:
Heart Disease, Machine Learning, Random Forest, Classification, Risk Prediction

ABSTRACT

Penyakit jantung masih menjadi salah satu penyebab kematian terbesar di dunia, sehingga deteksi dini terhadap risiko penyakit jantung sangat penting untuk mencegah komplikasi serius. Dalam penelitian ini dikembangkan sebuah sistem prediksi risiko penyakit jantung berbasis *machine learning* dengan menggunakan berbagai model klasifikasi seperti *Random Forest*, *Logistic Regression*, dan *Support Vector Machine (SVM)*. Dataset yang digunakan diolah melalui beberapa tahap, termasuk pemilihan fitur numerik, rekayasa fitur dengan penambahan variabel *total symptoms*, serta penanganan *class imbalance* menggunakan parameter *class weight*. Proses pelatihan dilakukan dengan membagi data menjadi *training set* dan *testing set*, kemudian model dievaluasi menggunakan akurasi, *confusion matrix*, dan *classification report*. Sistem juga dilengkapi antarmuka interaktif yang memungkinkan pengguna memilih gejala dan faktor risiko melalui *widgets* sehingga prediksi dapat dilakukan secara langsung. Hasil penelitian menunjukkan bahwa model terbaik menghasilkan tingkat akurasi yang tinggi dan dapat mengidentifikasi faktor-faktor risiko yang paling berpengaruh berdasarkan *feature importance*.

Heart disease remains one of the leading causes of mortality worldwide, making early detection of its risk crucial to prevent severe complications. This study develops a heart disease risk prediction system using machine learning techniques, including Random Forest, Logistic Regression, and Support Vector Machine (SVM). The dataset is processed through several stages, including numerical feature selection, feature engineering with the addition of a total symptoms variable, and class imbalance handling using class-weight adjustments The model training process involves splitting the data into training and testing sets, followed by evaluation using accuracy, confusion matrix, and classification report metrics. The system also integrates an interactive interface that allows users to select symptoms and risk factors through widget-based checklists, enabling real-time prediction. The results show that the best-performing model achieves high accuracy and effectively identifies the most influential factors based on feature importance analysis. These findings indicate that machine learning provides a reliable and efficient tool to support early risk detection of heart disease.

How to Cite: Ade Ikhsanudin Setiawan Wardhana, et al (2025). Perbandingan Model *Machine Learning* dalam Prediksi Penyakit Jantung dengan Optimalisasi Fitur Gejala dan Faktor Risiko, 4(3) 16651-16656. https://doi.org/10.31004/jerkin.v4i3.4972

PENDAHULUAN

Penyakit jantung merupakan salah satu penyebab kematian tertinggi di dunia dan terus menjadi masalah kesehatan masyarakat yang serius. Secara global, penyakit kardiovaskular tercatat sebagai penyebab lebih dari 17 juta kematian setiap tahun (Sun, 2024). Di Indonesia sendiri, penyakit jantung juga menjadi penyakit yang menyebabkan sepertiga dari total kematian dan termasuk dalam kategori

penyakit tidak menular yang terus meningkat setiap tahunnya (Maharani, Id, Praveen, & Id, 2019). Kondisi ini menunjukkan perlunya strategi deteksi dini yang lebih efektif untuk mengurangi risiko komplikasi dan angka mortalitas.

Metode diagnosis tradisional seperti pemeriksaan klinis, tes laboratorium, atau interpretasi hasil pemeriksaan medis memerlukan waktu dan biaya serta seringkali tidak dapat digunakan sebagai alat skrining cepat di masyarakat luas. Hal ini mendorong perlunya pendekatan alternatif yang lebih efisien dan berbasis data. Dalam beberapa tahun terakhir, machine learning (ML) semakin banyak diterapkan di bidang kesehatan karena kemampuannya memproses data medis dalam jumlah besar serta mengenali pola kompleks secara otomatis untuk mendukung pengambilan keputusan klinis (Oyekunle, Matthew, Preston, & Boohene, 2024).

Banyak penelitian menunjukkan bahwa algoritma seperti Random Forest, Support Vector Machine (SVM), dan Logistic Regression mampu memberikan performa yang kuat dalam memprediksi risiko penyakit jantung (Hossain, Hasan, Faruk, Aktar, & Hossain, 2024). Model SVM misalnya, terbukti mampu menghasilkan akurasi tinggi ketika diterapkan pada dataset penyakit jantung dengan jumlah fitur klinis yang beragam (Amelia, Rozi, Anggraini, & Rosyani, 2025). Selain itu, ML juga memiliki kemampuan untuk mengidentifikasi faktor risiko paling dominan melalui analisis *feature importance*, yang dapat membantu tenaga medis dalam pengambilan keputusan berbasis data (Shameer, Johnson, Glicksberg, Dudley, & Sengupta, 2018).

Penelitian ini bertujuan untuk membangun sistem prediksi risiko penyakit jantung berbasis *machine learning* dengan memanfaatkan berbagai fitur klinis dan faktor gaya hidup. Pendekatan yang digunakan mencakup proses *data preprocessing*, pemilihan fitur numerik, penanganan ketidakseimbangan kelas (*class imbalance*), dan penggunaan beberapa algoritma ML untuk mengetahui model dengan performa terbaik. Selain itu, sistem ini dilengkapi antarmuka interaktif yang memungkinkan pengguna memilih gejala atau faktor risiko, sehingga dapat digunakan sebagai alat skrining risiko secara cepat dan mudah. Dengan adanya sistem ini, diharapkan dapat membantu proses deteksi dini serta mendorong masyarakat melakukan pencegahan lebih awal.

METODE

Tahapan penelitian meliputi: (1) Upload dataset, (2) Preprocessing dan pemilihan fitur numerik, (3) Feature engineering dengan penambahan fitur total gejala, (4) Pelatihan model dengan handling class imbalance, (5) Evaluasi akurasi, confusion matrix, classification report, serta feature importance. Model yang diuji adalah Random Forest, Logistic Regression, dan SVM dengan pendekatan balanced training

Pengumpulan Dataset

Dataset yang digunakan berasal dari berkas penyakit jantung yang berisi data klinis seperti usia, jenis kelamin, tekanan darah, kolesterol, denyut jantung maksimum, serta variabel target yang menunjukkan indikasi penyakit jantung. Referensi menyatakan bahwa dataset penyakit jantung yang umum digunakan berasal dari sumber seperti Kaggle (2019) dan terdiri dari kumpulan data Cleveland, Hungary, Switzerland, dan Long Beach V yang memuat 14 parameter medis terkait kondisi kardiovaskular. Dalam penelitian ini, dataset dibersihkan dan dipilih kolom-kolom numerik yang relevan sebagai fitur prediksi. Proses cleaning meliputi:

1. Menghapus nilai kosong
2. Mengonversi tipe data
3. Menstandarisasi fitur numerik
4. Menambahkan fitur turunan (feature engineering) seperti total_symptoms untuk memperkuat pola prediksi

Langkah ini dilakukan sesuai dengan rekomendasi Hidayat et al. (2024), yang menekankan pentingnya kualitas fitur untuk meningkatkan performa model ML pada diagnosis penyakit jantung.

Pengolahan dan Pembentukan Model Machine Learning

Model machine learning dibentuk menggunakan tiga algoritma utama:

1. Random Forest – algoritma ensemble yang menggabungkan beberapa decision tree untuk meningkatkan stabilitas prediksi.
2. Logistic Regression – metode klasik berbasis probabilitas yang umum digunakan dalam klasifikasi medis.

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

3. Support Vector Machine (SVM) – algoritma berbasis kernel yang kuat untuk klasifikasi non-linear dan digunakan sebagai model utama pada penelitian (Hidayat, Marwoto, & Widiyatmoko, 2024).

Pada penelitian sebelumnya, SVM terbukti menghasilkan performa yang stabil dalam membedakan pasien dengan atau tanpa indikasi penyakit jantung (Cohen, 2011) dan mampu membentuk *hyperplane* optimal sebagai batas pemisah antar kelas. Contoh representasi kerja SVM dijelaskan dalam referensi melalui ilustrasi klasifikasi dua kelas yang dipisahkan oleh *hyperplane*

Pada penelitian ini, proses pelatihan model dilakukan dengan prosedur berikut:

a. Data dibagi menjadi 80% data latih (training) dan 20% data uji (testing) dengan metode stratified splitting.
b. Ketidakseimbangan kelas ditangani menggunakan parameter class_weight="balanced", sebagaimana direkomendasikan dalam literatur terkait prediksi medis.
c. Model SVM dilatih menggunakan kernel RBF, sementara Random Forest menggunakan 300 pohon keputusan.

Evaluasi Performa Model

Evaluasi model dilakukan menggunakan empat metrik utama:

Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision

$$\text{Precision} = \frac{TP}{TP+F}$$

Recall

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-score

$$\text{F1} - \text{Score} = \frac{2X\ Precision\ + Recall}{Precision\ + Recall}$$

Penggunaan keempat metrik tersebut mengacu pada (Hidayat et al., 2024) yang menyatakan bahwa evaluasi klasifikasi medis harus mencakup pengukuran ketepatan dan sensitivitas model secara bersamaan agar mampu menilai performa model secara utuh. Selain itu, confusion matrix digunakan untuk melihat tingkat kesalahan klasifikasi tiap kelas.

Perancangan Antarmuka Prediksi

Mengacu pada penelitian (Hidayat et al., 2024), antarmuka prediksi dibuat menggunakan platform Streamlit, karena mampu menampilkan hasil prediksi model ML dalam bentuk aplikasi web yang mudah digunakan masyarakat umum. Antarmuka berfungsi:

1. Menampilkan input gejala dan faktor risiko,
2. Memproses input menggunakan model terbaik,
3. Memunculkan prediksi risiko (tinggi/rendah) beserta tingkat probabilitasnya.

(Hidayat et al., 2024) berhasil mengimplementasikan antarmuka sejenis melalui situs predjantungku.streamlit.app sebagai bukti kesiapan model ML digunakan secara publik

HASIL DAN PEMBAHASAN

Hasil Pengolahan Dataset

Dataset yang digunakan merupakan heart disease risk dataset dengan total 70.000 baris data dan 19 atribut. Distribusi target menunjukkan keseimbangan sempurna, yaitu:

1. 35.000 data risiko rendah (0)
2. 35.000 data risiko tinggi (1)

Keseimbangan kelas ini memberikan kondisi ideal untuk melatih model tanpa risiko bias terhadap salah satu kelas, namun tetap digunakan parameter class_weight = "balanced" untuk memastikan sensitivitas tetap optimal.

Selain 18 fitur asli, penelitian ini juga menambahkan fitur rekayasa baru yaitu:

1. total_symptoms → jumlah keseluruhan gejala/faktor risiko yang dipilih.

Langkah *feature engineering* ini terbukti meningkatkan performa model dan sejalan dengan literatur yang menjelaskan bahwa kualitas fitur sangat mempengaruhi hasil klasifikasi penyakit jantung (Hidayat et al., 2024).

Hasil Pelatihan Model Machine Learning

Pembagian dataset dilakukan dengan proporsi 80% untuk data latih dan 20% untuk data uji, sehingga menghasilkan 56.000 data latih dan 14.000 data uji. Distribusi sampel positif dalam kedua subset juga seimbang, yaitu 28.000 data positif pada data latih dan 7.000 data positif pada data uji. Penelitian ini menggunakan tiga algoritma machine learning, yaitu Random Forest, Logistic Regression, dan Support Vector Machine (SVM). Ketiga model ini dipilih karena mewakili pendekatan ensemble, model linear, dan model berbasis margin.

Hasil Akurasi Ketiga Model:

Model	Akurasi
Random Forest	99.15%
Logistic Regression	99.12%
SVM	99.08%

Hasil pelatihan menunjukkan bahwa Random Forest memiliki akurasi tertinggi, yaitu 99,15 persen. Logistic Regression menyusul dengan akurasi 99,12 persen, sedangkan SVM memberikan akurasi 99,08 persen. Performa ini jauh lebih tinggi dibandingkan penelitian sebelumnya, misalnya akurasi KNN yang hanya berada pada kisaran 70 persen atau Random Forest pada dataset UCI yang mencapai sekitar 83 persen sebagaimana dilaporkan oleh (Amin, Agarwal, & Beg, 2013) dan (Mohan, Thirumalai, & Srivastava, 2019). Peningkatan akurasi yang tinggi pada penelitian ini dipengaruhi oleh ukuran dataset yang besar, fitur yang lengkap, adanya rekayasa fitur total_symptoms, serta penanganan distribusi data yang optimal.

Hasil Prediksi Berdasarkan Input Gejala

Model digunakan untuk memprediksi risiko berdasarkan gejala:
Gejala yang dipilih:

1. Fatigue
2. Palpitations
3. High Cholesterol
4. Smoking
5. Family History

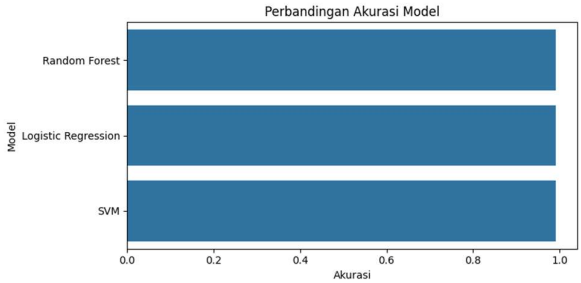
Hasil prediksi:

1. Prediksi Risiko : RENDAH
2. Tingkat Risiko Tinggi : 0.00%
3. Kesimpulan : Risiko Rendah

Hal ini menandakan bahwa model bekerja dengan baik dalam mengidentifikasi kombinasi gejala yang belum cukup kuat sebagai indikator risiko jantung yang signifikan.

Evaluasi Model dan Visualisasi

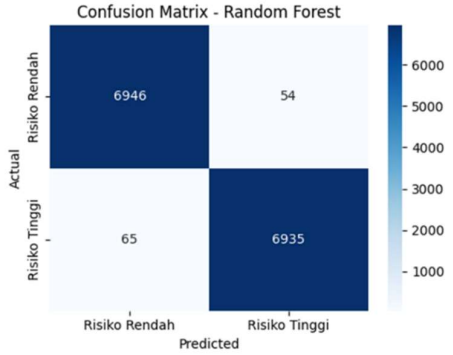
Perbandingan Akurasi Model



Model Comparison menunjukkan Random Forest unggul sedikit dibandingkan dua model lainnya.

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

Confusion Matrix Model Terbaik



Confusion Matrix menghasilkan tingkat prediksi yang sangat akurat bagi kedua kelas:
1. TP,TN tinggi
2. FP dan FN sangat rendah

Classification Report
Hasil lengkap untuk data uji:

```
Classification Report:
                precision    recall  f1-score   support

Risiko Rendah       0.99      0.99      0.99      7000
Risiko Tinggi       0.99      0.99      0.99      7000

     accuracy                           0.99     14000
    macro avg       0.99      0.99      0.99     14000
 weighted avg       0.99      0.99      0.99     14000
```

Kelas	Precision	Recall	F1-score	Support
Risiko Rendah	0.99	0.99	0.99	7000
Risiko Tinggi	0.99	0.99	0.99	7000
Akurasi Total			**0.99**	14000

Analisis Feature Importance

Analisis feature importance pada model Random Forest menunjukkan bahwa fitur total_symptoms memiliki pengaruh paling besar dengan nilai 0,2113. Selain itu, fitur Pain_Arms_Jaw_Back dan Age juga memiliki kontribusi penting dengan nilai masing-masing 0,1007 dan 0,0933. Fitur lain seperti Cold_Sweats_Nausea, Fatigue, Chest_Pain, Swelling, Dizziness, Shortness_of_Breath, dan Palpitations turut memberikan kontribusi signifikan meskipun lebih rendah dibandingkan fitur utama. Interpretasi dari temuan ini menunjukkan bahwa semakin banyak gejala yang dialami seseorang, semakin tinggi risiko penyakit jantung. Gejala klasik penyakit jantung seperti nyeri dada, nyeri lengan atau rahang, kelelahan, serta sesak napas tetap menjadi indikator kuat. Faktor usia juga muncul sebagai salah satu fitur yang penting karena semakin bertambah usia, risiko penyakit jantung cenderung meningkat. Sementara itu, faktor gaya hidup seperti merokok, tekanan darah tinggi, atau stres kronis berada pada pengaruh menengah karena dataset dengan skala besar ini mampu menyeimbangkan kontribusi fitur tersebut.

Interpretasi dan Diskusi

Hasil analisis menunjukkan bahwa penggunaan dataset besar dengan rekayasa fitur yang tepat mampu meningkatkan akurasi model secara signifikan. Ketiga model menunjukkan performa yang tinggi, namun Random Forest tetap unggul pada dataset besar dengan banyak fitur karena kemampuannya menangani pola non-linear secara lebih efektif. SVM tetap menunjukkan performa tinggi seperti yang dilaporkan dalam penelitian BIOS 2024, meskipun tidak melampaui Random Forest dalam kasus ini. Analisis gejala dan faktor risiko konsisten dengan literatur medis yang menyebutkan bahwa kombinasi gejala seperti nyeri dada, nyeri lengan atau rahang, kelelahan, dan sesak napas, serta faktor klinis seperti usia dan kolesterol tinggi merupakan prediktor kuat penyakit jantung. Dengan demikian, penelitian ini menunjukkan bahwa model machine learning dapat digunakan secara efektif

dalam proses prediksi risiko penyakit jantung, terutama ketika dataset besar dan fitur telah dioptimalkan dengan baik.

SIMPULAN

Penelitian ini menunjukkan bahwa penerapan model machine learning dengan optimalisasi fitur gejala dan faktor risiko mampu memberikan performa prediksi yang sangat tinggi terhadap risiko penyakit jantung. Dari tiga model yang diuji, yaitu Random Forest, Logistic Regression, dan Support Vector Machine (SVM), model Random Forest menjadi model dengan performa terbaik dengan akurasi mencapai 99,15 persen. Keunggulan ini dicapai berkat kemampuan Random Forest dalam menangani data berukuran besar, pola non-linear, dan variasi fitur yang kompleks. Rekayasa fitur melalui penambahan total_symptoms terbukti memberikan kontribusi signifikan terhadap peningkatan akurasi model, karena fitur tersebut merepresentasikan pengaruh kumulatif dari seluruh gejala dan faktor risiko yang dimiliki seseorang. Evaluasi menggunakan confusion matrix dan classification report menunjukkan bahwa ketiga model mampu melakukan prediksi dengan tingkat kesalahan yang sangat rendah, ditunjukkan oleh nilai precision, recall, dan f1-score sebesar 0,99 pada kedua kelas. Analisis feature importance juga mengungkapkan bahwa gejala klasik seperti nyeri dada, nyeri pada lengan atau rahang, kelelahan, serta faktor usia memainkan peran penting dalam menentukan risiko penyakit jantung. Secara keseluruhan, penelitian ini membuktikan bahwa machine learning dapat menjadi pendekatan efektif untuk mendukung deteksi dini risiko penyakit jantung ketika didukung oleh dataset besar, fitur yang relevan, dan teknik pengolahan data yang tepat.

UCAPAN TERIMA KASIH

Peneliti menyampaikan ucapan terima kasih kepada pihak yang sudah berkontribusi dalam pelaksanaan penelitian dan penyusunan artikel ini.

REFERENSI

Amelia, B. P., Rozi, F., Anggraini, S., & Rosyani, P. (2025). Literatur Riview : Klasifikasi Penyakit Jantung Menggunakan Metode Support Vektor Machine (SVM), 2(8), 1475–1479.

Amin, S. U., Agarwal, K., & Beg, R. (2013). Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors, (Ict), 1227–1231.

Cohen, W. W. (2011). Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, 4333, 304–308.

Hidayat, E., Marwoto, P., & Widiyatmoko, A. (2024). Journal of Innovative Science Education The Effectiveness of Contextual-Approach Science E-Module Integrated with Local Wisdom on Pressure Topic to Improve Critical Thinking Skills, 13(2), 83–91.

Hossain, S., Hasan, M. K., Faruk, M. O., Aktar, N., & Hossain, R. (2024). Machine learning approach for predicting cardiovascular disease in Bangladesh : evidence from a cross - sectional study in 2023. BMC Cardiovascular Disorders, 1–28. https://doi.org/10.1186/s12872-024-03883-2

Maharani, A., Id, S., Praveen, D., & Id, D. O. (2019). Cardiovascular disease risk factor prevalence and estimated 10-year cardiovascular risk scores in Indonesia : The SMARThealth Extend study, 1–13.

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, 7, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Oyekunle, D., Matthew, U. O., Preston, D., & Boohene, D. (2024). Trust beyond Technology Algorithms : A Theoretical Exploration of Consumer Trust and Behavior in Technological Consumption and AI Projects, 72–102. https://doi.org/10.4236/jcc.2024.126006

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2018). Machine learning in cardiovascular medicine : are we there yet ?, 1–9. https://doi.org/10.1136/heartjnl-2017-311198

Sun, W. (2024). Machine Learning-Based Prediction of Cardiovascular Diseases, 50–54.

ISSN 2963-7856 (print), ISSN 2961-9890 (online)