Vol. 8, No. 1, April 2026, pp 78-89
https://doi.org/10.36590/jika.v8i1.1929
http://salnesia.id/index.php/jika
jika@salnesia.id, p-ISSN: 2337-9847, e-ISSN: 2686-2883
Penerbit: Sarana Ilmu Indonesia (salnesia)

ARTIKEL PENELITIAN

Klasifikasi Risiko Penyakit Jantung dan *Feature Importance* Berbasis *Machine Learning* Menggunakan Data Mikro SKI 2023

Heart Disease Risk Classification and Feature Importance Based on Machine Learning Using Microdata SKI 2023

Zahra Mulki Syari'ati[1*], Maula Ismail Muhamad[1], Lina Khasanah[1], Bambang Karmanto[1], Suratmi[2]

[1]Program Studi Rekam Medis dan Informasi Kesehatan, Poltekkes Kemenkes Tasikmalaya, Tasikmalaya, Indonesia

[2]Program Studi Terapan Kebidanan, Poltekkes Kemenkes Tasikmalaya, Tasikmalaya, Indonesia

Abstract

Cardiovascular disease remained the leading cause of mortality in Indonesia, with death rates increasing by more than 25% and national health expenditure reaching Rp17.92 trillion. Its complex risk profile requires population-based predictive approaches. The 2023 Indonesian Health Survey (SKI) provided extensive data suitable for Machine Learning-based risk modelling. This study aimed to develop a classification model for heart-disease risk and identify dominant risk factors using the Random Forest algorithm. A case-control design was applied, with data divided into training and testing sets using an 80:20 ratio to ensure objective model evaluation. The analysis followed the Knowledge Discovery in Database (KDD) framework, including data selection, preprocessing, transformation, modelling, and evaluation. Random Forest was used for classification, while feature importance was assessed using Information Gain and Gain Ratio. Model performance was evaluated using accuracy, sensitivity, and specificity. Age, hypertension, and Body Mass Index (BMI) were identified as the most influential predictors. The model achieved an accuracy of 72.13%, sensitivity of 73.11%, and specificity of 71.26%, indicating stable classification performance on large population data. However, this study is limited by the use of secondary data and the absence of external validation. These findings highlight the potential of Machine Learning to support population-based risk stratification and inform targeted prevention strategies, contributing to evidence-based policy development and early screening programs in primary healthcare settings.

Keywords: *heart disease, health survey, machine learning*

Article history:

Submitted 15 Januari 2026
Accepted 30 April 2026
Published 30 April 2026

PUBLISHED BY:
Sarana Ilmu Indonesia (salnesia)
Address:
Jl. Dr. Ratulangi No. 75A, Baju Bodoa, Maros Baru, Kab. Maros, Provinsi Sulawesi Selatan, Indonesia

Email:
info@salnesia.id, jika@salnesia.id
Phone:
+62 85255155883

Abstrak

Penyakit jantung merupakan penyebab utama kematian di Indonesia dengan peningkatan mortalitas lebih dari 25% dan beban pembiayaan kesehatan yang tinggi, tercermin dari klaim JKN mencapai Rp17,92 triliun. Kompleksitas faktor risiko menuntut pendekatan prediktif berbasis populasi. Survei Kesehatan Indonesia (SKI) 2023 menyediakan data yang luas untuk pengembangan model risiko berbasis *Machine Learning.* Penelitian ini bertujuan mengembangkan model klasifikasi risiko penyakit jantung serta mengidentifikasi faktor risiko dominan menggunakan algoritma *Random Forest*. Desain *case-control* diterapkan dalam penelitian ini dengan pembagian data training dan testing menggunakan rasio 80:20 untuk memastikan evaluasi performa model yang objektif. Analisis mengikuti tahapan *Knowledge Discovery in Database* (KDD) yang meliputi seleksi data, prapemrosesan, transformasi, pemodelan, dan evaluasi. *Random Forest* digunakan untuk klasifikasi, sedangkan *feature importance* dianalisis menggunakan *Information Gain* dan *Gain Ratio.* Evaluasi model dilakukan menggunakan akurasi, sensitivitas, dan spesifisitas. Hasil menunjukkan bahwa usia, hipertensi, dan Indeks Massa Tubuh (IMT) merupakan prediktor paling dominan. Model menghasilkan akurasi 72,13%, sensitivitas 73,11%, dan spesifisitas 71,26% yang menunjukkan performa klasifikasi yang stabil pada data populasi besar. Namun, penelitian ini memiliki keterbatasan berupa penggunaan data sekunder dan belum dilakukannya validasi eksternal. Temuan ini menunjukkan potensi *Machine Learning* dalam mendukung stratifikasi risiko berbasis populasi serta pengembangan strategi pencegahan terarah, sehingga dapat berkontribusi pada perumusan kebijakan berbasis bukti dan implementasi program skrining dini di layanan kesehatan primer.

Kata Kunci: *machine learning*, survey kesehatan, penyakit jantung

*Penulis Korespondensi:
Zahra Mulki Syari'ati, email: zahramusy21@gmail.com

Highlight:

- Berdasarkan analisis kontribusi fitur (*feature importance*), faktor usia, hipertensi, dan Indeks Massa Tubuh (IMT) diidentifikasi sebagai prediktor atau faktor risiko yang paling berpengaruh terhadap tingkat risiko penyakit jantung di Indonesia.
- Model berbasis algoritma *Random Forest* yang dikembangkan menunjukkan performa klasifikasi yang stabil pada populasi data besar, dengan capaian tingkat **a**kurasi sebesar 72,13%, sensitivitas 73,11%, dan spesifisitas 71,26%.
- Pemanfaatan *Machine Learning* ini memiliki potensi besar untuk membantu stratifikasi risiko berbasis populasi. Temuan ini dapat mendukung perumusan kebijakan kesehatan berbasis bukti serta implementasi program skrining (deteksi) dini penyakit jantung pada layanan kesehatan primer.

PENDAHULUAN

Penyakit jantung masih menjadi penyebab utama kematian di seluruh dunia dan menyumbang proporsi kematian yang terus meningkat dalam dua dekade terakhir. Data *Global Burden of Disease* menunjukkan bahwa penyakit jantung iskemik merupakan kontributor kematian terbesar secara global, sejalan dengan perubahan gaya hidup, transisi demografis, dan meningkatnya prevalensi obesitas serta hipertensi (Islam et al.,

2023). Tren ini menggambarkan pergeseran epidemiologi global menuju dominasi Penyakit Tidak Menular (PTM), yang kini menyebabkan lebih dari 74% kematian dunia (Vinet dan Zhedanov, 2011). Di Indonesia, beban penyakit jantung menunjukkan pola yang serupa dan bahkan lebih mengkhawatirkan. Laporan *Global Burden of Disease* tahun 2020 mencatat bahwa angka kematian akibat penyakit jantung meningkat lebih dari 25%, menjadikannya penyebab kematian tertinggi di Indonesia (BKPK, 2021).

Selain itu, data *Institute for Health Metrics and Evaluation* (IHME) mencatat bahwa penyakit jantung menyumbang 53,2% kematian nasional, memperlihatkan dominasi yang sangat signifikan dalam struktur mortalitas (Rusyda, 2025). Beban ekonomi yang ditimbulkan juga sangat besar; pada tahun 2022, total pembiayaan penyakit jantung melalui Jaminan Kesehatan Nasional (JKN) mencapai Rp17,92 triliun, menjadikannya sebagai penyakit katastropik dengan biaya tertinggi (Solida et al., 2021). Kondisi ini tidak hanya mencerminkan tingginya prevalensi, tetapi juga menunjukkan bahwa banyak kasus terdiagnosis pada tahap lanjut sehingga memerlukan intervensi medis yang kompleks dan mahal (Tampubolon et al., 2023).

Urgensi peningkatan deteksi dini diperkuat dengan tersedianya sumber data berskala nasional yang lebih mutakhir dan komprehensif. Peluncuran Survei Kesehatan Indonesia (SKI) 2023 sebagai integrasi Riskesdas dan SSGI berhasil mengumpulkan sebanyak lebih dari 586.000 rumah tangga di seluruh Indonesia, menghasilkan populasi sebanyak 602.982 subjek dewasa yang dapat dianalisis. Survei Kesehatan Indonesia (SKI) tahun 2023 memiliki cakupan variabel yang jauh lebih luas dibandingkan survei sebelumnya, mencakup variabel klinis, perilaku kesehatan, sosial ekonomi, dan diagnosis tenaga kesehatan hingga tingkat kabupaten/kota (SKI, 2023). Namun, meskipun dataset ini sangat komprehensif, pemanfaatannya untuk analisis prediktif penyakit jantung dengan pendekatan *Machine Learning* masih terbatas (Nazari et al., 2024).

Sejalan dengan perkembangan ilmu data, *Machine Learning* (ML) telah banyak digunakan untuk meningkatkan akurasi prediksi berbagai penyakit kronis (Khalisatifa et al., 2024; Naser et al., 2024). *Random Forest* merupakan salah satu algoritma populer, mampu menangkap pola nonlinear dan interaksi kompleks antarvariabel yang sulit ditangkap oleh metode statistik tradisional (Saputra, 2023). Penelitian internasional menunjukkan bahwa ML dapat meningkatkan akurasi prediksi penyakit jantung secara signifikan melalui analisis *big data* kesehatan (Tasnim et al., 2025). Meski demikian, sebagian besar penelitian sebelumnya menggunakan dataset klinis, berskala kecil, atau tidak representatif secara populasi, seperti UCI *Heart Dataset* atau data institusional rumah sakit (Alwakid et al., 2025; Reddy et al., 2024). Belum banyak penelitian di Indonesia yang menggabungkan data mikro SKI 2023, algoritma *Random Forest,* serta analisis *feature importance* untuk mengidentifikasi faktor risiko penyakit jantung secara populasi nasional yang menjadi aspek kebaruan penelitian ini.

Kebaruan penelitian ini terletak pada pemanfaatan dataset besar SKI 2023 yang representatif secara nasional, dengan mengintegrasikan pendekatan *Machine Learning* untuk membangun model klasifikasi risiko penyakit jantung yang akurat. Selain itu, penelitian ini mengidentifikasi faktor risiko dominan (usia, hipertensi, IMT, dan faktor perilaku) berdasarkan nilai *Information Gain* (IG) dan *Gain Ratio* (GR) yang diuji pada sampel besar berjumlah 5.580 subjek hasil *matching case-control.* Penelitian sebelumnya di Indonesia belum mengombinasikan dataset dengan metode ML secara mendalam, sehingga hasilnya dapat memberikan kontribusi signifikan dalam pengembangan skrining risiko berbasis data pada layanan kesehatan primer (Khalisatifa et al., 2024; Wardhana et al., 2023). Berdasarkan urgensi dan kesenjangan penelitian

tersebut, penelitian ini bertujuan mengembangkan model klasifikasi risiko penyakit jantung menggunakan algoritma *Random Forest* berbasis data mikro SKI 2023 serta mengidentifikasi faktor risiko paling berpengaruh melalui analisis *feature importance.*

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan desain *case-control* berbasis data sekunder dari Survei Kesehatan Indonesia (SKI) 2023. Penelitian dilaksanakan pada Oktober–Desember 2025 dengan cakupan analisis tingkat nasional sesuai struktur data SKI. Populasi penelitian mencakup seluruh subjek dewasa yang tercatat dalam modul penyakit tidak menular. Sampel penelitian diperoleh melalui proses matching antara subjek yang memiliki penyakit jantung dan subjek tanpa penyakit jantung sehingga menghasilkan total 5.580 sampel yang memenuhi kriteria inklusi. Teknik *sampling* yang digunakan adalah *purposive sampling* berdasarkan ketersediaan variabel yang lengkap dalam *dataset.* Variabel yang dianalisis meliputi usia, status hipertensi, indeks massa tubuh, kebiasaan merokok, konsumsi alkohol, tingkat aktivitas fisik, konsumsi buah dan sayur, serta variabel sosiodemografi yang relevan dengan risiko penyakit jantung. Pengumpulan data dilakukan secara tidak langsung melalui pemanfaatan data sekunder yang telah melewati proses verifikasi, validasi, dan pengendalian kualitas oleh Badan Kebijakan Pembangunan Kesehatan (BKPK, 2021).

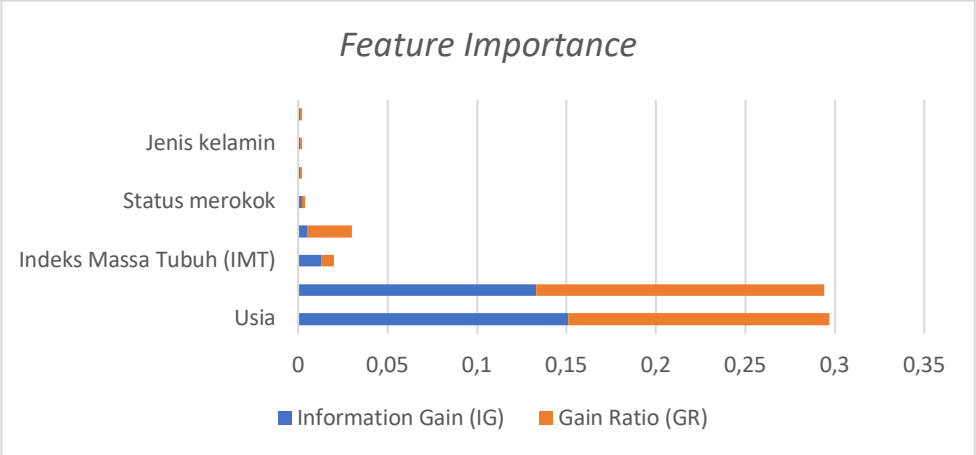
Proses *matching* pada desain *case-control* dilakukan dengan pendekatan proporsi seimbang antara kelompok kasus (subjek dengan diagnosis penyakit jantung) dan kontrol (tanpa penyakit jantung) berdasarkan karakteristik usia dan jenis kelamin untuk meminimalkan bias seleksi. Penanganan missing value dilakukan melalui metode listwise deletion pada variabel dengan proporsi data hilang rendah (<5%) serta imputasi sederhana pada variabel tertentu yang masih relevan secara analitis. Dataset dibagi menjadi data *training* dan *testing* dengan rasio 80:20 untuk memastikan evaluasi performa model yang objektif. Selain itu, validasi model dilakukan menggunakan teknik *10-fold cross-validation* untuk meningkatkan stabilitas dan generalisasi model. Parameter utama *Random Forest* yang digunakan meliputi jumlah pohon *(n_estimators = 100),* kedalaman maksimum pohon *(max_depth),* serta pemilihan subset fitur secara acak pada setiap split untuk mengoptimalkan performa model.

Proses analisis mengikuti tahapan *Knowledge Discovery in Database (KDD)* yang terdiri atas seleksi data, pembersihan data, transformasi data, pemodelan, dan evaluasi. Tahap seleksi data dilakukan dengan memilih variabel yang secara teoritis berhubungan dengan kejadian penyakit jantung serta memastikan kelengkapan data untuk kebutuhan pemodelan. Tahap pembersihan data meliputi penanganan nilai hilang, penyesuaian tipe data, dan pengkodean ulang agar seluruh variabel dapat diproses secara optimal oleh algoritma *Machine Learning.* Transformasi data dilakukan melalui normalisasi dan konversi variabel kategorik menjadi format numerik sesuai kebutuhan model. Proses pemodelan dilakukan dengan algoritma *Random Forest* untuk membangun model klasifikasi risiko penyakit jantung. Tahap evaluasi menggunakan *confusion matrix,* akurasi, sensitivitas, dan spesifisitas untuk menilai performa model, sedangkan analisis *feature importance* dilakukan menggunakan *information gain* dan *gain ratio* untuk menentukan variabel yang memberikan kontribusi terbesar terhadap klasifikasi. Seluruh hasil analisis disajikan secara deskriptif melalui tabel dan narasi untuk menggambarkan performa model serta pola kontribusi variabel terhadap risiko penyakit jantung.

HASIL DAN PEMBAHASAN

Analisis faktor risiko dominan model

Struktur *feature importance* pada model *Random Forest* pada bagian ini diawali dengan pemetaan posisi setiap variabel berdasarkan kekuatan kontribusinya terhadap proses klasifikasi. Visualisasi tersebut menghadirkan urutan atribut berdasarkan nilai *information gain* dan *gain ratio,* sehingga memberikan gambaran awal mengenai bagaimana algoritma menyusun logika pemisahan kelas berdasarkan sinyal informasi yang paling kuat. Penyajian ini berfungsi sebagai fondasi interpretatif yang menegaskan kerangka kerja internal model sebelum analisis diarahkan pada pembahasan variabel-variabel dengan kontribusi terbesar terhadap risiko penyakit jantung.

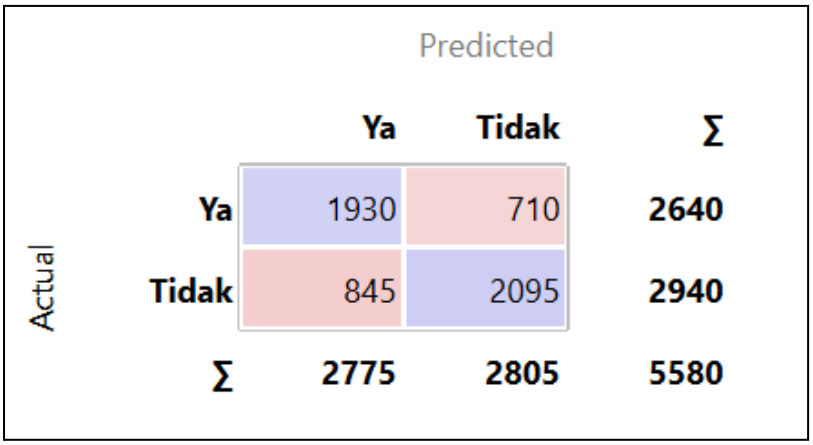


Gambar 1. Analisis *feature importance*

Gambar 1. analisis *feature importance* menunjukkan bahwa variabel usia muncul sebagai determinan paling dominan dengan kontribusi moderat (IG 0,151; GR 0,146), diikuti oleh variabel hipertensi yang menempati posisi kedua dengan kontribusi kuat dan efisiensi pemisahan tertinggi dalam model (IG 0,133; GR 0,161). Sementara itu, indeks massa tubuh (IMT) berada pada peringkat ketiga dengan kontribusi rendah namun tetap relevan dalam struktur klasifikasi (IG 0,013; GR 0,007), menandakan bahwa ketiga variabel ini merupakan faktor yang paling berpengaruh dalam pemodelan risiko penyakit jantung.

Evaluasi kinerja model *Machine Learning* (ML)

Evaluasi performa algoritma *Random Forest* dilakukan menggunakan *confusion matrix* sebagai instrumen dasar untuk menilai ketepatan klasifikasi terhadap status penyakit jantung pada 5.580 subjek. *Confusion matrix* memberikan struktur evaluatif yang menggambarkan distribusi prediksi benar dan tidak tepat, serta menjadi fondasi penghitungan akurasi, sensitivitas, dan spesifisitas. Nilai *True Positive* (TP), *False Negative* (FN), *False Positive* (FP), dan *True Negative* (TN) merepresentasikan kemampuan model dalam mengenali pola risiko maupun non-risiko yang relevan.



Gambar 2. *Confusion matrix*

Gambar 2 di atas memperlihatkan distribusi hasil klasifikasi model *Random Forest* dalam membedakan subjek yang memiliki penyakit jantung dan yang tidak. Matriks ini memberikan gambaran awal mengenai bagaimana model memproses pola data dan mengidentifikasi dua kategori tersebut berdasarkan karakteristik masing-masing individu. Berdasarkan Gambar 2, terdapat 1.930 subjek yang benar-benar memiliki penyakit jantung dan berhasil diklasifikasikan dengan tepat oleh model sebagai positif (*True Positive,* TP). Nilai ini menunjukkan kemampuan model dalam mengenali sebagian besar individu dengan kondisi penyakit jantung secara akurat sesuai label aktualnya.

Sebaliknya, terdapat 2.095 subjek yang tidak memiliki penyakit jantung dan berhasil diklasifikasikan dengan benar sebagai negatif (*True Negative,* TN). Temuan ini menandakan bahwa model cukup efektif dalam mengenali kelompok non-risiko dan mempertahankan ketepatan klasifikasi pada kategori sehat. Pada sisi lain, terdapat 845 subjek yang sebenarnya tidak memiliki penyakit jantung, tetapi diprediksi sebagai memiliki penyakit (*False Positive,* FP). Kondisi ini menggambarkan adanya sebagian individu sehat yang masuk kategori risiko oleh model, sehingga perlu perhatian pada pola fitur yang mungkin menyerupai karakteristik kelompok berisiko.

Terakhir, sebanyak 710 subjek yang sebenarnya memiliki penyakit jantung justru diprediksi sebagai tidak memiliki penyakit (*False Negative,* FN). Angka ini menunjukkan sebagian kasus positif yang luput terdeteksi oleh model, yang secara klinis penting karena terkait dengan potensi keterlambatan identifikasi individu yang seharusnya masuk kelompok risiko. Hal ini menjadi dasar untuk menghitung dan menafsirkan akurasi, sensitivitas, serta spesifisitas secara lebih sistematis pada bagian berikutnya. Perhitungan metrik dilakukan berdasarkan standar evaluasi model klasifikasi, yaitu akurasi, sensitivitas, dan spesifisitas. Ketiga parameter tersebut dihitung menggunakan formula baku sebagai berikut.

Tabel 1. Hasil perhitungan akurasi, sensitivitas, dan spesifisitas

Metrik	Nilai (%)
Akurasi	72,13
Sensitivitas	73,11
Spesifisitas	71,26

Sumber: Data sekunder: 2023

Hasil evaluasi kinerja algoritma *Random Forest* (Tabel 1) menunjukkan bahwa model memiliki kapasitas klasifikasi yang kuat, tercermin dari akurasi 72,13% yang menandakan kemampuan model dalam membaca pola risiko penyakit jantung pada populasi besar dengan karakteristik yang kompleks seperti SKI 2023. Pada taraf akurasi tersebut, sebagian besar subjek, baik yang berpenyakit maupun tidak berpenyakit, berhasil diklasifikasikan secara tepat. Nilai sensitivitas 73,11% dan spesifisitas 71,26% menggambarkan keseimbangan performa model dalam mengenali individu berisiko maupun tidak berisiko (Pal dan Patel, 2020). Sensitivitas yang tinggi menunjukkan ketepatan model dalam mendeteksi subjek dengan penyakit jantung berdasarkan fitur-fitur utama seperti tekanan darah tinggi, usia lanjut, dan IMT yang berada pada kategori *overweight* atau obesitas.

Sementara itu, spesifisitas yang cukup kuat menandakan kemampuan model mengidentifikasi kelompok sehat secara konsisten tanpa menghasilkan overestimasi prediksi *(over-detection).* Keseimbangan kedua metrik ini menegaskan bahwa mekanisme *ensemble learning,* melalui kombinasi banyak *decision tree* dan pemilihan fitur secara acak memberikan struktur prediksi yang kokoh dan tahan terhadap multikolinearitas (Ramadhan et al., 2023). Secara keseluruhan, performa ini menunjukkan bahwa *Random Forest* merupakan algoritma yang andal untuk pemodelan risiko penyakit jantung pada skala populasi nasional, dan relevan untuk mendukung pengembangan sistem skrining kesehatan berbasis data besar di Indonesia.

Ketiga metrik evaluasi yang stabil memperlihatkan keselarasan antara struktur pembelajaran model dan karakteristik data, sekaligus menegaskan kapasitas *Random Forest* dalam menghasilkan klasifikasi yang kuat, adaptif, dan dapat diandalkan (Chandra dan Prasetyo, 2024). Dengan kemampuan menangani data besar dan multivariabel secara alami, model ini relevan untuk diterapkan dalam epidemiologi digital, pengembangan sistem skrining berbasis *big* data*,* serta mendukung pengambilan keputusan berbasis bukti *(evidence-based decision making)* dalam sistem kesehatan modern (Ahmed dan Husien, 2024). Hasil evaluasi ini mengkonfirmasi bahwa *Random Forest* dapat menjadi fondasi penting dalam analisis risiko penyakit jantung di Indonesia, terutama ketika digunakan dalam skala nasional yang membutuhkan pendekatan metodologis yang solid dan responsif terhadap keragaman populasi.

Hasil penelitian ini menunjukkan bahwa usia, hipertensi, dan indeks massa tubuh (IMT) merupakan determinan utama dalam klasifikasi risiko penyakit jantung pada populasi dewasa Indonesia, dan temuan ini selaras dengan landasan ilmiah mengenai patogenesis penyakit kardiovaskular (Nazari et al., 2024). Usia muncul sebagai faktor paling dominan karena proses penuaan secara fisiologis meningkatkan kekakuan arteri, menurunkan fungsi endotel, dan memicu akumulasi plak aterosklerotik, sehingga risiko penyakit jantung meningkat secara progresif seiring bertambahnya usia, sebagaimana dijelaskan dalam berbagai literatur epidemiologi *modern* (Ramadhanti et al., 2024). Hipertensi menempati posisi kedua dengan kontribusi yang sangat kuat, yang dapat dijelaskan melalui mekanisme kerusakan dinding pembuluh darah akibat tekanan darah tinggi kronis, memicu remodeling vaskular dan hipertrofi ventrikel kiri, yang akhirnya meningkatkan kerentanan terhadap kejadian kardiovaskular berat (Hidayat et al., 2023). Temuan ini konsisten dengan studi patologi vaskular terbaru yang menekankan peran tekanan darah sebagai penyebab utama gangguan sistemik yang berhubungan dengan penyakit jantung (PERKI, 2022).

Sementara itu, IMT yang berada pada peringkat ketiga menunjukkan bahwa obesitas dan overweight tetap menjadi komponen penting risiko penyakit jantung, terutama melalui jalur inflamasi kronis dan resistensi insulin, yang telah lama diakui

sebagai faktor kunci dalam perkembangan aterosklerosis (Delavera et al., 2021). Dominasi ketiga variabel ini dalam struktur *feature importance* menunjukkan bahwa faktor biologis memiliki pola kontribusi yang lebih kuat dibandingkan faktor perilaku seperti merokok, konsumsi alkohol, atau aktivitas fisik, yang dalam penelitian ini muncul dengan nilai IG dan GR yang lebih rendah. Kinerja model yang tercermin dalam *confusion matrix* lebih lanjut menguatkan interpretasi bahwa struktur data kesehatan populasi Indonesia memiliki pola risiko yang terutama dipengaruhi oleh faktor klinis dan demografis (Kurniawati et al., 2025). Nilai *True Positive* (TP) yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar individu berisiko tinggi, terutama karena fitur usia dan hipertensi memiliki pola distribusi yang sangat berbeda antara kelompok kasus dan kontrol (Görtler et al., 2022).

Sensitivitas 73,11% mengindikasikan bahwa mekanisme pembelajaran ensemble pada *Random Forest* berhasil mengekstraksi ciri-ciri risiko yang paling menonjol pada kelompok positif, seperti tekanan darah tinggi dan IMT tinggi, yang secara epidemiologis merupakan prediktor kuat penyakit jantung (Guha et al., 2025). Di sisi lain, nilai *True Negative* (TN) yang tinggi dan spesifisitas 71,26% mencerminkan bahwa model juga efektif mengenali kelompok sehat dengan ciri stabil seperti tekanan darah normal dan IMT seimbang, sehingga risiko *over-detection* dapat diminimalkan (Saptawan et al., 2024). Distribusi *False Positive* (FP) dan *False Negative* (FN) menggambarkan tantangan tipikal dalam pemodelan penyakit kronis berbasis survei populasi, terutama FN yang masih muncul pada sebagian subjek berpenyakit yang tidak terdeteksi oleh model, kondisi ini dapat terjadi karena variabilitas diagnosis mandiri, keterbatasan informasi klinis, atau pola risiko yang tidak sepenuhnya tertangkap oleh atribut survei (Ratan, 2022).

Keseimbangan antara sensitivitas dan spesifisitas menunjukkan bahwa *Random Forest* mampu bekerja secara proporsional pada kedua kategori kelas tanpa bias signifikan, suatu keunggulan yang sejalan dengan teori *ensemble learning* yang mengutamakan stabilitas prediksi melalui agregasi banyak pohon keputusan (Nuraeni, 2024). Secara keseluruhan, pembahasan ini menunjukkan bahwa integrasi faktor biologis dominan dan struktur pembelajaran model menghasilkan performa klasifikasi yang kuat, sekaligus menegaskan relevansi algoritma *Random Forest* sebagai pendekatan prediktif yang handal untuk mendukung skrining dan pemetaan risiko penyakit jantung berbasis data populasi nasional (Alwakid et al., 2025).

Selain itu, pola kontribusi variabel pada penelitian ini mengindikasikan bahwa beban penyakit jantung di Indonesia masih sangat dipengaruhi oleh faktor risiko klasik yang telah lama dikenal, tetapi dengan dinamika yang semakin kompleks seiring perubahan gaya hidup dan transisi epidemiologi. Fenomena ini terlihat dari rendahnya nilai *feature importance* pada variabel perilaku seperti aktivitas fisik, konsumsi alkohol, dan porsi buah serta sayur, yang dalam banyak studi internasional sering muncul sebagai determinan kuat. Kondisi ini dapat dijelaskan melalui karakteristik populasi Indonesia yang memiliki variasi perilaku kesehatan yang lebih sempit pada survei populasi, sehingga sinyal risiko yang ditimbulkan lebih lemah dibandingkan faktor biologis yang sifatnya lebih stabil dan terukur.

Temuan ini sejalan dengan laporan Global Burden of *Disease* (GBD) yang menegaskan bahwa penyakit kardiovaskular di negara berpenghasilan menengah cenderung lebih dipicu oleh faktor klinis yang kurang terkontrol, terutama hipertensi dan obesitas, dibandingkan faktor perilaku individual (Si et al., 2025). Oleh karena itu, hasil penelitian ini tidak hanya menggambarkan struktur risiko individu, tetapi juga mencerminkan pola kesehatan masyarakat Indonesia yang menempatkan faktor klinis

sebagai prioritas penanganan paling mendesak dalam pengendalian penyakit jantung secara nasional.

Jika dilihat dari perspektif metodologis, keberhasilan *Random Forest* dalam menghasilkan performa evaluasi yang stabil menunjukkan bahwa algoritma ini mampu beradaptasi terhadap karakteristik data kesehatan populasi yang seringkali tidak memenuhi asumsi linearitas, normalitas, dan homogenitas yang dibutuhkan pada model statistik klasik (Hidayat et al., 2023). Mekanisme *bagging* dan pemilihan fitur acak memungkinkan model mempelajari pola variasi risiko pada subkelompok populasi tanpa terdistorsi oleh multikolinearitas atau distribusi variabel yang tidak seimbang (Guha et al., 2025). Hal ini terbukti penting ketika menangani data survei seperti SKI yang memiliki kompleksitas tinggi dan representasi multidimensi, sehingga model lebih mampu menangkap interaksi antarvariabel yang bersifat non-linear, misalnya hubungan simultan antara usia lanjut, tekanan darah tinggi, dan IMT tinggi sebagai triad risiko metabolik utama.

Kinerja model yang seimbang pada metrik akurasi, sensitivitas, dan spesifisitas juga menunjukkan bahwa pemanfaatan algoritma ensemble dapat menjadi pendekatan strategis dalam epidemiologi digital, terutama untuk mendukung sistem skrining otomatis, pemetaan risiko wilayah, atau penargetan intervensi kesehatan masyarakat berbasis bukti (Tasnim et al., 2025). Temuan ini memberikan implikasi bahwa integrasi *Machine Learning* pada surveilans kesehatan nasional bukan hanya memungkinkan deteksi risiko yang lebih tepat, tetapi juga dapat meningkatkan efisiensi proses pengambilan keputusan pada level populasi.

Meskipun model menunjukkan performa yang cukup baik, hasil ini tidak dapat diinterpretasikan sebagai hubungan kausal antara variabel prediktor dan kejadian penyakit jantung. Pendekatan Machine Learning dalam penelitian ini bersifat prediktif dan eksploratif, sehingga hasil yang diperoleh lebih menekankan pada kemampuan model dalam mengenali pola data dibandingkan menjelaskan hubungan sebab-akibat secara langsung.

KESIMPULAN

Penelitian ini berhasil mengembangkan model klasifikasi risiko penyakit jantung berbasis algoritma *Random Forest* menggunakan data mikro Survei Kesehatan Indonesia (SKI) 2023. Model menunjukkan performa prediktif yang kuat dengan akurasi 72,13%, sensitivitas 73,11%, dan spesifisitas 71,26%, menandakan stabilitas klasifikasi pada data populasi besar. Analisis *feature importance* mengidentifikasi usia, hipertensi, dan Indeks Massa Tubuh (IMT) sebagai faktor risiko paling dominan dalam membedakan kelompok berisiko dan tidak berisiko. Dengan demikian, tujuan penelitian untuk membangun model klasifikasi dan menentukan determinan utama risiko penyakit jantung telah tercapai secara tegas. Hasil ini menegaskan bahwa faktor biologis memiliki peran sentral dalam risiko penyakit jantung pada populasi dewasa Indonesia. Namun demikian, penelitian ini memiliki keterbatasan, antara lain penggunaan data sekunder yang bergantung pada kualitas input survei, potensi bias klasifikasi, serta belum dilakukannya validasi eksternal pada dataset lain. Oleh karena itu, penelitian selanjutnya disarankan untuk melakukan validasi model secara eksternal serta menggunakan desain prospektif untuk meningkatkan kekuatan inferensi. Secara praktis, hasil penelitian ini dapat dimanfaatkan sebagai dasar pengembangan sistem skrining risiko penyakit jantung berbasis *Machine Learning* pada layanan kesehatan primer, khususnya dalam mendukung deteksi dini berbasis populasi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan dalam proses pelaksanaan penelitian, termasuk institusi penyedia data, lingkungan akademik, serta semua pihak yang turut membantu dalam analisis, penyusunan, dan penyempurnaan artikel. Kontribusi dan bantuan tersebut sangat berarti dalam terselesaikannya penelitian ini.

DAFTAR PUSTAKA

Ahmed, M., Husien, I., 2024. Heart Disease Prediction using Hybrid Machine Learning : A Brief Review. Journal of Robotic and Control 5(3), 884–892. https://journal.umy.ac.id/index.php/jrc/article/view/21606

Alwakid, G., Ul Haq, F., Tariq, N., Humayun, M., Shaheen, M., Alsadun, M., 2025. Optimized Machine Learning Framework for Cardiovascular Disease Diagnosis: A Novel Ethical Perspective. BMC Cardiovascular Disorders 25(1), 1-28. Https://Doi.Org/10.1186/S12872-025-04550-W

[BKPK] Badan Kebijakan Pembangunan Kesehatan., 2021. National Health Accounts Indonesia Tahun 2020. BKPK Kemenkes RI, Jakarta.

Chandra, K., Prasetyo, J.S., 2024. Prediksi Penyakit Jantung Koroner Menggunakan Metode K-NN dan Regresi Logistik Berdasarkan Kerangka Kerja CRISP-DM. [Prosiding]. Seminar Nasional Ma Chung Sistem Informasi & Informatika, 4, 241–248.

Delavera, A., Siregar, K.N., Jazid, R., Eryando, T., 2021. Hubungan Kondisi Psikologis Stress dengan Hipertensi pada Penduduk Usia di atas 15 Tahun di Indonesia. Jurnal Biostatistik, Kependudukan, dan Informatika Kesehatan 1(3), 148–159. https://scholarhub.ui.ac.id/bikfokes/vol1/iss3/2/

Görtler, J., Hohman, F., Moritz, D., Kirchner, M., 2022. Neo: Generalizing Confusion Matrix Visualization to Hierarchical and Multi-Output Labels. [Prosiding]. CHI Conference on Human Factors in Computing Systems 408, 1-13. Https://Doi.Org/10.1145/3491102.3501823

Guha, A., Shah, V., Nahle, T., Singh, S., Kunhiraman, H.H., Shehnaz, F., Nain, P., Makram, O.M., Mahmoudi, M., Al-Kindi, S., Madabhushi, A., Shiradkar, R., Daoud, H., 2025. Artificial Intelligence Applications in Cardio-Oncology: A Comprehensive Review. Current Cardiology Reports 27(1), 1-22. Https://Doi.Org/10.1007/S11886-025-02215-W

Hidayat, H., Sunyoto, A., Al-Fatta, H., 2023. Klasifikasi Penyakit Jantung Menggunakan Random Forest Classifier. Jurnal Sistem Komputer dan Kecerdasan Buatan 7(1), 31–40. Https://Doi.Org/10.47970/Siskom-Kb.V7i1.464

Islam, M.M., Alam, M.J., Maniruzzaman, M., Ahmed, N.A.M.F., Ali, M.S., Rahman, M.J., Roy, D.C., 2023. Predicting The Risk of Hypertension using Machine Learning Algorithms: A Cross Sectional Study in Ethiopia. Plos One 18(8), 1–20. Https://Doi.Org/10.1371/Journal.Pone.0289613

Khalisatifa, A., Arum, H.D., Jambak, M.I., 2024. Klasifikasi Risiko Penyakit Serangan Jantung dengan Menggunakan Algoritma C4.5. Jurnal Ilmiah dan Teknologi 14(1), 57–64. Https://Doi.Org/10.32699/Device.V14i1.6869

Kurniawati, L., Priyanto, D., Sulistianingsih, N., Syahrir, M., Rismawati, R., 2025. Perbandingan Metode Berbasis Decision Tree untuk Mendeteksi Penyakit Paru Comparison of Decision Tree-Based Methods in Lung Disease Detection 7(1),

51–62. Https://Doi.Org/10.30812/Bite.V7i1.4909

Naser, M.A., Majeed, A.A., Alsabah, M., Al-Shaikhli, T.R., Kaky, K.M., 2024. A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges. Algorithms 17(2), 1–33. Https://Doi.Org/10.3390/A17020078

Nazari, M., Emami, H., Rabiei, R., Hosseini, A., Rahmatizadeh, S., 2024. Detection of Cardiovascular Diseases Using Data Mining Approaches: Application of An Ensemble-Based Model. Cognitive Computation 16(5), 2264–2278. Https://Doi.Org/10.1007/S12559-024-10306-Z

Nuraeni, N., 2024. Klasifikasi Data Mining untuk Prediksi Penyakit Kardiovaskular. Jurnal Teknik Informasi dan Komputer 7(1), 161-169. https://jurnal.murnisadar.ac.id/index.php/Tekinkom/article/view/1276

Pal, K., Patel, B.V., 2020. Data Classification with K-Fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques. 2020 Fourth International Conference on Computing Methodologies and Communication 83–87. Https://Doi.Org/10.1109/ICCMC48092.2020.ICCMC-00016

[PERKI] Perhimpunan Dokter Spesialis Kardiovaskular Indonesia., 2022. Panduan Prevensi Penyakit Kardiovaskular Arterosklerosis. Perhimpunan Dokter Spesialis Kardiovaskular Indonesia, Jakarta.

Ramadhan, B., Firdaus, D., Rafi, A.R., 2023. Teknik SMOTE sebagai Solusi Imbalance Class dalam Model Deteksi Intrusi DDoS dengan Metode PCA-Random Forest. Multimedia Artificial Intelligence Networking Database 8(1), 52–64. https://ejurnal.itenas.ac.id/index.php/mindjournal/article/view/8161

Ramadhanti, I., Izzati, M.N., Nurcandra, F., Apriningsih, A., 2024. Studi Kualitatif: Program Penanggulangan Penyakit Jantung dan Pembuluh Darah di Kementerian Kesehatan RI. Jurnal Kesehatan Tambusai 5(3), 757-824. https://journal.universitaspahlawan.ac.id/index.php/jkt/article/view/29796

Ratan, U., 2022. Applied Machine Learning for Healthcare and Life Sciences using AWS. Packt Publishing Ltd, Birmingham.

Reddy, S.P., Reddy, C.V.K., Sambath, M., Thangakumar, J., 2024. Heart Disease Prediction using Machine Learning Techniques. Communications in Computer and Information Science, 27–35. Https://Doi.Org/10.1007/978-3-031-75957-4_3

Rusyda, A.L., 2025. Exploring The Non-Communicable Disease Burden in Indonesia Findings from The 2023 Health Survey. Indonesia Journal of Public Health Nutrition 5(2), 1-16. Https://Doi.Org/10.7454/Ijphn.V5i2.1064

Saptawan, F., David, D., Wijaya, T., Kosasi, S., Kuway, S.M., 2024. Prediksi Epidemiologi Penyakit Tidak Menular Menggunakan Algoritma Random Forest pada Puskesmas. Jurnal Times 13(2), 192–201. Https://Doi.Org/10.51351/Jtm.13.2.2024788

Saputra, I., 2023. Belajar Mudah Data Mining untuk Pemula. Penerbit Informatika.

[SKI] Survei Kesehatan Indonesia., 2023. Laporan Survei Kesehatan Indonesia (SKI) 2023. Survei Kesehatan Indonesia, Jakarta.

Si, Y., Guo, L., Chen, S., Zhang, X., Dai, X., Wang, D., Liu, Y., Tran, B. X., Pronyk, P. M., Tang, S., 2025. Progressing Towards The 2030 Health-Related SDGs in ASEAN: A Systematic Analysis. Plos Medicine 22(4), 1–20. Https://Doi.Org/10.1371/Journal.Pmed.1004551

Solida, A., Noerjoedianto, D., Mekarisce, A.A., Widiastuti, F., 2021. Pola Belanja Kesehatan Katastropik Peserta Jaminan Kesehatan di Kota Jambi. Jurnal

Kebijakan Kesehatan Indonesia 10(4), 209–215. https://journal.ugm.ac.id/jkki/article/view/68736

Tampubolon, L.F., Ginting, A., Turnip, F.E., 2023. Gambaran Faktor yang Mempengaruhi Kejadian Penyakit Jantung Koroner (PJK) di Pusat Jantung Terpadu (PJT). Jurnal Ilmiah Permas: Jurnal Ilmiah STIKES Kendal, 13(3), 1043–1052. Https://Doi.Org/10.32583/Pskm.V13i3.1077

Tasnim, A.F., Rahman, R., Prabha, M., Hossain, M.A., Nilima, S.I., Mahmud, M.A., Erdei, T.I., 2025. Explainable Machine Learning Algorithms to Predict Cardiovascular Strokes. Engineering, Technology and Applied Science Research 15(1), 20131–20137. Https://Doi.Org/10.48084/Etasr.9152

Vinet, L., Zhedanov, A., 2011. A "Missing" Family of Classical Orthogonal Polynomials. Journal of Physics A: Mathematical and Theoretical 44(8), 1–14. Https://Doi.Org/10.1088/1751-8113/44/8/085201

Wardhana, R.G., Wang, G., Sibuea, F., 2023. Penerapan Machine Learning dalam Prediksi Tingkat Kasus Penyakit di Indonesia. Journal of Information System Management 5(1), 40–45. Https://Doi.Org/10.24076/Joism.2023v5i1.1136