Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3234~3243
e-ISSN: 2548-6861

Comparative Analysis of Random Forest, SVM, and Naive Bayes for Cardiovascular Disease Prediction

Windy Aldora Rayadhani [1]*, Majid Rahardi [2]*
* Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta
windyaldoraya@students.amikom.ac.id [1], majid@amikom.ac.id [2]

Article Info

Article history:

Received 2025-10-07
Revised 2025-10-30
Accepted 2025-11-08

Keyword:

Cardiovascular Disease,
Random Forest,
SVM,
Naïve Bayes,
Clinical Decision Support.

ABSTRACT

Cardiovascular disease is one of the leading causes of death worldwide; therefore, accurate early detection is essential to reduce fatal risks. This study aims to compare the performance of three machine learning algorithms — Random Forest, Support Vector Machine (SVM), and Naïve Bayes — in predicting cardiovascular disease risk using the Mendeley Cardiovascular Disease Dataset, which contains 1,000 patient records and 14 clinical attributes. The models were evaluated using accuracy, precision, recall, and F1-score metrics, and their performance differences were statistically tested using the paired t-test. The experimental results indicate that the Random Forest algorithm achieved the best performance with 99% accuracy, 100% recall, 98% precision, and an F1-score of 99%. The SVM model followed with 98% accuracy and 100% recall, while the Naïve Bayes algorithm obtained 94.5% accuracy and an F1-score of 95%. The p-value < 0.05 confirmed that the performance differences among the three models were statistically significant. From a clinical perspective, a model with high recall, such as Random Forest, is more desirable because it reduces the likelihood of false negatives, which are critical in heart disease diagnosis. The feature importance analysis also revealed that age, resting blood pressure, and cholesterol level were the most influential factors in predicting cardiovascular risk. These findings suggest that machine learning algorithms, particularly Random Forest, have strong potential to be implemented in Clinical Decision Support Systems (CDSS) for accurate and efficient early detection of cardiovascular disease.

I. PENDAHULUAN

Berdasarkan data dari World Health Organization (WHO), Penyakit Kardiovaskular (PKV) merupakan penyebab utama kematian di seluruh dunia. Setiap tahunnya, lebih dari 17 juta orang meninggal akibat penyakit jantung dan pembuluh darah, dan angka ini diperkirakan akan terus meningkat seiring dengan perubahan gaya hidup yang tidak sehat serta peningkatan faktor risiko [1]. Penyakit tidak menular seperti PKV menunjukkan peningkatan yang signifikan dan juga menjadi tantangan utama dalam sistem kesehatan. Perkembangan teknologi informasi yang pesat telah menghasilkan volume data yang sangat besar di berbagai sektor, termasuk sektor kesehatan [2]. Dalam konteks ini, data yang tersedia semakin banyak, mulai dari rekam medis pasien, data gaya hidup, riwayat penyakit, hingga hasil tes laboratorium. Namun, data tersebut belum seluruhnya dimanfaatkan secara maksimal untuk mendukung pengambilan keputusan medis yang lebih baik dan cepat. Oleh karena itu, dibutuhkan metode yang mampu menggali informasi tersembunyi dari kumpulan data tersebut agar dapat dimanfaatkan secara optimal. Salah satu metode yang dapat digunakan untuk tujuan ini adalah data mining [3].

Data mining adalah proses menemukan pola data terpilih dengan menggunakan metode tertentu, yang merupakan bagian dari proses Knowledge Discovery in Database (KDD). Melalui proses KDD, informasi penting dari data yang besar dapat disederhanakan dan dijadikan pengetahuan yang bermanfaat. Salah satu teknik dalam data mining yang banyak digunakan adalah klasifikasi, yaitu proses pengelompokan data ke dalam kategori tertentu berdasarkan atribut-atribut yang dimilikinya [3].

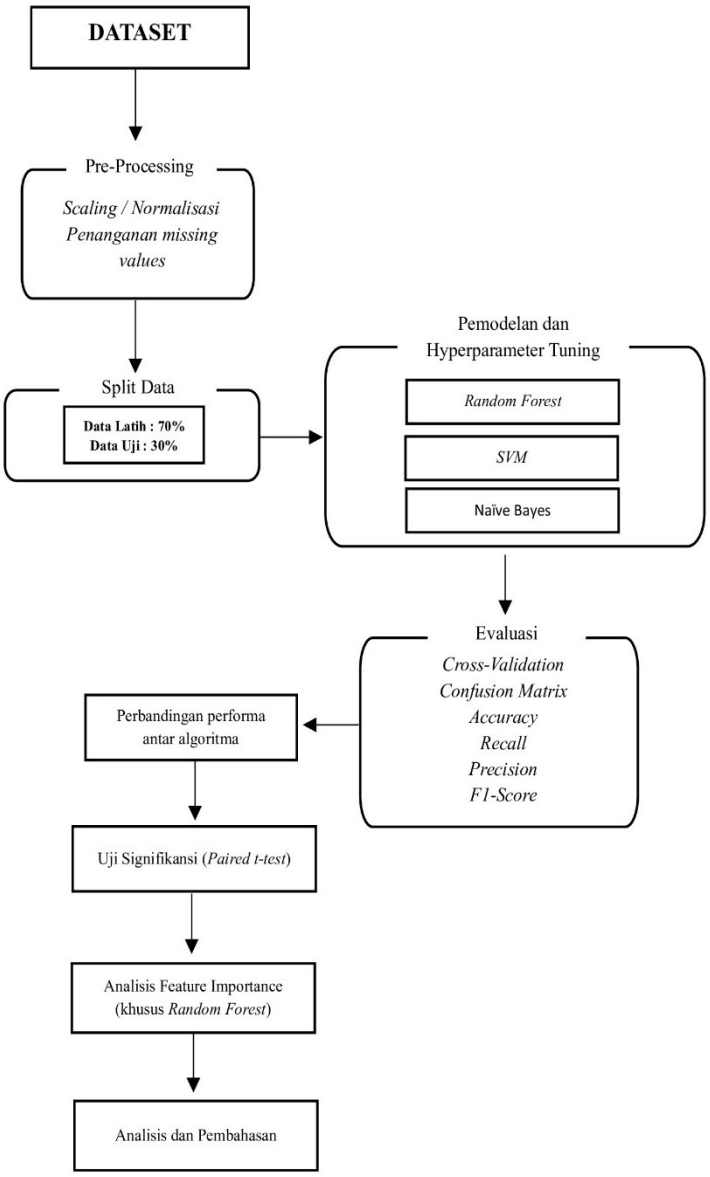
Di Indonesia, salah satu tantangan utama dalam pembangunan sektor kesehatan adalah beban ganda penyakit, di mana di satu sisi masih banyak kasus penyakit infeksi yang perlu diatasi, sementara di sisi lain, penyakit tidak menular, khususnya penyakit jantung dan pembuluh darah, mengalami peningkatan yang signifikan[4][5]. Angka kematian akibat penyakit tidak menular meningkat dari 41,7% pada tahun 1995 menjadi 59,5% pada tahun 2007 [4].

Berdasarkan data yang ada, penyakit jantung koroner merupakan salah satu penyakit yang sangat berbahaya dan memerlukan perhatian serius. Salah satu cara untuk mengatasi hal ini adalah dengan mendeteksi penyakit jantung koroner sejak dini melalui pengembangan sistem berbasis Machine Learning yang dapat menganalisis kondisi tubuh seseorang dengan akurasi yang tinggi. Oleh karena itu, penting untuk menggunakan algoritma yang mampu memprediksi penyakit jantung koroner secara tepat [6].

Sejumlah penelitian terbaru menunjukkan bahwa algoritma machine learning telah banyak digunakan dalam memprediksi penyakit kardiovaskular. Penelitian oleh Nasution et al. [7] pada tahun 2025 melakukan evaluasi pada beberapa algoritma, termasuk Random Forest dan SVM, menggunakan dataset penyakit jantung UCI. Hasilnya menunjukkan bahwa Random Forest mencapai akurasi 89,7% sedangkan SVM memperoleh akurasi 87,0%, sehingga Random Forest lebih unggul dalam penelitian tersebut. Penelitian lainnya, seperti yang dilakukan Hasanah [8] pada tahun 2022, menggunakan algoritma Naïve Bayes dan menghasilkan akurasi pengujian sebesar 83,78% untuk pasien yang memiliki penyakit jantung dan 87,50% untuk pasien yang tidak memiliki penyakit jantung. Selanjutnya, Adinulhaq dan Sam'an [9] pada tahun 2023, membandingkan beberapa model *machine learning*, termasuk Logistic Regression, Random Forest, dan SVM, pada dataset yang sama dengan hasil akurasi tertinggi sebesar 90%. Namun, penelitian tersebut belum mencakup penerapan *hyperparameter tuning* maupun *cross-validation* yang penting untuk meningkatkan stabilitas dan kemampuan generalisasi model. Penelitian oleh Kholish et al. [10] pada tahun 2024 membandingkan algoritma Random Forest dan Naïve Bayes pada kasus medis, yaitu prediksi penyakit diabetes, dan hasilnya menunjukkan bahwa Random Forest memberikan performa yang lebih baik dan stabil dibandingkan Naïve Bayes. Temuan tersebut memperkuat bukti bahwa Random Forest juga berpotensi unggul pada kasus prediksi penyakit jantung yang memiliki karakteristik klinis serupa dalam konteks data medis.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk memprediksi ada atau tidaknya penyakit kardiovaskular pada seseorang menggunakan tiga metode *supervised machine learning*, yaitu Random Forest, Support Vector Machine (SVM), dan Naïve Bayes. Penelitian ini menggunakan Mendeley Cardiovascular Disease Dataset [11] yang berisi 1.000 data pasien dengan 14 atribut klinis. Tahapan penelitian meliputi pengumpulan dataset, proses *preprocessing* berupa penanganan nilai kosong (*missing values*), serta normalisasi fitur menggunakan *StandardScaler* untuk menyeragamkan skala data. Model kemudian dibangun menggunakan ketiga algoritma tersebut dan dievaluasi berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score* dengan menerapkan 5-Fold Cross-Validation untuk menjamin kestabilan hasil. Selanjutnya, dilakukan *paired t-test* untuk menguji signifikansi perbedaan performa antar algoritma serta analisis *feature importance* pada Random Forest untuk mengidentifikasi atribut klinis yang paling berpengaruh terhadap prediksi. Hasil analisis tersebut dibahas dalam konteks medis guna menilai potensi penerapan model sebagai sistem pendukung keputusan dalam deteksi dini penyakit kardiovaskular.

II. METODE



Gambar 1. Alur Penelitian

A. Pengambilan Dataset

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari situs *Mendeley Data* pada tautan https://data.mendeley.com/datasets/dzz48mvjht/1 [11]. Dataset ini terdiri atas 1.000 entri data pasien dan mencakup 14 atribut klinis, termasuk informasi seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, hasil elektrokardiografi, serta status risiko penyakit kardiovaskular. Dataset ini banyak digunakan dalam penelitian terkait prediksi penyakit jantung karena memiliki variabel klinis yang representatif terhadap kondisi pasien. Selain itu, dataset telah melalui proses pembersihan dasar oleh penyedia sehingga tidak mengandung *missing value* signifikan. Distribusi kelas terdiri atas pasien dengan risiko penyakit kardiovaskular dan pasien tanpa risiko dengan proporsi yang relatif seimbang. Tabel 1 menyajikan informasi detail mengenai fitur-fitur yang digunakan dalam penelitian ini.

TABEL 1
DESKRIPSI DATASET

S.No	Attribute	Assigned Code	Unit	Type of the Data
1	Patient Identification Number	patientid	Number	Numeric
2	Age	age	In Years	Numeric
3	Gender	gender	1,0(0= female, 1 = male)	Binary
4	Chest pain type	chestpain	0,1,2,3 (Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic)	Nominal
5	Resting blood pressure	restingBP	94-200 (in mm HG)	Numeric
6	Serum cholesterol	serumcholestrol	126-564 (in mg/dl)	Numeric
7	Fasting blood sugar	Fastingbloodsugar	0,1 > 120 mg/dl (0 = false , 1 = true)	Binary
8	Resting electrocardiogram results	restingrelectro	0,1,2 (Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)	Nominal
9	Maximum heart rate achieved	maxheartrate	71-202	Numeric
10	Exercise induced angina	exerciseangia	0,1 (0 = no, 1 = yes)	Binary
11	Oldpeak =ST	oldpeak	0-6.2	Numeric
12	Slope of the peak exercise ST segment	slope	1,2,3 (1-upsloping, 2-flat, 3-downsloping)	Nominal
13	Number of major vessels	noofmajorvessels	0,1,2,3	Numeric
14	Classification	target	0,1 (0= Absence of Heart Disease, 1= Presence of Heart Disease)	Binary

B. Pre-Prossesing Data

Tahapan *pre-processing* dilakukan untuk memastikan kualitas data sebelum proses pelatihan model. Pemeriksaan terhadap nilai hilang (*missing values*) dilakukan menggunakan fungsi isnull() untuk setiap atribut, agar hasilnya menunjukkan bahwa dataset tidak memiliki nilai yang hilang sehingga seluruh data dapat digunakan tanpa proses imputasi tambahan. *pre-processing* berperan penting dalam menjamin kualitas data dengan mengatasi permasalahan seperti data yang hilang maupun tidak konsisten [12]. Selanjutnya, dilakukan proses normalisasi menggunakan metode *StandardScaler* untuk menyeragamkan skala antar fitur numerik agar setiap variabel memiliki kontribusi yang seimbang dalam proses pembelajaran model. Normalisasi diperlukan karena algoritma seperti *Support Vector Machine (SVM)* sensitif terhadap perbedaan skala antar fitur, dan scaling terbukti dapat meningkatkan stabilitas serta performa [13].

C. Split Data

Dataset dibagi menjadi dua bagian utama, yaitu data latih (training data) dan data uji (testing data). Pembagian dilakukan dengan rasio 70% untuk pelatihan dan 30% untuk pengujian. Tujuan pembagian ini adalah untuk melatih model pada sebagian data dan menguji akurasi prediksinya terhadap data yang belum pernah dilihat. Tabel 2 merupakan hasil split dataset.

TABEL 2
HASIL SPLIT DATA LATIH DAN DATA UJI

Deskripsi	Data Latih	Data Uji	Total
Proporsi (%)	70 %	30%	100%
Jumlah (entri)	700	300	1000

D. Hyperparameter Tuning GridSearchCV / RandomizedSearchCV

Dalam penelitian ini, dilakukan optimasi hyperparameter menggunakan GridSearchCV atau RandomizedSearchCV untuk memperoleh kombinasi parameter terbaik pada masing-masing algoritma klasifikasi. GridSearchCV digunakan untuk mengevaluasi seluruh kombinasi parameter yang telah ditentukan, sedangkan RandomizedSearchCV digunakan untuk mempercepat pencarian pada ruang parameter yang lebih luas dengan pemilihan kombinasi secara acak. Proses tuning dilakukan melalui validasi silang (cross-validation) untuk menghindari bias dan memastikan model yang diperoleh memiliki performa yang konsisten[14].

E. Pengujian Algoritma Random Forest

Algoritma Random Forest digunakan sebagai metode klasifikasi dalam penelitian ini. Random Forest merupakan algoritma berbasis ensemble learning yang membentuk banyak pohon keputusan (decision tree) dari subset data yang berbeda. Hasil akhir prediksi ditentukan berdasarkan voting mayoritas dari seluruh pohon. Model dilatih menggunakan pustaka Scikit-learn dalam bahasa Python dengan beberapa parameter utama yang diatur sebagai Tabel 3:

TABEL 3
PARAMETER RANDOM FOREST

Parameter	Keterangan
n_estimators	Jumlah pohon dalam hutan (semakin besar dapat meningkatkan akurasi)
max_depth	Kedalaman maksimum setiap pohon
min_samples_split	Jumlah minimum sampel untuk memisahkan node
min_samples_leaf	Jumlah minimum sampel di setiap daun (leaf)
criterion	Fungsi pengukuran untuk kualitas split (contoh: gini, entropy)
random_state	Seed untuk memastikan hasil yang konsisten pada setiap eksekusi

Parameter-parameter ini dapat disesuaikan (tuning) untuk mendapatkan performa model yang lebih optimal.

F. Pengujian Algoritma Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan salah satu metode klasifikasi yang populer dalam pembelajaran mesin karena kemampuannya mencari *hyperplane* terbaik yang memisahkan kelas dengan margin maksimal. Keunggulan utama SVM terletak pada fleksibilitas penggunaan berbagai fungsi kernel sehingga dapat menangani data yang linear maupun non-linear [15]. Proses normalisasi dan seleksi fitur sangat penting untuk mendukung performa SVM, sebab algoritma ini sensitif terhadap skala data dan kualitas atribut yang digunakan [16]. SVM mampu mengolah data dengan dimensi tinggi, misalnya pada data genetik penyakit kardiovaskular, tanpa mengalami masalah serius terkait *overfitting* yang sering muncul pada algoritma lain [17]. Hal ini menunjukkan bahwa SVM tidak hanya unggul dalam aspek akurasi, tetapi juga memiliki kekuatan dalam generalisasi, kemampuan menangani kompleksitas data, serta ketahanan terhadap data berdimensi besar.

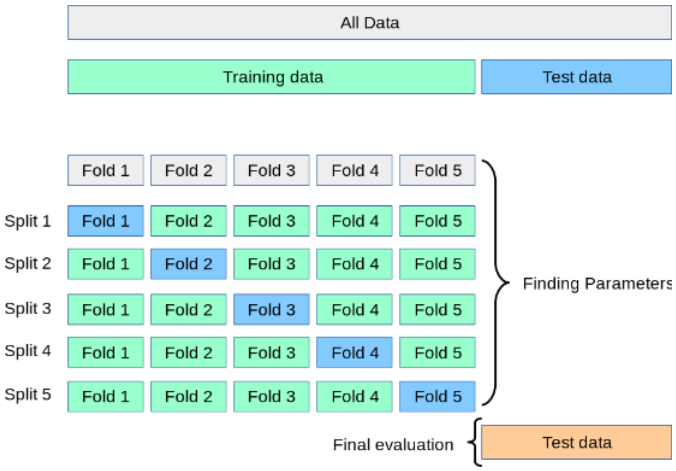
G. Pengujian Algoritma Naïve Bayes

Naïve Bayes merupakan algoritma klasifikasi berbasis probabilistik yang sederhana namun efektif, dengan asumsi bahwa setiap atribut saling independen dalam memengaruhi kelas. Keunggulan utama Naïve Bayes adalah kemampuannya bekerja dengan baik pada dataset berukuran besar maupun kecil, serta efisiensi komputasinya yang tinggi karena perhitungan peluang relatif sederhana. Naïve Bayes dapat menjadi alternatif yang kompetitif dalam prediksi penyakit diabetes, terutama karena mampu menangani data dengan variabel klinis yang beragam [9]. Selain itu, Naïve Bayes sering memberikan hasil stabil meskipun tanpa tuning parameter yang kompleks, sehingga cocok digunakan pada penelitian kesehatan dengan sumber daya terbatas [10]. Namun demikian, kelemahan dari algoritma ini adalah asumsi independensi antar fitur yang tidak selalu sesuai dengan kondisi nyata, sehingga preprocessing data seperti seleksi fitur dan normalisasi tetap penting agar model dapat menghasilkan prediksi yang lebih akurat. Dengan karakteristik tersebut, Naïve Bayes layak diuji bersamaan dengan algoritma lain seperti Random Forest dan SVM untuk memprediksi risiko penyakit kardiovaskular.

H. Evaluasi Metode

Evaluasi terhadap performa model dilakukan menggunakan klasifikasi, yaitu:

1) Cross Validation. K-fold cross validation adalah metode statistik yang digunakan untuk menilai kinerja suatu model yang akan dikembangkan [18]. Penelitian ini menggunakan metode k-fold cross-validation dengan nilai k = 5 untuk mengevaluasi performa model dan mencegah overfitting. Teknik ini dilakukan dengan membagi dataset menjadi lima bagian (fold) yang sama besar setiap bagian secara bergantian untuk digunakan sebagai data uji, sementara sisanya digunakan untuk pelatihan. Proses ini diulang sebanyak lima kali, dan nilai rata-rata dari seluruh hasil pengujian digunakan sebagai evaluasi akhir performa model. Gambar 2 merupakan Alur kerja cross-validation



Gambar 2. Contoh simulasi crossvalidation

2) Accuracy (Akurasi). Menunjukkan efektivitas keseluruhan hasil klasifikasi. Akurasi dihitung dengan rumus:

$$Accuracy\% = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

3) Precision (Presisi). Menunjukkan persentase label positif yang diklasifikasikan dengan benar.

$$Precision\% = \frac{TP}{TP + FP} \times 100\%$$

4) Recall (Sensitivitas). Menunjukkan kemampuan model untuk menangkap semua data yang sebenarnya positif.

$$Recall\% = \frac{TP}{TP + FN} \times 100\%$$

5) F1-Score. Merupakan rata-rata harmonis antara precision dan recall:

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

I. Paired Test

Uji *Paired Sample t-Test* digunakan dalam penelitian ini untuk membandingkan performa tiga algoritma klasifikasi: Random Forest, SVM, dan Naive Bayes. Setiap algoritma dievaluasi menggunakan metode 5-fold cross validation pada dataset yang sama, sehingga diperoleh nilai performa berpasangan untuk masing-masing fold. Untuk menentukan apakah terdapat perbedaan signifikan secara statistik, dilakukan uji paired t-test secara berpasangan antara setiap kombinasi algoritma, yaitu Random Forest vs SVM, Random Forest vs Naive Bayes, dan SVM vs Naive Bayes, dengan tingkat signifikansi (α) sebesar 0,05. Jika nilai *p-value* < 0,05, perbedaan performa antar algoritma dianggap signifikan secara statistik; sebaliknya, jika *p-value* ≥ 0,05, tidak terdapat perbedaan signifikan antara kedua algoritma yang diuji.\

J. Analisis Feature Importance (khusus Random Forest)

Analisis feature importance dilakukan untuk menilai kontribusi setiap fitur terhadap performa model Random Forest. Nilai kepentingan fitur dihitung berdasarkan besarnya pengurangan impuritas (Gini Importance) yang terjadi ketika suatu fitur digunakan dalam proses pemisahan node pada pohon Keputusan. Model Random Forest dilatih menggunakan dataset yang telah melalui tahap preprocessing dan evaluasi menggunakan 5-Fold Cross Validation, sehingga setiap fitur memperoleh nilai kepentingannya berdasarkan kontribusinya terhadap hasil prediksi yang akurat. Hasil analisis feature importance selanjutnya dibahas pada bagian Hasil dan Pembahasan untuk mengidentifikasi fitur-fitur yang paling berpengaruh terhadap performa model dalam memprediksi risiko penyakit kardiovaskular.

Analisis *feature importance* difokuskan pada model Random Forest, karena algoritma ini secara bawaan mampu menghitung tingkat kepentingan fitur berdasarkan pengurangan impuritas (*Gini Importance*). Model SVM dan Naïve Bayes tidak memiliki mekanisme serupa, sehingga interpretasi fitur hanya dilakukan pada Random Forest.

III. HASIL DAN PEMBAHASAN

A. Pengambilan Dataset

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diambil dari *Mendeley Data Repository* melalui tautan https://data.mendeley.com/datasets/dzz48mvjht/1 [11]. Dataset ini berisi 1.000 data pasien dengan 14 atribut klinis, antara lain usia, jenis kelamin, tekanan darah, kolesterol, hasil elektrokardiografi, serta status risiko penyakit kardiovaskular. Dataset telah melalui proses pembersihan awal oleh penyedia, sehingga tidak terdapat *missing values* maupun data duplikat. Distribusi kelas antara pasien berisiko dan tidak berisiko relatif seimbang, Variabel target bersifat biner, dengan nilai 1 menunjukkan pasien terdiagnosis penyakit jantung, dan 0 menunjukkan tidak terdiagnosis. Dari total data, terdapat 580 data dengan target 1 (58%) dan 420 data dengan target 0 (42%). Distribusi tersebut menunjukkan bahwa dataset relatif seimbang sehingga tidak diperlukan penanganan khusus terhadap ketidakseimbangan kelas. sehingga dataset ini layak digunakan untuk penelitian klasifikasi berbasis *machine learning*. Gambar 3 menampilkan sebagian contoh struktur data yang digunakan dalam penelitian ini.

patientid	age	gender	chestpain	restingBP	serumcholestrol	fastingbloodsugar
103368	53	1	2	171	0	0
119250	40	1	0	94	229	0
119372	49	1	2	133	142	0
132514	43	1	0	138	295	1
146211	31	1	1	199	0	0
148462	24	1	1	173	0	0
168686	79	1	2	130	240	0
170498	52	1	0	127	345	0
188225	62	1	0	121	357	0

restingrelectro	maxheartrate	exerciseangia	oldpeak	slope	noofmajorvessels	target
1	147	0	53	3	3	1
1	115	0	37	1	1	0
0	202	1	5	1	0	0
1	153	0	32	2	2	1
2	136	0	53	3	2	1
0	161	0	47	3	2	1
2	157	0	25	2	1	1
0	192	1	49	1	0	0
1	138	0	28	0	0	0

Gambar 3. Sebagian data pada Dataset.

B. Preprocessing Data

Tahap preprocessing data bertujuan untuk memastikan bahwa data yang digunakan dalam proses pelatihan model berada dalam kondisi bersih, terstandar, dan siap diolah. Proses ini mencakup serangkaian langkah untuk meningkatkan kualitas data dan menghindari bias pada hasil analisis. Pada penelitian ini, tahapan preprocessing yang dilakukan difokuskan pada dua langkah utama, yaitu pemeriksaan missing values dan proses normalisasi data, dengan penjelasan sebagai berikut.

1) Pemeriksaan dan penanganan missing values. Berdasarkan hasil eksekusi kode pada Gambar 4, diperoleh keluaran "No missing data", yang menunjukkan bahwa dataset telah bersih dari missing values serta siap digunakan untuk tahap selanjutnya.

```
# Checking for missing data
print("No missing data") if sum(df.isna().sum()) == 0 else df.isna().sum()

No missing data
```

Gambar 4. Hasil pengecekan terdapat missing data atau tidak.

2) Scalling/Normalization. Untuk SVM dan Naïve Bayes Pada dataset ini, fitur numerik seperti age, restingBP, serumcholestrol, maxheartrate, dan oldpeak memiliki skala berbeda yang dapat menimbulkan bias pada algoritma, terutama SVM. Untuk mengatasinya, dilakukan normalisasi dengan StandardScaler sehingga setiap fitur memiliki rata-rata 0 dan standar deviasi 1. Tabel 4 menampilkan perbandingan nilai rata-rata dan standar deviasi sebelum dan sesudah normalisasi.

TABEL 4
RATA-RATA DAN STANDAR DEVISIASI FITUR SEBELUM DAN SESUDAH NORMALISASI

Fitur	Mean Sebelum	Std Sebelum	Mean Sesudah	Std Sesudah
Age	49.24	17.86	0.00	1.00
RestingBP	151.75	29.97	-0.00	1.00
SerumCholestrol	311.45	132.44	-0.00	1.00
MaxHeartRate	145.48	34.19	-0.00	1.00
OldPeak	2.71	1.72	0.00	1.00

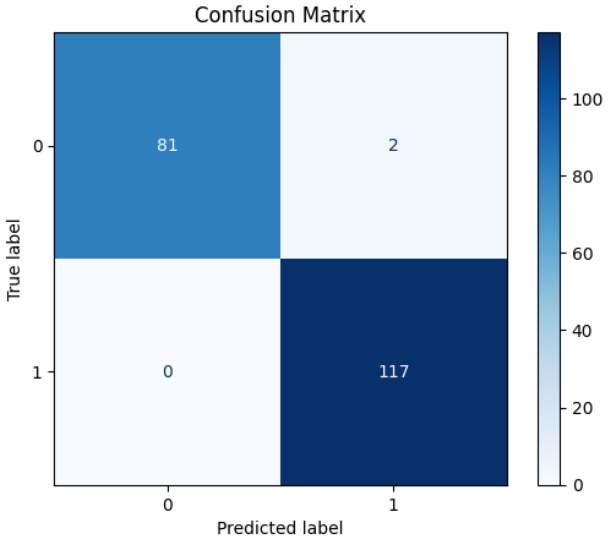
Setelah normalisasi, seluruh fitur memiliki rata-rata mendekati 0 dan standar deviasi mendekati 1, sehingga tidak ada fitur yang mendominasi akibat perbedaan skala. Proses ini diterapkan pada algoritma SVM dan Naïve Bayes. Pada SVM, Normalisasi meningkatkan performa model secara signifikan, sedangkan pada Naïve Bayes membantu menjaga konsistensi distribusi fitur untuk menghasilkan prediksi yang lebih stabil.

C. *Split Data*

Pada penelitian ini, dataset dibagi menjadi dua bagian, yaitu 70% sebagai data latih (training data) dan 30% sebagai data uji (testing data). Proporsi 70:30 dipilih karena standar yang umum digunakan dalam penelitian machine learning untuk memastikan adanya data latih yang cukup banyak agar model dapat mengenali pola dengan baik, sekaligus menyediakan data uji yang memadai untuk mengevaluasi performa model secara objektif. Pembagian data dilakukan secara acak (random splitting) sehingga setiap entri data memiliki peluang yang sama untuk masuk ke dalam data latih maupun data uji. Pemilihan metode ini bertujuan untuk mengurangi potensi bias dan menjaga representasi distribusi kelas antara data latih dan data uji, sehingga hasil evaluasi kinerja algoritma dapat lebih reliabel. Hasil dari pembagian data terdapat pada Tabel 5.

TABEL 5
HASIL SPLIT DATA

Data	Length
Xtrain	700
Xtest	300



Gambar 6. Confusion matrix Random Forest Setelah Tuning Hyperparameter

Pembagian ini dilakukan agar model dapat mempelajari pola dari mayoritas data sekaligus tetap dapat diuji secara objektif menggunakan data yang belum pernah dilihat sebelumnya.

D. *Hyperparameter Tuning GridSearchCV / RandomizedSearchCV*

Untuk mengatasi potensi overfitting, dilakukan proses tuning hyperparameter menggunakan teknik Grid Search Cross Validation (GridSearchCV). Metode ini memungkinkan pencarian kombinasi parameter terbaik dari berbagai opsi yang tersedia untuk meningkatkan performa model secara optimal. Pada tahap ini, data dibagi menjadi dua bagian, yaitu data training sebesar 70% dan data testing sebesar 30%. Proses tuning dilakukan hanya pada data training, dengan menerapkan 5-fold cross-validation di dalam GridSearchCV. Artinya, data training dibagi menjadi lima subset (fold), di mana model dilatih pada empat subset dan divalidasi pada satu subset secara bergantian hingga lima kali. Pendekatan ini bertujuan untuk memastikan bahwa model yang dihasilkan tidak hanya menghafal pola pada data tertentu (overfitting), tetapi mampu melakukan generalisasi yang baik terhadap data baru. Hasil tuning hyperparameter menunjukkan bahwa kombinasi parameter optimal diperoleh

dengan: n_estimators = 100, min_samples_split = 5, min_samples_leaf = 1, max_depth = 20, dan bootstrap = False. Setelah diperoleh parameter terbaik, dilakukan validasi tambahan menggunakan 5-fold cross-validation pada model hasil tuning untuk memastikan kestabilan performa. Hasil validasi menunjukkan nilai akurasi rata-rata sebesar 97,37%, menandakan model memiliki generalisasi yang baik dan tidak mengalami overfitting.

```
param_grid = {
    'n_estimators': [100, 200, 300, 500],
    'max_depth': [None, 5, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}
```

```
Cross-validation scores: [0.9875  0.95625 0.95625 0.9875  0.98125]
Mean CV accuracy: 0.9737500000000001
```

Gambar 5. Hyperparameter Tunning

1) *Pengujian Algoritma Random Forest*

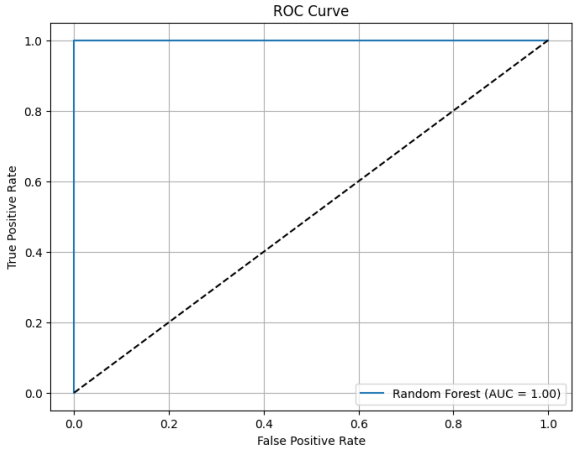
Metode klasifikasi Random Forest tidak memerlukan asumsi tertentu sehingga data hasil dari pre-processing dapat langsung digunakan dalam proses pemodelan. Hasil evaluasi menggunakan confusion matrix pada data testing setelah dilakukan hyperparameter Tuning ditunjukkan pada Gambar 6.

Berdasarkan hasil confusion matrix, maka akan didapatkan nilai akurasi, recall, precision, dan F1-score sebagaimana tertera pada Tabel 6.

TABEL 6
HASIL EVALUASI MODEL RANDOM FOREST SETELAH TUNING HYPERPARAMETER

METRIK EVALUASI	NILAI
Confusion Matrix	[[81, 2], [0,117]]
Accuracy	99%
Recall	100%
Precision	98%
F1-score	99%

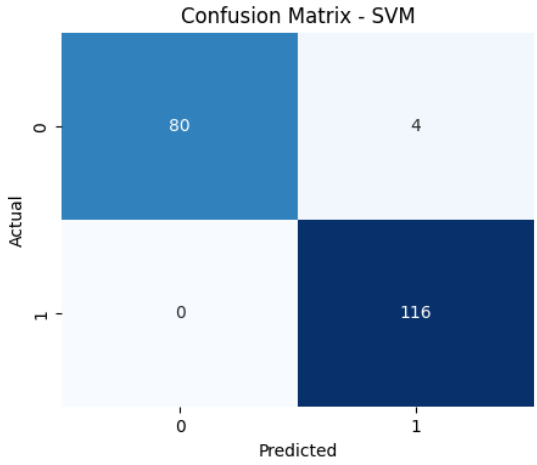
Adapun Gambar 7 menunjukkan kurva ROC dari pemodelan Random Forest setelah dilakukannya tuning hyperparameter.



Gambar 7. Kurva ROC Random Forest Setelah Tuning Hyperparameter

2) *Pengujian Algoritma Support Vector Machine (SVM)*

Metode klasifikasi terakhir yang digunakan adalah Support Vector Machine (SVM). Pada penelitian ini, metode SVM yang digunakan adalah SVM dengan kernel linier. Model ini menunjukkan performa yang sangat baik dengan nilai akurasi sebesar 98%. Nilai akurasi data testing juga mencapai 98%, sedangkan data training memiliki akurasi yang hampir sama sehingga tidak terjadi overfitting pada model. Untuk memperoleh performa yang optimal, dilakukan tuning hyperparameter menggunakan Grid Search CV. Hasil tuning hyperparameter menunjukkan bahwa parameter terbaik diperoleh dengan nilai Cost = 10. Berdasarkan hasil evaluasi, diperoleh nilai recall sebesar 100%, precision sebesar 96,7%, f1-score sebesar 98%, dan ROC-AUC sebesar 99,8%. Hal ini menunjukkan bahwa model mampu mengklasifikasikan data dengan sangat baik, khususnya dalam mendeteksi kelas positif. Confusion matrix data testing untuk Support Vector Machine (SVM) kernel linier setelah dilakukan tuning hyperparameter dapat dilihat pada Gambar 8.



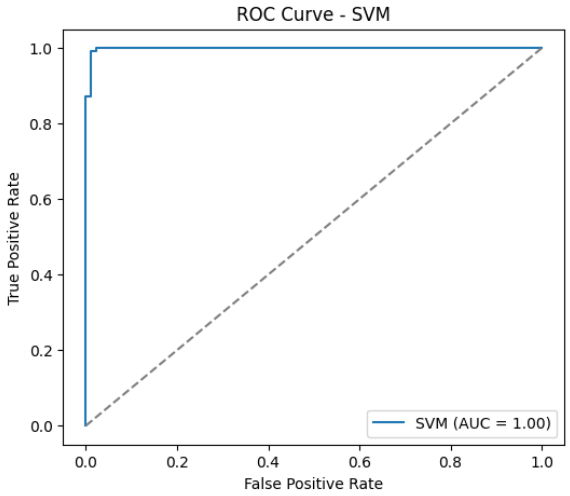
Gambar 8. Confusion matrix SVM Setelah Tuning Hyperparameter

Berdasarkan hasil confusion matrix, maka akan didapatkan nilai akurasi, recall, precision, dan F1-score sebagaimana tertera pada Tabel 7.

TABEL 7
HASIL EVALUASI MODEL SVM SETELAH TUNING HYPERPARAMETER

METRIK EVALUASI	NILAI
Confusion Matrix	[[80, 4], [0,116]]
Accuracy	98%
Recall	100%
Precision	97%
F1-score	98%

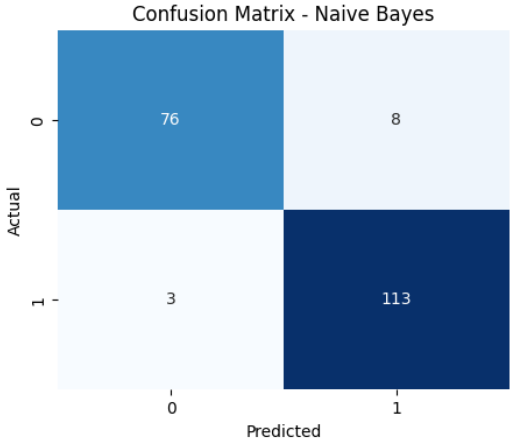
Adapun Gambar 9 menunjukkan kurva ROC dari pemodelan SVM setelah dilakukannya tuning hyperparameter.



Gambar 9. Kurva ROC SVM Setelah Tuning Hyperparameter

3) Pengujian Algoritma Naïve Bayes

Berdasarkan hasil evaluasi, model Naive Bayes memperoleh nilai akurasi sebesar 94,5%. Nilai recall sebesar 97% menunjukkan bahwa model mampu mengenali hampir seluruh data positif dengan baik, sedangkan precision sebesar 93% mengindikasikan masih terdapat beberapa kesalahan prediksi positif (false positive). Nilai f1-score sebesar 95% mencerminkan keseimbangan antara precision dan recall, sehingga secara keseluruhan performa model dapat dikategorikan baik. Hasil confusion matrix yang ditampilkan pada Gambar 10 memperlihatkan bahwa sebanyak 76 data negatif dan 113 data positif berhasil diprediksi dengan benar, sementara hanya terdapat 11 data yang mengalami kesalahan klasifikasi. Hal ini membuktikan bahwa metode Naive Bayes dapat memberikan hasil prediksi yang cukup andal, meskipun kinerjanya sedikit lebih rendah dibandingkan dengan Random Forest dan Support Vector Machine.



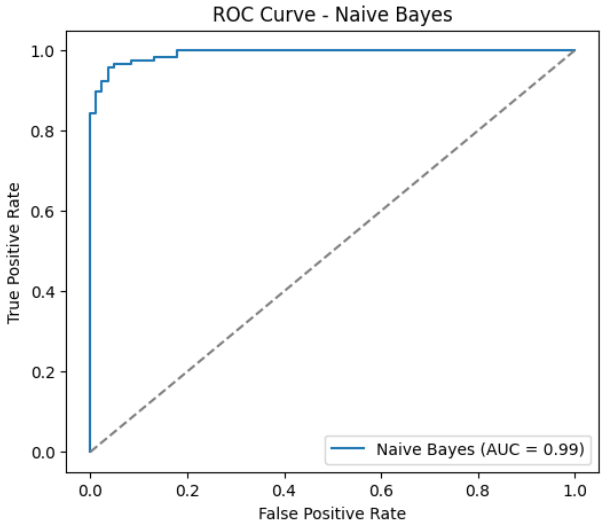
Gambar 10. Confusion matrix Naïve Bayes Setelah Tuning Hyperparameter

Berdasarkan hasil confusion matrix, maka akan didapatkan nilai akurasi, recall, precision, dan F1-score sebagaimana tertera pada Tabel 8.

TABEL 8
HASIL EVALUASI MODEL NAÏVE BAYES SETELAH TUNING HYPERPARAMETER

METRIK EVALUASI	NILAI
Confusion Matrix	[[76, 8], [3,113]]
Accuracy	94.5%
Recall	97%
Precision	93%
F1-score	95%

Adapun Gambar 11 menunjukkan kurva ROC dari pemodelan SVM setelah dilakukannya tuning hyperparameter.



Gambar 11. Kurva ROC Naïve Bayes Setelah Tuning Hyperparameter

E. Perbandingan Antar Metode

Setelah dilakukan klasifikasi dengan menggunakan metode Random Forest, Support Vector Machine (SVM), dan Naïve Bayes didapatkan performa dari setiap model yang digunakan, langkah selanjutnya adalah membandingkan performa yang telah didapatkan. Tabel 9 menunjukkan performa hasil analisis dari setiap metode.

TABEL 9
PERBANDINGAN SELURUH METODE

Metode	Accuracy	Recall	Precision	F1-score
Random Forest	99%	100%	98%	99%
SVM	98%	100%	97%	98%
Naïve Bayes	94.5%	97%	93%	95%

Berdasarkan Tabel 9, metode Random Forest memberikan hasil terbaik dengan akurasi 99% dan recall 100%. Metode SVM juga menunjukkan performa tinggi dengan akurasi 98% dan recall 100%, meskipun precision sedikit lebih rendah dibandingkan Random Forest. Sementara itu, metode Naïve Bayes memiliki akurasi 94,5% dengan recall 97%, sehingga performanya lebih rendah dibandingkan dua metode lainnya namun tetap efisien dari sisi komputasi.

F. Uji Signifikansi (Paired t-test)

Untuk memastikan bahwa perbedaan performa antar model tidak terjadi secara kebetulan, dilakukan uji statistik menggunakan *Paired t-test* berdasarkan hasil akurasi dari *5-fold cross-validation*. Uji ini membandingkan performa tiga model klasifikasi, yaitu Random Forest (RF), Support Vector Machine (SVM), dan Naïve Bayes (NB), dengan hasil sebagai berikut:

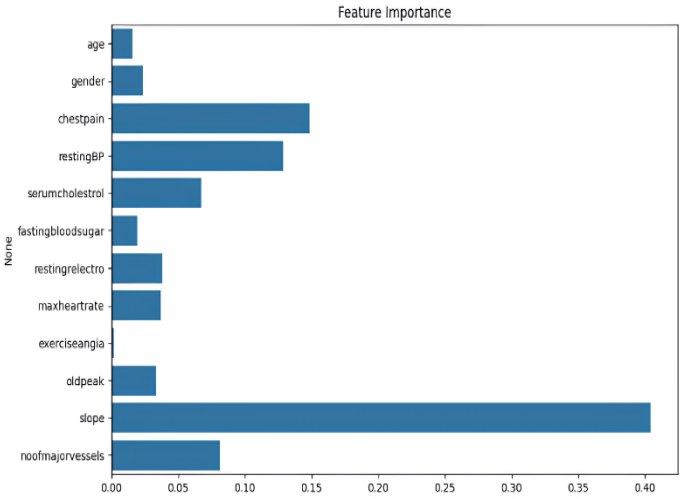
TABEL 10.
PERBANDINGAN KINERJA MODEL BERDASARKAN UJI PAIRED T-TEST

Perbandingan Model	t-value	p-value	Keterangan
Random Forest vs SVM	3.5386	0.0240	Signifikan
Random Forest vs Naïve Bayes	3.3029	0.0299	Signifikan
SVM vs Naïve Bayes	2.2953	0.0834	Tidak signifikan

Berdasarkan hasil uji paired t-test, diperoleh bahwa nilai p-value untuk perbandingan Random Forest vs SVM dan Random Forest vs Naïve Bayes masing-masing lebih kecil dari 0.05. Hal ini menunjukkan bahwa perbedaan performa antara model Random Forest dengan dua model lainnya signifikan secara statistic. Dengan demikian, dapat disimpulkan bahwa Random Forest memberikan performa yang secara konsisten dan signifikan lebih baik dibandingkan SVM maupun Naïve Bayes pada data ini. Konsistensi nilai akurasi yang tinggi di setiap fold juga memperkuat bahwa model ini lebih stabil dan mampu melakukan generalisasi lebih baik terhadap data baru. Sementara itu, perbandingan antara SVM dan Naïve Bayes menunjukkan p-value sebesar 0.0834 (> 0.05), sehingga perbedaannya tidak signifikan. Artinya, performa kedua model tersebut relatif sebanding dan keduanya masih berada di bawah capaian Random Forest.

G. Analisis Feature Importance (khusus Random Forest)

Analisis feature importance dilakukan untuk mengetahui sejauh mana masing-masing variabel berkontribusi terhadap hasil prediksi model Random Forest. Nilai kepentingan fitur dihitung berdasarkan tingkat penurunan impuritas (Gini Importance) yang dihasilkan ketika fitur digunakan untuk memisahkan data dalam setiap node pohon keputusan. Berdasarkan hasil pelatihan model Random Forest dengan evaluasi menggunakan 5-fold cross validation, diperoleh tingkat kepentingan fitur seperti ditunjukkan pada Gambar 12.



Gambar 12. Feature importance

Gambar 12 menunjukkan hasil analisis *feature importance* menggunakan algoritma Random Forest. Berdasarkan grafik tersebut, fitur slope memiliki tingkat kepentingan tertinggi (≈0,40), diikuti oleh chestpain dan restingBP. Hal ini menunjukkan bahwa ketiga fitur tersebut berperan besar dalam menentukan hasil prediksi risiko penyakit kardiovaskular. Sementara itu, fitur seperti exerciseangina, fastingbloodsugar, dan gender memiliki pengaruh yang relatif kecil terhadap performa model.

IV. KESIMPULAN

Penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma machine learning — Random Forest, Support Vector Machine (SVM), dan Naïve Bayes — dalam memprediksi risiko penyakit kardiovaskular menggunakan Mendeley Cardiovascular Disease Dataset. Berdasarkan hasil pengujian, algoritma Random Forest menunjukkan performa terbaik dengan akurasi sebesar 99%, recall 100%, precision 98%, dan F1-score 99%, diikuti oleh SVM dan Naïve Bayes. Uji paired t-test menunjukkan bahwa perbedaan performa

antar model signifikan secara statistik (p-value < 0,05), sehingga dapat disimpulkan bahwa Random Forest merupakan model paling andal dan konsisten.

Dari sisi relevansi klinis, model dengan nilai recall tinggi seperti Random Forest memiliki keunggulan dalam mendeteksi pasien yang benar-benar berisiko (true positive), sehingga dapat mengurangi kemungkinan terjadinya false negative. Kesalahan tipe ini sangat penting untuk dihindari dalam diagnosis penyakit jantung, karena dapat menyebabkan keterlambatan dalam penanganan medis. Sebaliknya, kesalahan tipe false positive memiliki dampak klinis yang lebih ringan, karena pasien masih dapat menjalani pemeriksaan lanjutan. Dengan demikian, model dengan kemampuan deteksi tinggi dan tingkat kesalahan rendah dinilai lebih sesuai untuk implementasi pada sistem pendukung keputusan medis (Clinical Decision Support System) guna membantu proses deteksi dini penyakit kardiovaskular.

Hasil analisis feature importance juga menunjukkan bahwa fitur usia, tekanan darah istirahat, dan kadar kolesterol merupakan faktor paling berpengaruh dalam prediksi risiko penyakit jantung, yang sejalan dengan literatur medis sebelumnya. Temuan ini menegaskan bahwa algoritma machine learning, khususnya Random Forest, dapat berperan penting dalam membantu tenaga medis melakukan identifikasi dini pasien berisiko tinggi, sehingga pencegahan dapat dilakukan lebih cepat dan tepat sasaran.

DAFTAR PUSTAKA

[1] W. H. Organization, "Cardiovascular diseases." [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] B. Ristevski and M. Chen, "Big Data Analytics in Medicine and Healthcare," *J. Integr. Bioinform.*, vol. 15, no. 3, pp. 1–5, 2018, doi: 10.1515/jib-2017-0030.

[3] J. P. Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2012.

[4] M. Wahidin, R. I. Agustiya, and G. Putro, "Beban Penyakit dan Program Pencegahan dan Pengendalian Penyakit Tidak Menular di Indonesia," *J. Epidemiol. Kesehat. Indones.*, vol. 6, no. 2, pp. 105–112, 2023, doi: 10.7454/epidkes.v6i2.6253.

[5] W. H. Organization, *Noncommunicable Diseases Country Profiles 2014*. Geneva, Switzerland: World Health Organization, 2014. [Online]. Available: https://www.who.int/publications/i/item/9789241507509

[6] R. Detrano *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989, doi: 10.1016/0002-9149(89)90524-9.

[7] N. Nasution, M. A. Hasan, and F. Bakri Nasution, "Predicting Heart Disease Using Machine Learning: An Evaluation of Logistic Regression, Random Forest, SVM, and KNN Models on the UCI Heart Disease Dataset," *IT J. Res. Dev.*, vol. 9, no. 2, pp. 140–150, 2025, doi: 10.25299/itjrd.2025.17941.

[8] S. Hadijah Hasanah, "Application of Machine Learning for Heart Disease Classification Using Naive Bayes," *J. Mat. MANTIK*, vol. 8, no. 1, pp. 68–77, 2022, doi: 10.15642/mantik.2022.8.1.68-77.

[9] J. M. Adinulhaq and M. Sam'an, "Perbandingan Kinerja Akurasi Model Mesin Learning Untuk Prediksi Penyakit Jantung," *J. Komput. Dan Teknol. Inf.*, vol. 1, no. 2, pp. 48–55, 2023, doi: 10.26714/jkti.v1i2.12918.

[10] M. Kholish, A. Herdianto, R. F. Setiawan, and R. Samsinar, "Perbandingan Algoritma Random Forest dan Naive Bayes dalam Memprediksi Penyakit Diabetes," *Hubisintek*, vol. 5, no. 1, pp. 322–328, 2024, [Online]. Available: https://ojs.udb.ac.id/index.php/HUBISINTEK/article/view/4757

[11] B. P. Doppala and D. Bhattacharyya, "Cardiovascular Disease Dataset." [Online]. Available: https://data.mendeley.com/datasets/dzz48mvjht/1/files/e4a4a2de-2783-4ea8-9958-0fc3c82cadd4

[12] V. Chernykh, A. Stepnov, and B. O. Lukyanova, "Data preprocessing for machine learning in seismology," *CEUR Workshop Proc.*, vol. 2930, no. October, pp. 119–123, 2021.

[13] J. M. H. Pinheiro *et al.*, "The Impact of Feature Scaling In Machine Learning: Effects on Regression and Classification Tasks," vol. XX, no. X, 2025, [Online]. Available: http://arxiv.org/abs/2506.08274

[14] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd Editio. Sebastopol, CA, USA: O'Reilly Media. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/

[15] L. N. Farida and S. Bahri, "Klasifikasi Gagal Jantung menggunakan Metode SVM (Support Vector Machine)," *Komputika J. Sist. Komput.*, vol. 13, no. 2, pp. 149–156, 2024, doi: 10.34010/komputika.v13i2.11330.

[16] Natasuwarna, "Seleksi Fitur Support Vector Machine pada Analisis Sentimen Keberlanjutan Pembelajaran Daring Support Vector Machine Feature Selection on Online Learning Sustainability Sentiment Analysis," vol. 19, no. 4, pp. 437–448, 2020.

[17] M. B. Anggara, F. T. Informasi, and U. B. Bandung, "Mohammad Bayu Anggara," vol. 20, pp. 32–42, 2025.

[18] W. Wijiyanto, A. I. Pradana, S. Sopingi, and V. Atina, "Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa," *J. Algoritm.*, vol. 21, no. 1, pp. 239–248, 2024, doi: 10.33364/algoritma/v.21-1.1618.