Jurnal Komtika (Komputasi dan Informatika),
Vol. 9 No. 1 | Mei 2025

Evaluasi Kinerja Algoritma *Random Forest* Dan *Gradient Boosting* Untuk Klasifikasi Penyakit Jantung

Ridwan[1*], Hanny Hikmayanti Handayani [2], Santi Arum Puspita Lestari [3], Yana Cahyana[4]
[1,2,3,4]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Buana Perjuangan Karawang

*email: if21.ridwan@mhs.ubpkarawang.ac.id

DOI: https://doi.org/10.31603/komtika.v9i1.13450

Received: 25-04-2025, Revised: 20-05-2025, Accepted: 26-05-2025

ABSTRACT

Cardiovascular disease still accounts for a major portion of deaths worldwide, emphasizing the need for accurate and early diagnosis to reduce associated risks. Advances in machine learning provide opportunities to assist medical professionals in predicting heart disease more efficiently. This study evaluates and compares the performance of two popular supervised learning algorithms, Random Forest and Gradient Boosting, for heart disease classification. Using a dataset comprised of 1000 records highlighting various heart disease risk indicators, the models were developed and assessed. Their performance was determined based on accuracy, precision, recall, and f1-score metrics. The findings indicate that Random Forest consistently outperforms Gradient Boosting across all evaluation metrics. Specifically, Random Forest achieved an accuracy of 99.5%, surpassing Gradient Boosting's 98.5%. Moreover, Random Forest achieved perfect scores (100%) in class 0 precision, class 1 recall, and class 1 F1-score, demonstrating its robustness in handling heart disease classification. The developed model presents strong potential as a decision-support tool in healthcare settings, especially in early screening and patient risk assessment. By identifying key features and patterns associated with heart disease, such models can support healthcare professionals in making quicker and more targeted clinical decisions, ultimately contributing to improved patient outcomes.
Keywords: *heart disease, classification, random forest, gradient boosting.*

ABSTRAK

Penyakit jantung tetap menjadi salah satu faktor utama penyebab kematian di berbagai belahan dunia, sehingga diperlukan diagnosis dini yang akurat untuk mengurangi risiko yang ditimbulkan. Kemajuan teknologi machine learning memberikan peluang baru untuk membantu tenaga medis dalam memprediksi penyakit jantung secara lebih efisien dan tepat. Kajian ini mengarah pada evaluasi dan perbandingan terhadap kinerja dua algoritma pembelajaran terawasi yang populer, yaitu *Random Forest* dan *Gradient Boosting*, dalam klasifikasi penyakit jantung. Dataset berisikan 1.000 baris data dengan sejumlah fitur yang merepresentasikan berbagai faktor risiko penyakit jantung. Penilaian kinerja dilakukan dengan memanfaatkan metrik seperti akurasi, presisi, *recall* dan *f1-score*. Dari hasil analisis diperoleh bahwa *Random Forest* unggul dibandingkan *Gradient Boosting* dalam seluruh metrik evaluasi. *Random Forest* memperoleh akurasi sebesar 99,5%, sementara *Gradient Boosting* memperoleh 98,5%. Selain itu, *Random Forest* mencapai nilai sempurna (100%) pada presisi kelas 0, *recall* kelas 1, dan *F1-score* kelas 1, menunjukkan kemampuannya yang tinggi dalam klasifikasi penyakit jantung. Model yang dikembangkan ini memiliki potensi besar untuk diterapkan sebagai alat bantu pengambilan keputusan dalam sistem layanan kesehatan, terutama pada tahap skrining awal dan penilaian risiko pasien. Dengan mengidentifikasi pola dan fitur kunci yang berhubungan dengan penyakit jantung, model ini dapat membantu tenaga kesehatan dalam memberikan hasil klinis yang lebih cepat juga tepat sasaran.
Keywords: penyakit jantung, klasifikasi, *random forest*, *gradient boosting.*

PENDAHULUAN

Jantung termasuk organ krusial yang menjalankan fungsi penting dalam sistem tubuh manusia. Irama detak jantung menjadi salah satu indikator utama dalam menilai kondisi kesehatan seseorang. Baik detak jantung yang terlalu lambat maupun terlalu cepat dapat berdampak negatif terhadap kesehatan dan menjadi salah satu tanda penyakit jantung [1]. Penyakit jantung adalah kondisi medis yang memengaruhi fungsi jantung, baik secara langsung maupun tidak langsung. Penyakit ini mencakup berbagai gangguan, seperti penyakit gagal jantung, jantung koroner, aritmia, dan penyakit jantung bawaan [2]. Penyakit pada jantung termasuk salah satu penyebab utama yang berkontribusi terhadap tingginya angka kematian di Indonesia. Kondisi ini menjadi perhatian serius karena dapat menyerang siapa saja, terutama untuk mereka yang gaya hidupnya tidak sehat atau memiliki riwayat penyakit tertentu [3]. Kondisi ini dipengaruhi oleh beragam penyebab, salah satu diantaranya adalah pola hidup kurang baik, seperti kebiasaan makan yang kurang baik, jarangnya beraktivitas fisik, serta kebiasaan kebiasaan merokok. Faktor-faktor tersebut secara signifikan meningkatkan risiko gangguan pada jantung dan dapat berujung pada komplikasi yang lebih serius [4].

Dalam mendiagnosis penyakit jantung, terdapat beberapa aspek yang diperhatikan, di antaranya kadar kolesterol, tekanan darah tinggi, kurangnya aktivitas fisik, serta obesitas. Penyebab risiko penyakit pada jantung sendiri terbagi dalam dua kategori, yaitu penyebab yang tak bisa diubah dan penyebab yang bisa dikendalikan. Penyebab yang tak bisa diubah meliputi gender, riwayat penyakit keluarga, serta umur. Kemudian itu, penyebab yang masih bisa diubah meliputi kebiasaan merokok (baik aktif maupun pasif), tekanan darah tinggi, kadar kolesterol yang berlebihan, serta kurangnya aktivitas fisik. Selain itu, proses analisis yang dilakukan oleh dokter menjadi tantangan tersendiri dalam menentukan tingkat risiko pasien terhadap serangan jantung, apakah risikonya tergolong rendah atau tinggi [2].

Pada penelitian samosir dkk tentang penyakit jantung, dengan menggunakan algoritma *K-Nearest Neighbor*, *Naïve Bayes*, dan *Random Forest,* memperlihatkan *Naïve Bayes* memiliki hasil akurasi terbaik yakni 91% [5]. Kemudian, Saputra dkk pernah melakukan klasifikasi juga tetapi klasifikasi nominal mata uang, dengan hasil akurasi 99% dari total 700 data pengujian [6]. Wardhana dkk juga pernah meneliti kacang kering dengan menggunakan algoritma *Gradient Boosting Machine* (GBM), *Random Forest* (RF), *light GBM* dengan model evaluasi *repeated k-folds*. Model Light GBM memperoleh akurasi 99% pada data *training* namun hanya mencapai 91% saat diuji pada data validasi [7]. Setelah itu, Penelitian Yaman dkk membandingkan kinerja algoritma *Random Forest* dengan *Decision Tree* untuk klasifikasi nutrisi pada makanan cepat saji dengan hasil diungguli oleh *Random Forest [8]*.Selanjutnya, Purnomo dkk menggunakan algoritma SVM dan *Random Forest* untuk memprediksi banjir, mendapatkan akurasi tertinggi 99.6% pada algoritma *Random Forest* [9].

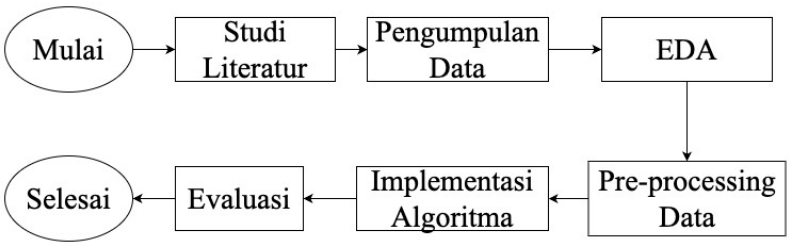
Penelitian-penelitian mengenai klasifikasi yang membedakan antara kelas positif dan negatif telah dilaksanakan oleh peneliti-peneliti terdahulu dengan objek dan metode yang bervariasi. Purnamawati dkk, misalnya, melakukan analisis sentimen terhadap aplikasi TikTok dan berhasil memperoleh akurasi sebesar 83,33%[10]. Sementara itu, Suwitono dan Kaunang menerapkan algoritma *Convolutional Neural Network* (CNN) untuk mengklasifikasikan jenis daun, yang menghasilkan akurasi tinggi sebesar 98%[11]. Penelitian lain oleh Fikriah dkk memfokuskan pada klasifikasi penyakit daun pada tanaman bawang merah dengan

memanfaatkan metode ekstraksi fitur GLMC (*Gray Level Co-Occurrence Matrix)* yang dipadukan dengan algoritma Naïve Bayes, dan mencapai akurasi sebesar 84%[12].

Penerapan algoritma *machine learning* menawarkan solusi yang menjanjikan untuk prediksi penyakit, termasuk penyakit jantung. Algoritma *Random Forest* telah mendapat hasil terbaik dalam penelitian Apriliah et al., dengan akurasi 97.88%, dimana algoritma ini akan menjadi algoritma pertama dalam penelitian [13]. Sedangkan, algoritma kedua yang akan digunakan adalah Gradient Boosting, dimana pada penelitian Andryan & Fajri., mendapatkan hasil 95.12% yang merupakan hasil terbaik [14]. Kedua algoritma ini telah menunjukkan potensi yang tinggi dalam berbagai studi kasus, namun perbandingan kinerjanya dalam konteks prediksi penyakit jantung masih terbatas dan memerlukan penelitian lebih lanjut.

METODE

Metodologi yang diterapkan dalam studi ini adalah metode eksperimental yang berbasis pada Machine Learning. Tahap awal penelitian ini meliputi identifikasi permasalahan, telaah Pustaka, pengumpulan data, pra-pemrosesan dan implementasi klasifikasi dengan *Random Forest* dan *Gradient Boosting*. Langkah selanjutnya adalah melakukan evaluasi terhadap algoritma yang dipilih, sebagaimana terlihat pada Gambar 1.



Gambar 1. Metode Penelitian

1. Studi Literatur

Pada proses ini menjalankan pencarian sumber bahan Pustaka yang terkait dengan objek dan topik penelitian. Proses ini mencakup pencarian jurnal, artikel dan sumber informasi lainnya untuk mendalami topik penelitian. Setelah dilakukan studi literatur dapat diambil kesimpulan bahwa masih sedikitnya studi yang membahas penyakit jantung dengan machine learning dan belum ada studi yang membahas tentang kasus penyakit jantung dengan algoritma *Gradient Boosting* dan *Random Forest*.

2. Pengumpulan Data

Setelah mengidentifikasi permasalahan dan melakukan studi literatur, Langkah selanjutnya adalah mengumpulkan, pencarian dan mempersiapkan dataset. Data yang akan digunakan adalah data dari website Mendeley dengan link https://data.mendeley.com/datasets/dzz48mvjht/1 data tersebut berisi 1000 baris dan 14 fitur kolom, data tersebut diambil pada tanggal 11 November 2024. Untuk rincain dataset bisa dilihat pada Tabel 1.

Tabel 1. Informasi dataset

Feature Name	***Description***
patientid	*Patient Identification Number*
age	*Age In Years*
gender	*gender (1 = male, 0= female)*
chestpain	*Chest pain type (3: asymptomatic, 2: non-anginalpain, , 1: atypical angina 0: typical angina)*
restingBP	*Resting blood pressure 94-200 (in mm HG)*
serumcholestrol	*Serum cholesterol 126-564 (in mg/dl)*
fastingbloodsugar	*Fasting blood sugar 0,1 > 120 mg/dl (1 = true, 0 = false)*
restingrelectro	*Resting electrocardiogram results 0,1,2 (2: Possible/definite LVH, 1: Abnormal ST-T wave, 0: normal)*
maxheartrate	*Maximum heart rate achieved (71-202)*
exerciseangia	*Exercise induced angina (1 = yes, 0 = no)*
oldpeak	*Oldpeak = ST (0-6.2)*
slope	*Slope of the peak exercise ST segment (3-downsloping, 2-flat, 1-upsloping)*
noofmajorvessels	*Number of major vessels (0,1,2,3)*
target	*Classification Heart Disease (0= Absence , 1= Presence)*

3. *Exploratory Data Analysis* (EDA)

Exploratory Data Analysis (EDA) diperkenalkan pertama kali pada tahun 1961 oleh Tukey. EDA yakni proses untuk menganalisis dan memahami data guna menggali wawasan serta mengidentifikasi karakteristik utama dari data tersebut. Secara umum, EDA dikategorikan ke dalam dua metode, yaitu analisis grafis dan analisis non-grafis. Proses ini memiliki peran penting karena membantu dalam memahami pernyataan masalah serta hubungan antara berbagai fitur data sebelum data digunakan dalam pemodelan. EDA sendiri merupakan bagian dari tahapan dalam *data science* [15].

4. *Preprocessing Data*

Tahap seleksi data menjadi langkah penting karena tidak semua data yang ada akan dimanfaatkan dalam penelitian ini. Pemilihan atribut dilakukan berdasarkan relevansi dan kesesuaian untuk perhitungan, sehingga atribut-atribut terpilih lebih signifikan sebagai acuan dalam menilai kelayakan dibandingkan atribut lainnya. Setelah data berhasil diseleksi, proses berikutnya adalah menormalisasi data , di mana data dalam bentuk numerik diubah menjadi bentuk numerik yang setara agar lebih mudah diolah menggunakan algoritma *Random Forest* dan *Gradient Boosting*.

5. *Implementasi Algoritma*

Setelah tahap seleksi data selesai, langkah selanjutnya adalah menerapkan algoritma *Random Forest* dan *Gradient Boosting* untuk melakukan perhitungan, yang dilakukan menggunakan bahasa pemrograman Python. *Random Forest* (RF) merupakan model yang mampu meningkatkan akurasi dengan membentuk simpul anak secara acak disetiap node guna mengoptimalkan kinerjanya. Teknik ini menyusun pohon keputusan yang terdiri dari *node* akar,

node internal dan *node* daun dengan memanfaatkan atribut atau data yang dipilih secara acak berdasarkan aturan tertentu. RF adalah metode machine learning yang berfungsi untuk mengklasifikasikan data, di mana algoritma ini mengelompokkan data berdasarkan kecenderungannya. RF terdiri dari kumpulan decision tree yang bekerja secara kolektif sebagai satu kesatuan fungsional, serta dapat menangani dalam jumlah besar data dengan baik [16]. *Gradient Boosting* merupakan model machine learning yang sangat efektif untuk tugas klasifikasi. Untuk menilai kinerja model, digunakan metrik seperti akurasi dan f-measure. Akurasi mengukur seberapa banyak prediksi yang tepat dari total prediksi yang dilakukan, sementara f-measure menggabungkan precision dan recall. Hal ini memberikan penilaian yang lebih menyeluruh terhadap performa model, khususnya pada data dengan distribusi kelas yang tidak seimbang. Dengan kombinasi ini, *Gradient Boosting* menjadi pilihan yang tepat untuk mengatasi berbagai tantangan dalam analisis data yang kompleks [17]. *Random Forest* dan juga *Gradient Boosting* digunakan karena menunjukan performa paling unggul dibanding model lain pada studi sebelumnya.

6. Evaluasi

Proses evaluasi metode klasifikasi dilakukan untuk mengetahui seberapa baik kinerja masing-masing algoritma dalam melakukan prediksi terhadap data [7]. Dalam penelitian ini, jenis evaluasi yang digunakan meliputi *Confusion Matrix* dan beberapa metrik evaluasi yakni presisi, *F1-score*, *recall* dan akurasi yang dihitung berdasarkan hasil dari *Confusion Matrix* tersebut. *Confusion matrix* berbentuk sebuah tabel yang digunakan untuk menggambarkan hasil klasifikasi dari data uji. Tabel ini menyajikan data mengenai jumlah prediksi yang tepat dan jumlah kesalahan dalam klasifikasi. Dalam konteks pembelajaran mesin, *confusion matrix* memberikan Gambaran yang lebih mendetail mengenai performa model klasifikasi dengan menunjukan sebaran antara prediksi yang tepat dan yang keliru [18].

Tabel 2. Informasi dataset

Kelas	Prediksi Positif	Prediksi Negatif
Aktual Positif	TP	FN
Aktual Negatif	FP	TN

Confusion matrix terdiri dari beberapa elemen utama yang digunakan untuk mengevaluasi kinerja model klasifikasi. *True Positive (TP)* merupakan situasi ketika model berhasil mengidentifikasi data sebagai kelas positif secara tepat . *True Negative (TN)* menggambarkan keberhasilan model dalam prediksi data yang memang masuk dalam kelas negatif. Sementara itu, *False Positive (FP)* muncul saat model keliru mengklasifikasikan data menjadi kelas positif padahal seharusnya negatif. Sebaliknya *False Negative (FN)* terjadi ketika model salah memprediksi data yang seharusnya termasuk dalam kelas positif sebagai kelas negatif. Seluruh komponen tersebut secara kolektif memberikan penilaian lengkap terhadap kemampuan model dalam mengolah data uji [12].

Matriks evaluasi didasarkan pada *confusion matrix*, yang secara luas digunakan untuk menggambarkan kinerja model klasifikasi. Dalam metrik tersebut, baris menggambarkan data actual dari dataset pengujian, sedangkan kolom menampilkan prediksi model. Nilai-nilai yang terdapat dalam *confusion matrix* (matriks kebingungan) dipakai untuk mengukur metrik

(1)

evaluasi seperti *f1-score*, *recall*, presisi dan akurasi, yang masing-masing rumusnya dijelaskan dalam persamaan 1 hingga 4.

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Setelah algoritma *Random Forest* dan *Gradient Boosting* diterapkan, langkah setelahnya yakni mengukur kinerja model yang dihasilkan. Penilaian kinerja model dilakukan dengan menggunakan metriks evaluasi seperti *f1-score*, *recall*, presisi dan akurasi guna mengetahui tingkat efektivitas model algoritma dalam melakukan prediksi terhadap data yang telah diproses. Selain itu, evaluasi juga mencakup analisis terhadap *confusion matrix* untuk mengetahui penyebaran kesalahan prediksi pada masing-masing kelas.

HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini, bertitel "Cardiovascular_Disease_Dataset," merupakan kumpulan data yang dirancang untuk menganalisis faktor-faktor yang berkaitan dengan penyakit kardiovaskular, mencakup berbagai atribut kesehatan pasien. Dataset ini terdiri dari 14 kolom, dengan 13 kolom bertipe data integer yang mewakili variabel kategorikal atau diskrit, seperti status kebiasaan merokok atau tingkat aktivitas fisik, dan 1 kolom bertipe data float yang menggambarkan variabel kontinu, seperti kadar kolesterol. Total dataset mencakup 1000 baris, masing-masing merepresentasikan catatan individu, sebagaimana ditampilkan dalam ikhtisar pada Gambar 3. Gambar 2 menyajikan lima baris pertama dari dataset, memperlihatkan semua kolom untuk memberikan gambaran awal tentang struktur dan isi data. Struktur dataset yang komprehensif ini memungkinkan analisis mendalam terhadap pola dan hubungan antarvariabel yang relevan dengan penyakit kardiovaskular, mendukung tujuan penelitian untuk mengidentifikasi faktor risiko utama. Gambar 2 Menunjukan 5 baris pertama dari dataset lengkap untuk semua kolom yang ada. Dataset yang berjudul "*Cardiovascular_Disease_Dataset*" ini sendiri tersusun dari 13 kolom bertipe data *integer* dan 1 kolom bertipe data *float*, sedangkan untuk barisnya terdiri dari 1000 baris, sebagaimana ditampilkan pada Gambar 3.

	patientid	age	gender	chestpain	restingBP	serumcholestrol	fastingbloodsugar	restingrelectro	maxheartrate	exerciseangia	oldpeak	slope	noofmajorvessels	target
0	103368	53	1	2	171	0	0	1	147	0	5.3	3	3	1
1	119250	40	1	0	94	229	0	1	115	0	3.7	1	1	0
2	119372	49	1	2	133	142	0	0	202	1	5.0	1	0	0
3	132514	43	1	0	138	295	1	1	153	0	3.2	2	2	1
4	146211	31	1	1	199	0	0	2	136	0	5.3	3	2	1

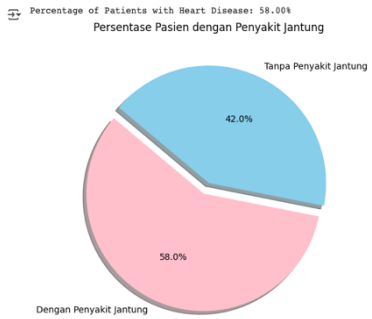
Gambar 2. *Preview* Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   patientid          1000 non-null   int64
 1   age                1000 non-null   int64
 2   gender             1000 non-null   int64
 3   chestpain          1000 non-null   int64
 4   restingBP          1000 non-null   int64
 5   serumcholestrol    1000 non-null   int64
 6   fastingbloodsugar  1000 non-null   int64
 7   restingrelectro    1000 non-null   int64
 8   maxheartrate       1000 non-null   int64
 9   exerciseangia      1000 non-null   int64
 10  oldpeak            1000 non-null   float64
 11  slope              1000 non-null   int64
 12  noofmajorvessels   1000 non-null   int64
 13  target             1000 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 109.5 KB
```

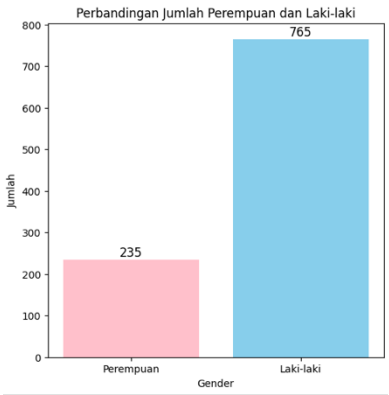
Gambar 3. Informasi tipe data setiap kolom dataset

Exploratory Data Analysis (EDA)

Eksplorasi data dilakukan untuk mengetahui data lebih dalam, yaitu dengan mencari beberapa informasi dari data. Pertama, rentang umur pasien yang ada pada dataset ini adalah dari 20 tahun sampai 80 tahun. Selanjutnya, pada Gambar 4 menampilkan presentase dari pasien dengan penyakit jantung dan tidak, terlihat bahwa data dengan penyakit jantung lebih banyak dengan 58% disbanding data yang tanpa penyakit jantung dengan 42%. Jenis kelamin dari data yang tersedia kemudian dibandingkan untuk mengetahui jumlah dari masing-masing kategori, yaitu perempuan dan laki-laki. Berdasarkan hasil visualisasi, terdapat jumlah data laki-laki yang lebih besar daripada perempuan dengan perbedaan yang cukup mencolok, yaitu sekitar perbandingan 1 berbanding 4. Hal tersebut dapat dilihat pada Gambar 5.



Gambar 4. Persentase Penyakit Jantung



Gambar 5. Perbandingan Jenis Kelamin

Preprocessing

Pada tahap ini, dilakukan serangkaian proses untuk memastikan kualitas dataset sebelum diterapkan pada model. Langkah pertama adalah pemeriksaan *missing value*, di mana data diperiksa untuk mengetahui apakah terdapat nilai yang hilang. Jika ditemukan *missing value*, maka dilakukan penanganan yang sesuai. Namun, dalam dataset kali ini tidak terdapat nilai yang hilang, hasil pemeriksaan *missing value* ditampilkan dalam Gambar 6.

```
patientid           0
age                 0
gender              0
chestpain           0
restingBP           0
serumcholestrol     0
fastingbloodsugar   0
restingrelectro     0
maxheartrate        0
exerciseangia       0
oldpeak             0
slope               0
noofmajorvessels    0
target              0
dtype: int64
```

Gambar 6. Pemeriksaan *Missing Value*

```
Jumlah data duplikat: 0
```

Gambar 7. Pemeriksaan Data Duplikat

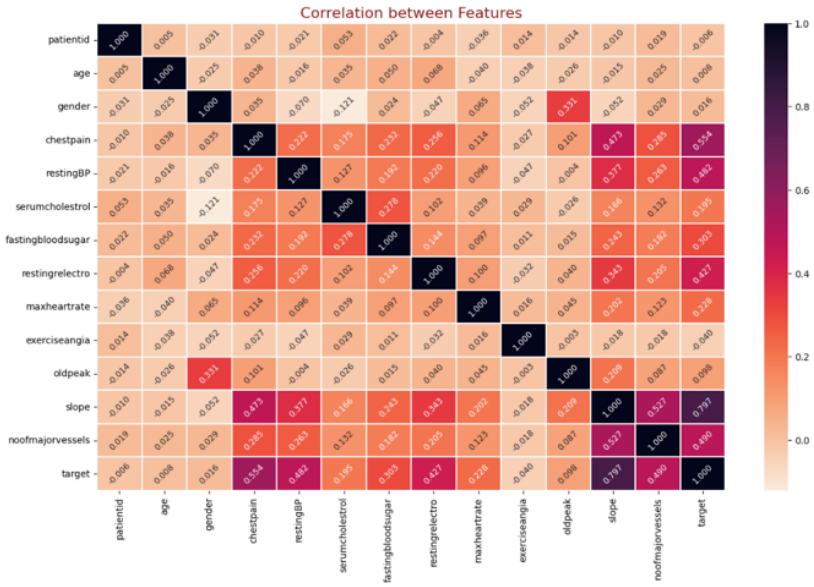
Selanjutnya, dilakukan pemeriksaan data duplikat untuk memastikan tidak ada entri yang berulang. Jika ditemukan duplikasi, maka data yang berulang akan dihapus dan hanya satu entri yang dipertahankan. Dalam kasus ini, dataset tidak mengandung data duplikat, seperti terlihat pada Gambar 7.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   patientid          1000 non-null   int64
 1   age                1000 non-null   int64
 2   gender             1000 non-null   int64
 3   chestpain          1000 non-null   int64
 4   restingBP          1000 non-null   int64
 5   serumcholestrol    1000 non-null   int64
 6   fastingbloodsugar  1000 non-null   int64
 7   restingrelectro    1000 non-null   int64
 8   maxheartrate       1000 non-null   int64
 9   exerciseangia      1000 non-null   int64
 10  oldpeak            1000 non-null   float64
 11  slope              1000 non-null   int64
 12  noofmajorvessels   1000 non-null   int64
 13  target             1000 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 109.5 KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   gender             1000 non-null   int64
 1   chestpain          1000 non-null   int64
 2   restingBP          1000 non-null   int64
 3   serumcholestrol    1000 non-null   int64
 4   fastingbloodsugar  1000 non-null   int64
 5   restingrelectro    1000 non-null   int64
 6   maxheartrate       1000 non-null   int64
 7   oldpeak            1000 non-null   float64
 8   slope              1000 non-null   int64
 9   noofmajorvessels   1000 non-null   int64
dtypes: float64(1), int64(9)
memory usage: 78.2 KB
```

Gambar 8. (a)Dataset sebelum *feature selection* (b) Dataset setelah *feature selection*

Tahap berikutnya adalah *feature selection*, dengan tujuan untuk menyeleksi fitur-fitur yang paling berkaitan dengan target klasifikasi. Dalam proses ini, variabel target (*y*) dipisahkan dari variabel independen (*x*), pemilihan fitur dilakukan dengan memilih 10 kolom yang memiliki korelasi tertinggi terhadap target, kolom yang memiliki korelasi tertinggi terhadap target dapat dilihat pada *Heatmap correlation* pada Gambar 9. Untuk Gambar 8(a) merupakan Kolom dataset sebelum *feature selection* dimana kolomnya masih berjumlah 14 termasuk kolom target, sedangkan Gambar 8(b) memperlihatkan sisa 10 kolom yang akan digunakan dalam Klasifikasi yang sudah dikurangi 3 kolom dalam *feature selection* dan 1 kolom Target



Gambar 9. *Heatmap Correlatioz*

	count	mean	std	min	25%	50%	75%	max
gender	1000.0	0.7650	0.424211	0.0	1.00	1.0	1.00	1.0
chestpain	1000.0	0.9800	0.953157	0.0	0.00	1.0	2.00	3.0
restingBP	1000.0	151.7470	29.965228	94.0	129.00	147.0	181.00	200.0
serumcholestrol	1000.0	311.4470	132.443801	0.0	235.75	318.0	404.25	602.0
fastingbloodsugar	1000.0	0.2960	0.456719	0.0	0.00	0.0	1.00	1.0
restingrelectro	1000.0	0.7480	0.770123	0.0	0.00	1.0	1.00	2.0
maxheartrate	1000.0	145.4770	34.190268	71.0	119.75	146.0	175.00	202.0
oldpeak	1000.0	2.7077	1.720753	0.0	1.30	2.4	4.10	6.2
slope	1000.0	1.5400	1.003697	0.0	1.00	2.0	2.00	3.0
noofmajorvessels	1000.0	1.2220	0.977585	0.0	0.00	1.0	2.00	3.0

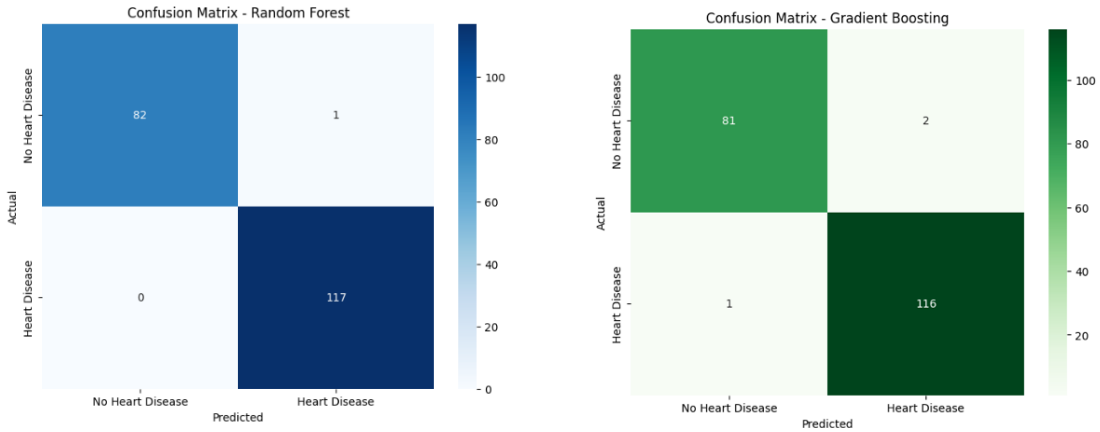
Gambar 10. Hasil Normalisasi data

Setelah itu, dilakukan normalisasi data menggunakan metode *StandardScaler*, proses ini mengatur ulang skala data numerik sehingga memiliki mean nol dan standar devisiasi satu, dengan tujuan menyamakan skali antar fitur agar performa model dapat ditingkatkan. Terakhir, dataset dipecah menjadi dua kelompok untuk tahap *training* dan juf=ga *testing* model. Sebanyak 80% data dipakai sebagai data *training*, sementara 20% dipakai sebagai data *testing*. Dalam penelitian ini, data training terdiri dari 800 baris, sedangkan data *testing* mencakup 200 baris, yang terdiri dari 117 data penderita penyakit jantung dan 83 data yang dinyatakan sehat. Artinya, proporsi data *testing* ini, jumlah penderita penyakit jantung lebih besar dibandingkan yang tidak mengalaminya, yaitu sekitar 58,5% berbanding 41.5%. Pembagian tersebut memiliki tujuan untuk memastikan model dapat belajar dari sebagian banyak data lalu diuji menggunakan data lain yang tidak digunakan dalam proses pelatihan..

Implementasi Algoritma

Algoritma yang diugunakan pada penelitian ini yakni Algoritma *Random Forest* dan *Gradient Boosting*. Kedua algoritma ini digunakan menggunakan *library* dari Scikit-learnn, untuk algoritma *Random Forest* menggunakan *library RandomForestClassifier* dan *Gradient*

Boosting Menggunakan *library GradientBoostingClassifier.* Pada penelitian ini kedua algoritma juga menggunakan *n_estimators = 100* dan *random_state = 42.* Perameter *n_estimators* ditetapkan sebesar 100 karena jumlah tersebut sudah cukup untuk menghasilkan model yang stabil dan performa yang baik, tanpa memerlukan waktu komputasi yang berlebihan. Sedangkan random Sedangkan random_state diatur ke 42 agar hasil model penelitian model dapat direproduksi secara konsisten. Nilai 42 dipilih secara konvensional dan tidak memengaruhi kinerja model secara langsung. Gambar 11 menunjukkan confusion matrix RF dan GB yang ada.



Gambar 11. (a) *Confusion Matrix RF* (b) Confusion Matrix *GB*

Evaluasi

Setelah Implementasi algoritma dilakukan didapatlah hasil evaluasi yaitu berupa *confusion matrix* dan metrik evaluasi. Untuk hasil *confusion matrix* seperti terlihat pada Gambar 11(a) dan 11(b). Pada *Confusion Matrix Random Forest* dari data 83 data 0*(No Heart Disease)* berhasil memprediksi benar sebanyak 82 dan memprediksi salah sebanyak 1. Sedangkan untuk 117 data 1*(Heart Disease)* berhasil memprediksi semuanya dengan benar. Pada *Confusion Matrix Gradient Boosting* dari data 83 data 0*(No Heart Disease)* berhasil memprediksi benar sebanyak 81 dan memprediksi salah 2. Sedangkan untuk 117 data 1*(Heart Disease)* berhasil memprediksi benar sebanyak 116 dan memprediksi salah sebanyak 1.

Sedangkan untuk metrik evaluasi didapatkan hasilnya tetap didominasi oleh keunggulan *Random Forest.* Hasil metrik evaluasi ini menunjukan keunggulan Random Forest pada semua lini, baik presisi, *recall, f1-score* dan juga akurasi. Meskipun perbedaan diantara kedua algoritma hanya sebesar 1%, hasil ini menunjukan bahwa random forest memiliki performa yang lebih unggul dalam menangani dataset yang dipakai. Akurasi random forest mecapai 99.5% sedangkan gradient boosting memperoleh akurasi sebesar 98.5%. perbedaan ini kemungkinan disebabkan oleh kemampuan random forest dalam mengelola variable yang saling berkorelasi dan mengurangi resiko *overfitting.* Meskipun demikian, selisih 1% ini masih relatif kecil dan dapat dipengaruhi oleh pengujian lebih lanjut oleh variasi data atau jumlah sampel yang terbatas. Oleh karena itu, diperlukan pengujian lebih lanjut atau cross-validation

pada penelitian selanjutnya untuk memastikan performa kedua model. Rincian evaluasi masing-masing algoritma disajikan pada Tabel 3.

Tabel 3. Metrik Evaluasi

Model dan Akurasi	Normal(0)			Penyakit Jantung(1)		
	Presisi	Recall	F1-Score	Presisi	Recall	F1-Score
Random Forest(99.5%)	100%	99%	99%	99%	100%	100%
Gradient Boosting(98.5%)	99%	98%	98%	98%	99%	99%

KESIMPULAN

Penelitian ini mengeksplorasi klasifikasi penyakit jantung dengan menggunakan dua algoritma berjenis supervised learning, yaitu *Random Forest* dan juga *Gradient Boosting*. Beberapa temuan penting pada penelitian kali ini yaitu *Random Forest* Menunjukan Hasil yang lebih unggul dibandingkan dengan *Gradient Boosting*. 99.5% adalah hasil akurasi yang diperoleh untuk algoritma Random Forest, hasil ini unggul 1% dibandingkan dengan hasil dari Algoritma Gradient Boosting yaitu 98.5%. Random Forest mengungguli Gradient Boosting dalam semua Metrik evaluasi, terutama pada Presisi kelas 0, *Recall* kelas 1 dan *F1-score* kelas 1 yang mecapai 100%. Ini mengindikasikan bahwa *Random Forest* mempunyai kemampuan yang baik dalam mendeteksi kasus positif penyakit jantung secara tepat. Namun demikian, perbedaan kecil ini masih perlu ditinjau lebih lanjut untuk memastikan kestabilan performa model terhadap dataset yang lebih besar atau bervariasi. Model yang telah dikembangkan berpotensi diterapkan pada alat pendukung dalam rangkaian layanan medis guna memprediksi serta melakukan klasifikasikan penyakit jantung lebih akurat. Model ini diharapkan bisa membantu pada dunia medis dalam proses diagnosa awal dan membuka jalan yang lebih terarah didasarkan pada faktor-faktor risiko yang terdeteksi. Saran pada penelitian selanjut agar dibuatkan sistem semacam perangkat lunak untuk pendeteksian secara langsung baik berupa *website* ataupun aplikasi.

DAFTAR PUSTAKA

[1] J. Dian, F. D. Silalahi, and N. D. Setiawan, "Sistem Monitoring Detak Jantung Untuk Mendeteksi Tingkat Kesehatan Jantung Berbasis Internet Of Things Menggunakan Android," *JUPITER: Jurnal Penelitian Ilmu dan Teknologi Komputer* 13.2 (2021): 69-75.

[2] S. N. N. Arif, A. M. Siregar, S. Faisal, and A. R. Juwita, "Klasifikasi Penyakit Serangan Jantung Menggunakan Metode Machine Learning K-Nearest Neighbors (KNN) dan Support Vector Machine (SVM)," *J. MEDIA Inform. BUDIDARMA*, vol. 8, no. 3, p. 1617, Jul. 2024, doi: https://doi.org/10.30865/mib.v8i3.7844.

[3] A. Saputra Hs, D. Supriyanto, G. Yulisatria, and C. A. Malasari, "Jalan Kaki Sebagai Salah Satu Faktor Dalam Menjaga Kesehatan Jantung," *Griya Cendikia,* vol. 9, no. 2, pp. 221–228, Aug. 2024, doi: https://doi.org/10.47637/griyacendikia.v9i2.1454.

[4] E. I. Scandea, M. A. R. Sugiarto, F. Lestari, and D. Hartanti, "Penerapan Data Mining Untuk Menganalisis Data Faktor Resiko Penyakit Jantung Menggunakan Metode

Logistic Regression," In: *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis. 2023*(SENATIB). p. 683-688.

[5] A. Samosir, W. E. Justino, and T. Hariyono, "Komparasi Algoritma Random Forest, Naïve Bayes dan K- Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung," 2021, Institut Informatika dan Bisnis Darmajaya, 214-222.

[6] A. N. Saputra, H. H. Handayani, C. E. Sukmawati, and A. M. Siregar, "Model Klasifikasi Nominal Mata Uang Kertas Republik Indonesia Menggunakan Convolutional Neural Network," *Journal of Information System Research (JOSH)*, 6(1), 187-195 2024.

[7] I. Wardhana, Musi Ariawijaya, Vandri Ahmad Isnaini, and Rahmi Putri Wirman, "Gradient Boosting Machine, Random Forest dan Light GBM untuk Klasifikasi Kacang Kering," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.,* vol. 6, no. 1, pp. 92–99, Feb. 2022, doi: https://doi.org/10.29207/resti.v6i1.3682.

[8] N. I. Yaman, A. R. Juwita, S. A. P. Lestari, and S. Faisal, "Perbandingan Kinerja Algoritma Decision Tree dan Random Forest," *Jurnal Algoritma Institut Teknologi Garut,* 21(2), 184-196. https://doi.org/10.33364/algoritma/v.21-2.164

[9] I. A. Purnomo, J. Indra, E. E. Awal, and T. Rohana, "Analisis Prediksi Banjir di Indonesia Menggunakan Algoritma Support Vector Machine dan Random Forest," *Journal of Information System Research (JOSH)*, vol. 6, no. 1, 2024.

[10] A. Purnamawati, M. N. Winarto, and M. Mailasari, "Analisis Sentimen Aplikasi TikTok menggunakan Metode BM25 dan Improved K-NN Fitur Chi-Square," *J. Komtika Komputasi Dan Inform.,* vol. 7, no. 1, pp. 97–105, May 2023, doi: https://doi.org/10.31603/komtika.v7i1.8938.

[11] Y. A. Suwitono and F. J. Kaunang, "Implementasi Algoritma Convolutional Neural Network (CNN) Untuk Klasifikasi Daun Dengan Metode Data Mining SEMMA Menggunakan Keras," *J. Komtika Komputasi Dan Inform.,* vol. 6, no. 2, pp. 109–121, Nov. 2022, doi: https://doi.org/10.31603/komtika.v6i2.8054.

[12] F. K. Fikriah, M. Burhanis Sulthan, N. Mujahidah, and Moh. Khoirur Roziqin, "Naïve Bayes untuk Klasifikasi Penyakit Daun Bawang Merah Berdasarkan Ekstraksi Fitur Gray Level Cooccurrence Matrix (GLCM)," *J. Komtika Komputasi Dan Inform.*, vol. 6, no. 2, pp. 133–141, Nov. 2022, doi: https://doi.org/10.31603/komtika.v6i2.7925.

[13] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi," *J. Sist. Inf.*, vol. 10. doi: https://doi.org/10.32520/stmsi.v10i1.1129

[14] M. R. Andryan and M. Fajri, "Komparasi Kinerja Algoritma Xgboost Dan Algoritma Support Vector Machine (Svm) Untuk Diagnosa Penyakit Kanker Payudara," [1] *JIKO (Jurnal Informatika dan Komputer).* vol. 6, no. 1, 2022. DOI: http://dx.doi.org/10.26798/jiko.v6i1.500

[15] M. Z. Siambaton and A. M. Husein, "Menganalisis Data Kesehatan Global : Pendekatan Analisis Data Eksplorasi Visual". ". *dsi*, vol 1 , no 2, 2021. DOI: https://doi.org/10.47709/dsi.v1i2.1315

[16] A. Hidayanti, A. M. Siregar, S. A. P. Lestari, and Y. C. Cahyana, "Model Analisis Kasus Covid-19 Di Indonesia Menggunakan Algoritma Regresi Linier Dan Random Forest," *PETIR*, vol. 15, no. 1, pp. 91–101, Dec. 2021, doi: https://doi.org/10.33322/petir.v15i1.1487.

[17] E. Wonda, M. S. Wambrauw, R. Ferrari, R. G. Napitupulu, and R. H. D. Febrianti, "Literature Review: Penerapan Gradient Boosting Untuk Klasifikasi Penyakit Diabetes Tipe 2," *OKTAL : Jurnal Ilmu Komputer dan Science,* vol. 3, no. 10, 2024.

[18] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," ,". *Jurnal Sains Komputer & Informatika (J-SAKTI)*., vol. 5, 2021. DOI: http://dx.doi.org/10.30645/j-sakti.v5i2.369