Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833

Komparasi Algoritma Extreme Gradient Boosting, Support Vector Machine, dan Random Forest Klasifikasi Penyakit Jantung

Fadil Arrosyid Fayyadh[1], Damayanti[2,*]

[1] Fakultas Teknik dan Ilmu Komputer, Prodi Sistem Informasi, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
[2] Fakultas Teknik dan Ilmu Komputer, Prodi Ilmu komputer, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
Email: [1]fadil_arrosyid_fayyadh_@teknokrat.ac.id, [2*]damayanti@teknokrat.ac.id
Email Penulis Korespondensi: damayanti@teknokrat.ac.id
Submitted: **04/05/2026**; Accepted: **30/06/2026**; Published**: 30/06/2026**

Abstrak−Penyakit jantung merupakan salah satu penyebab utama kematian di dunia yang memerlukan deteksi dini untuk mengurangi risiko fatalitas. Penelitian ini bertujuan menganalisis dan membandingkan kinerja tiga algoritma machine learning, yaitu Support Vector Machine (SVM), Random Forest, dan XGBoost dalam klasifikasi penyakit jantung. Dataset diperoleh dari platform Kaggle dengan variabel seperti BMI, kebiasaan merokok, aktivitas fisik, kondisi kesehatan umum, dan atribut kesehatan lainnya, serta HeartDisease sebagai variabel target. Tahapan penelitian meliputi preprocessing data, encoding data kategorikal, normalisasi data, dan pembagian data menggunakan metode train-test split 80:20. Hasil penelitian menunjukkan bahwa SVM dan XGBoost memperoleh akurasi sebesar 0.90, sedangkan Random Forest sebesar 0.89. Berdasarkan metrik evaluasi lainnya, XGBoost menunjukkan performa terbaik dengan precision 0.88, recall 0.90, dan F1-score 0.88. Analisis feature importance juga menunjukkan bahwa beberapa faktor kesehatan memiliki pengaruh signifikan terhadap risiko penyakit jantung. Kontribusi penelitian ini terletak pada perbandingan performa algoritma machine learning untuk klasifikasi penyakit jantung serta identifikasi faktor kesehatan yang berpengaruh terhadap risiko penyakit jantung. Hasil penelitian diharapkan dapat menjadi referensi dalam pengembangan sistem pendukung keputusan untuk membantu deteksi dini penyakit jantung secara lebih akurat dan efisien.

Kata Kunci: Penyakit Jantung; Random Forest; Support Vector Machine; XGBoost

Abstract−Heart disease is one of the leading causes of death worldwide and requires early detection to reduce the risk of fatality. This study aims to analyze and compare the performance of three machine learning algorithms, namely Support Vector Machine (SVM), Random Forest, and XGBoost, in heart disease classification. The dataset was obtained from the Kaggle platform and consists of variables such as BMI, smoking habits, physical activity, general health conditions, and other health-related attributes, with HeartDisease as the target variable. The research stages include data preprocessing, categorical data encoding, data normalization, and data splitting using the 80:20 train-test split method. The results show that SVM and XGBoost achieved an accuracy of 0.90, while Random Forest achieved 0.89. Based on other evaluation metrics, XGBoost demonstrated the best performance with a precision of 0.88, recall of 0.90, and F1-score of 0.88. Feature importance analysis also revealed that several health factors significantly influence the risk of heart disease. The contribution of this study lies in comparing the performance of machine learning algorithms for heart disease classification and identifying influential health factors related to heart disease risk. The findings are expected to serve as a reference for developing decision support systems to assist early heart disease detection more accurately and efficiently.

Keywords: Heart Disease; Random Forest; Support Vector Machine; XGBoost

1. PENDAHULUAN

Jantung merupakan organ berotot berbentuk rongga yang berfungsi memompa darah ke seluruh tubuh melalui pembuluh darah dengan mekanisme kontraksi yang berlangsung secara ritmis dan berkesinambungan. Organ ini memiliki peran utama dalam sistem peredaran darah, yaitu mendistribusikan oksigen dan nutrisi ke seluruh jaringan tubuh serta membantu proses pembuangan sisa metabolisme. Secara anatomis, jantung terletak di dalam rongga dada dengan posisi cenderung ke sebelah kiri [1]. Mengingat fungsinya yang sangat vital, jantung termasuk organ esensial dalam tubuh manusia, sehingga setiap gangguan atau kelainan yang terjadi dapat meningkatkan risiko morbiditas hingga mortalitas. Gangguan pada jantung umumnya diklasifikasikan menjadi dua kategori, yaitu penyakit jantung dan serangan jantung. Berdasarkan data dari Organisasi Kesehatan Dunia (WHO), sekitar 7,3 juta kematian di dunia disebabkan oleh penyakit jantung. Kondisi tersebut umumnya terjadi akibat ketidakmampuan jantung dalam menjalankan fungsinya secara optimal, seperti adanya kelemahan pada otot jantung atau kelainan struktural, misalnya defek pada sekat antara serambi kanan dan serambi kiri [2].

Penyakit jantung dapat dipahami sebagai istilah umum yang mencakup berbagai jenis gangguan yang berasal dari organ jantung itu sendiri. Dalam lingkup yang lebih luas, penyakit kardiovaskular (CVD) menjadi salah satu penyebab kematian tertinggi di dunia jika dibandingkan dengan penyakit lainnya. Sebagai penyakit tidak menular, kondisi ini menyumbang angka kematian paling besar secara global. Berdasarkan laporan dari World Health Organization (WHO) tahun 2000, salah satu permasalahan utama adalah banyaknya individu yang tidak mampu mengenali gejala awal penyakit ini, sehingga penanganan atau pengobatan sering kali terlambat dilakukan [3].

Faktor risiko penyakit jantung tidak terlepas dari pola hidup yang kurang sehat, seperti kebiasaan mengonsumsi makanan tinggi karbohidrat maupun lemak. Selain itu, kondisi obesitas, kurangnya aktivitas fisik atau jarang berolahraga, serta kebiasaan merokok turut berkontribusi dalam meningkatkan risiko terjadinya penyakit jantung [4]. Masih banyak masyarakat yang memiliki keterbatasan pengetahuan terkait kondisi kesehatan jantung yang dimilikinya, sehingga sering kali tidak menyadari bahwa dirinya sedang mengalami gangguan pada jantung. Kondisi

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833

ini umumnya dipengaruhi oleh rendahnya tingkat kesadaran terhadap pentingnya melakukan pemeriksaan medis secara berkala, khususnya yang berkaitan dengan kesehatan jantung. Padahal, penyakit jantung dapat terjadi pada siapa saja tanpa terkecuali, termasuk pada individu yang sebelumnya tidak memiliki riwayat penyakit jantung atau tidak menunjukkan gejala yang jelas [5].

Penelitian ini dilakukan untuk membandingkan kinerja tiga algoritma klasifikasi, yaitu XGBoost, Support Vector Machine (SVM), dan Random Forest dalam mengklasifikasikan risiko penyakit jantung. Data yang digunakan dalam penelitian ini berjumlah 31.979 entri pasien yang diperoleh dari platform Kaggle. Pengujian dilakukan dengan menggunakan teknik random sampling yang dikombinasikan dengan skema pelatihan dan pengujian sebanyak sepuluh kali, dengan proporsi 80% data digunakan sebagai data latih pada setiap iterasi. Penelitian ini diharapkan dapat menghasilkan informasi terkait algoritma klasifikasi yang paling optimal dalam deteksi dini penyakit jantung, berdasarkan indikator evaluasi berupa akurasi, presisi, recall, dan F1-score, serta memberikan rekomendasi metode yang efektif untuk diimplementasikan pada sistem pendukung keputusan medis [1].

Klasifikasi penyakit jantung telah banyak menjadi objek penelitian dengan memanfaatkan berbagai metode dalam data mining. Berdasarkan penelitian terdahulu, Random Forest terbukti memiliki performa terbaik dengan akurasi 98,61% dan AUC-ROC 98,96% pada data klinis, sedangkan XGBoost juga menunjukkan kinerja yang kompetitif [6]. Pada penelitian sebelumnya, juga menunjukkan bahwa kedua algoritma memiliki kinerja yang baik, dengan rata-rata nilai Precision, Recall, dan F1-Score di atas 79%, serta tingkat akurasi yang melebihi 86%. Selain itu, algoritma Random Forest menunjukkan performa paling optimal dengan rata-rata nilai di atas 87% dan akurasi tertinggi mencapai 98,17% [7]. Temuan serupa juga diperoleh dalam penelitian tersebut, di mana metode klasifikasi Decision Tree, Random Forest, dan XGBoost masing-masing menghasilkan akurasi sebesar 87,5%, 91,7%, dan 93,1% pada data latih tanpa penerapan SMOTE, serta meningkat menjadi 88,9%, 93,1%, dan 97,2% setelah dilakukan penanganan data[8]. Penelitian tersebut mengungkapkan bahwa metode Stacking menghasilkan kinerja klasifikasi sempurna dengan tingkat akurasi 100% pada data pengujian, yang terverifikasi melalui confusion matrix tanpa kesalahan klasifikasi. Hasil ini sebanding dengan Random Forest serta melampaui performa XGBoost yang hanya mencapai 90% [9].

Berdasarkan penelitian terdahulu, masih terdapat perbedaan hasil dalam menentukan algoritma terbaik untuk klasifikasi penyakit jantung, di mana Random Forest, XGBoost, dan metode lainnya menunjukkan performa yang bervariasi. Selain itu, belum banyak penelitian yang membandingkan XGBoost, Support Vector Machine (SVM), dan Random Forest secara langsung dalam satu eksperimen dengan metode pengujian yang konsisten. Perbedaan dataset dan teknik pengolahan data juga membuat hasil penelitian sulit dibandingkan, serta sebagian penelitian masih berfokus pada performa model tanpa membahas penerapannya dalam sistem pendukung keputusan medis [10].

Tujuan dari penelitian ini adalah untuk menganalisis dan membandingkan kinerja tiga algoritma klasifikasi, yaitu XGBoost, Support Vector Machine (SVM), dan Random Forest dalam mengklasifikasikan risiko penyakit jantung. Penelitian ini bertujuan untuk mengidentifikasi algoritma yang paling optimal dalam mendeteksi penyakit jantung secara dini berdasarkan indikator evaluasi berupa akurasi, presisi, recall, dan F1-score. Selain itu, penelitian ini juga bertujuan untuk memberikan rekomendasi metode klasifikasi yang efektif dan akurat guna mendukung pengembangan sistem pendukung keputusan di bidang medis, khususnya dalam diagnosis penyakit jantung [11].

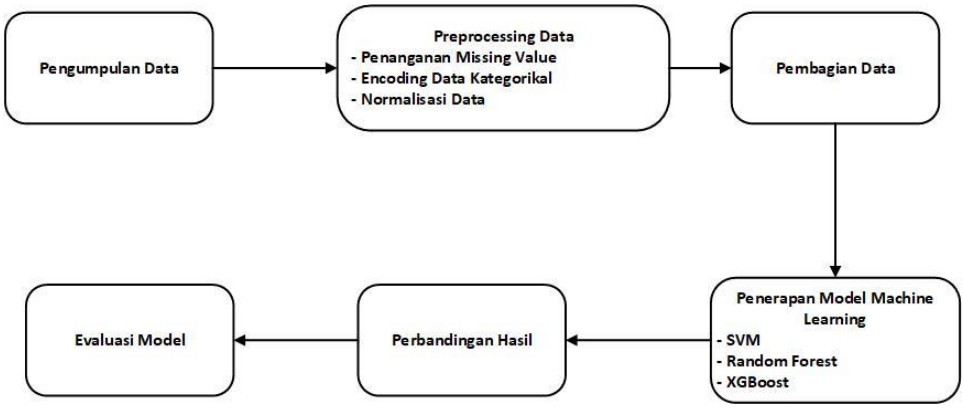
Kontribusi penelitian ini terletak pada perbandingan langsung tiga algoritma klasifikasi, yaitu XGBoost, Support Vector Machine (SVM), dan Random Forest menggunakan dataset dan metode pengujian yang konsisten, sehingga hasil evaluasi dapat dibandingkan secara lebih objektif. Penelitian ini juga memberikan analisis terhadap performa masing-masing algoritma berdasarkan metrik akurasi, presisi, recall, dan F1-score untuk menentukan model terbaik dalam klasifikasi penyakit jantung. Selain itu, hasil penelitian diharapkan dapat menjadi referensi dalam pengembangan sistem pendukung keputusan medis berbasis machine learning guna membantu proses deteksi dini penyakit jantung secara lebih cepat dan akurat.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilakukan dengan menggunakan pendekatan kuantitatif melalui implementasi dan evaluasi tiga algoritma machine learning, yaitu Support Vector Machine (SVM), Random Forest, dan XGBoost, dalam klasifikasi penyakit jantung berdasarkan data medis[6]. Tahapan penelitian diawali dengan pengumpulan dataset penyakit jantung yang diperoleh dari platform Kaggle sebagai sumber data terbuka. Selanjutnya dilakukan tahap preprocessing data yang meliputi penanganan missing value, encoding variabel kategorikal, dan normalisasi data untuk meningkatkan kualitas data serta performa model klasifikasi. Setelah proses preprocessing selesai, dataset dibagi menjadi data latih dan data uji menggunakan teknik random sampling dengan proporsi tertentu. Data latih digunakan untuk proses pelatihan model, sedangkan data uji digunakan untuk evaluasi performa model. Tahap evaluasi dilakukan menggunakan indikator accuracy, precision, recall, dan F1-score untuk menentukan algoritma dengan performa terbaik dalam klasifikasi penyakit jantung [12]. Adapun tahapan penelitian yang dilakukan pada penelitian ini dapat dilihat pada Gambar 1 yang menunjukkan alur proses penelitian mulai dari pengumpulan data, preprocessing data, pembagian data, penerapan algoritma machine learning, hingga tahap evaluasi model.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833



Gambar 1. Tahapan Metode penelitian

2.2 Pengumpulan Data

Tahap pengumpulan data dalam penelitian ini dilakukan dengan memanfaatkan data sekunder yang diperoleh dari platform *Kaggle*. Dataset yang digunakan merupakan data penyakit jantung yang tersedia secara publik dan dapat diakses oleh peneliti untuk keperluan analisis. Proses pengambilan data dilakukan dengan cara mengunduh dataset secara langsung dari Kaggle, kemudian data tersebut diseleksi dan disesuaikan dengan kebutuhan penelitian. Dataset ini memuat berbagai atribut yang relevan, seperti variabel target *Heart Disease* serta fitur pendukung lainnya yang digunakan dalam proses klasifikasi[13].

2.3 Preprosessing Data

Pada penelitian ini, tahap preprocessing data dilakukan untuk mempersiapkan dataset penyakit jantung yang diperoleh dari Kaggle agar siap digunakan dalam proses pemodelan menggunakan algoritma Support Vector Machine (SVM), Random Forest, dan XGBoost[14]. Tahapan preprocessing yang dilakukan meliputi:

2.3.1 Penanganan Missing Value

Dataset diperiksa untuk mengidentifikasi adanya nilai yang hilang. Jika ditemukan data yang tidak lengkap, maka dilakukan penanganan dengan metode tertentu, seperti pengisian nilai menggunakan rata-rata (mean) untuk data numerik atau modus untuk data kategorikal, sehingga tidak mengganggu proses analisis[15].

2.3.2 Encoding Data Kategorikal

Beberapa variabel dalam dataset seperti *Sex*, *Age Category*, *Race*, dan *Gen Health* merupakan data kategorikal. Oleh karena itu, dilakukan proses encoding untuk mengubah data tersebut menjadi bentuk numerik menggunakan teknik seperti *label encoding* atau *one-hot encoding*, agar dapat diproses oleh algoritma machine learning[16].

2.3.3 Normalisasi Data

Fitur numerik seperti *BMI*, *Physical Health*, *Mental Health*, dan *SleepTime* dinormalisasi untuk menyamakan skala data. Proses ini bertujuan agar tidak ada fitur yang mendominasi model, terutama pada algoritma seperti SVM yang sensitif terhadap skala data[15].

2.4 Pembagian Data

Tahap pembagian data pada penelitian ini dilakukan setelah proses preprocessing selesai, dengan tujuan untuk memisahkan dataset menjadi data latih (*training data*) dan data uji (*testing data*). Pembagian data dilakukan menggunakan teknik *random sampling* agar distribusi data tetap representatif dan tidak menimbulkan bias. Proporsi yang digunakan adalah 80:20, di mana 80% data digunakan untuk proses pelatihan model dan 20% sisanya digunakan untuk pengujian. Data latih dimanfaatkan untuk membangun model menggunakan algoritma Support Vector Machine (SVM), Random Forest, dan XGBoost, sedangkan data uji digunakan untuk mengevaluasi kinerja model dalam mengklasifikasikan data baru, sehingga dapat diketahui tingkat akurasi dan kemampuan generalisasi dari model yang dihasilkan[17].

2.5 Model *Support Vector Machine (SVM)*

Pada penelitian ini, algoritma Support Vector Machine (SVM) digunakan untuk melakukan klasifikasi penyakit jantung dengan mencari *hyperplane* terbaik yang mampu memisahkan dua kelas secara optimal. SVM bekerja dengan memaksimalkan margin atau jarak antar kelas sehingga menghasilkan batas keputusan yang paling akurat[18].

$$min_{w,b} \frac{1}{2} ||w||2 \ \ dengan\ syrat\ y_i(w.x_i + b) \geq 1 \quad (1)$$

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833

Rumus tersebut menunjukkan bahwa SVM berusaha meminimalkan nilai $frac12 \mid\mid w \mid\mid^2$, yaitu panjang vektor bobot (*weight vector*) yang menentukan kemiringan *hyperplane*. Syarat $y_i(wcdotx_i + b)geq1$ memastikan bahwa setiap data $x_i$diklasifikasikan dengan benar sesuai labelnya y_i. Dengan kata lain, model tidak hanya mencari pemisah, tetapi juga memastikan jarak antar kelas maksimum sehingga meningkatkan kemampuan generalisasi.

2.6 Model *Random Forest*

Random Forest merupakan algoritma berbasis *ensemble learning* yang membangun banyak pohon keputusan (*decision tree*) dan menggabungkan hasil prediksi untuk meningkatkan akurasi. Pada penelitian ini, Random Forest digunakan untuk mengurangi overfitting dan meningkatkan stabilitas model dalam klasifikasi penyakit jantung[19].

$$\hat{y} = \text{mode}\big(h_1(x), h_2(x), \dots, h_n(x)\big) \tag{2}$$

Rumus tersebut menjelaskan bahwa hasil prediksi akhir $haty$diperoleh dari nilai mayoritas (*mode*) dari seluruh prediksi pohon keputusan $h_1(x), h_2(x), \dots, h_n(x)$. Setiap pohon memberikan hasil prediksi sendiri berdasarkan subset data yang berbeda, kemudian hasil akhir ditentukan berdasarkan suara terbanyak. Pendekatan ini membuat model lebih robust dan mengurangi kesalahan prediksi dari masing-masing pohon tunggal.

2.7 Model *XGBoost (Extreme Gradient Boosting)*

XGBoost (Extreme Gradient Boosting) merupakan algoritma boosting yang membangun model secara bertahap, di mana setiap model baru memperbaiki kesalahan dari model sebelumnya. Pada penelitian ini, XGBoost digunakan karena kemampuannya dalam menghasilkan performa tinggi dan efisiensi komputasi yang baik[19].

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{3}$$

Rumus tersebut menunjukkan bahwa prediksi akhir $haty_i$merupakan hasil penjumlahan dari sejumlah fungsi $f_k(x_i)$yang berasal dari setiap model (biasanya berupa pohon keputusan). Setiap fungsi $f_k$berperan sebagai model lemah (*weak learner*) yang secara bertahap memperbaiki kesalahan prediksi sebelumnya. Dengan menjumlahkan kontribusi dari setiap model, XGBoost mampu menghasilkan prediksi yang lebih akurat dan meminimalkan error secara keseluruhan.

2.8 Evaluasi Model

Tahap evaluasi model pada penelitian ini dilakukan untuk mengukur kinerja algoritma Support Vector Machine (SVM), Random Forest, dan XGBoost dalam mengklasifikasikan penyakit jantung. Evaluasi dilakukan menggunakan data uji (*testing data*) yang tidak digunakan pada saat proses pelatihan model, sehingga dapat diketahui kemampuan model dalam melakukan prediksi terhadap data baru. Metode evaluasi yang digunakan adalah *confusion matrix*, yang menghasilkan beberapa metrik penilaian seperti akurasi (*accuracy*), presisi (*precision*), *recall*, dan *F1-score*[20].

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data Penelitian

Penelitian ini memanfaatkan dataset penyakit jantung yang diperoleh dari platform Kaggle melalui tautan https://www.kaggle.com/code/alkidiarete/heart-disease-roc-0-98/input. Dataset yang digunakan merupakan data sekunder yang bersifat publik dan telah terstruktur, sehingga dapat langsung dimanfaatkan dalam proses analisis. Data tersebut memuat sejumlah variabel penting, seperti *Heart Disease* sebagai variabel target, serta berbagai fitur pendukung seperti *BMI, Smoking, Alcohol Drinking, Stroke, Physical Health, Mental Health, Diff Walking, Sex, Age Category, Race, Diabetic, Physical Activity, Gen Health, SleepTime, Asthma, Kidney Disease*, dan *Skin Cancer*. Proses pengumpulan data dilakukan dengan cara mengunduh dataset secara langsung dari Kaggle, kemudian data diseleksi dan disesuaikan dengan kebutuhan penelitian. Dataset ini mencerminkan kondisi kesehatan individu berdasarkan berbagai faktor risiko yang relevan, sehingga dapat digunakan sebagai dasar dalam proses analisis klasifikasi. Dengan ketersediaan data yang lengkap dan terstruktur, penelitian ini dapat mengidentifikasi pola keterkaitan antar variabel serta mendukung tahapan selanjutnya, seperti preprocessing data, eksplorasi data, hingga penerapan algoritma machine learning untuk klasifikasi penyakit jantung. Selain itu, penelitian ini juga menerapkan metode perbandingan performa algoritma menggunakan nilai evaluasi accuracy, precision, recall, dan F1-score untuk membandingkan hasil analisis dari algoritma Support Vector Machine (SVM), Random Forest, dan XGBoost sehingga dapat diketahui algoritma dengan performa terbaik dalam klasifikasi penyakit jantung. Adapun data penyakit jantung yang digunakan dalam penelitian ini ditampilkan pada Tabel 1 untuk memberikan gambaran mengenai atribut dan variabel yang digunakan dalam proses klasifikasi penyakit jantung.

Tabel 1. Data *Heart Disease*

HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	Physical Health	MentalHealth	DiffWalking	Sex	AgeCategory	Diabetic	GenHealth	Asthma	KidneyDisease	SkinCancer
No	16.6	Yes	No	No	3.0	30.0	No	Female	55-59	Yes	Very good	Yes	No	Yes

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833

HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Diabetic	GenHealth	Asthma	KidneyDisease	SkinCancer
No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	No	Very good	No	No	No
No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	Yes	Fair	Yes	No	No
No	24.21	No	No	No	0.0	0.0	No	Female	75-79	No	Good	No	No	Yes
No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	No	Very good	No	No	No
Yes	28.87	Yes	No	No	6.0	0.0	Yes	Female	75-79	No	Fair	No	No	No
No	21.63	No	No	No	15.0	0.0	No	Female	70-74	No	Fair	Yes	No	Yes

Berdasarkan Tabel 1, data *Heart Disease* yang digunakan dalam penelitian ini terdiri dari berbagai atribut yang berkaitan dengan kondisi kesehatan dan gaya hidup pasien, seperti BMI, riwayat merokok, konsumsi alkohol, kesehatan fisik dan mental, tingkat kesulitan berjalan, jenis kelamin, kelompok usia, status diabetes, aktivitas fisik, asma, penyakit ginjal, dan riwayat kanker kulit. Variabel-variabel tersebut digunakan sebagai fitur dalam proses klasifikasi untuk mengidentifikasi potensi penyakit jantung. Keberagaman atribut pada dataset ini memberikan informasi yang cukup lengkap sehingga dapat mendukung proses analisis serta meningkatkan performa algoritma machine learning dalam melakukan prediksi penyakit jantung.

3.2 *Preprosessing Data*

Pada tahap ini dilakukan proses seleksi fitur untuk menentukan variabel yang digunakan dalam penelitian. Atribut yang dipilih terdiri dari *HeartDisease* sebagai variabel dependen (target) serta variabel independen yang meliputi *BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, DiffWalking, Sex, AgeCategory, Diabetic, GenHealth, KidneyDisease,* dan *Asthma*. Seleksi fitur dilakukan dengan tujuan untuk menyederhanakan dataset serta menghilangkan atribut yang tidak relevan, sehingga dapat meningkatkan efisiensi proses komputasi dan kinerja model dalam melakukan klasifikasi. Hasil seleksi fitur dapat dilihat pada Gambar 2.

```
### Seleksi Fitur ###
Total Baris & Kolom: (3116, 13)
  HeartDisease    BMI Smoking AlcoholDrinking Stroke  PhysicalHealth  \
0           No  16.60     Yes              No     No             3.0
1           No  20.34      No              No    Yes             0.0
2           No  26.58     Yes              No     No            20.0
3           No  24.21      No              No     No             0.0
4           No  23.71      No              No     No            28.0

  DiffWalking     Sex  AgeCategory Diabetic  GenHealth KidneyDisease Asthma
0          No  Female        55-59      Yes  Very good            No    Yes
1          No  Female  80 or older       No  Very good            No     No
2          No    Male        65-69      Yes       Fair            No    Yes
3          No  Female        75-79       No       Good            No     No
4         Yes  Female        40-44       No  Very good            No     No
```

Gambar 2. Hasil Seleksi Fitur

Berdasarkan Gambar 2, dataset yang digunakan setelah proses seleksi fitur memiliki total 3116 baris data dan 13 kolom atribut. Gambar tersebut juga menampilkan beberapa sampel data hasil seleksi fitur yang terdiri dari variabel target dan variabel independen yang digunakan pada penelitian. Dengan dilakukannya seleksi fitur, data yang digunakan menjadi lebih terstruktur dan relevan sehingga dapat mendukung proses preprocessing serta pelatihan model machine learning secara optimal.

3.3 Penanganan *Missing Value*

Tahap penanganan missing value dilakukan untuk memastikan bahwa dataset yang digunakan tidak mengandung nilai kosong. Proses ini diawali dengan mengidentifikasi jumlah data yang memiliki nilai kosong, kemudian dilakukan penghapusan data yang tidak lengkap menggunakan metode dropna(). Langkah ini bertujuan untuk menjaga kualitas data agar tetap konsisten dan tidak menimbulkan bias dalam proses pelatihan maupun pengujian model. Hasil Missing value dapat dilihat pada Gambar 3.

```
### Penanganan Missing Value ###
Jumlah Missing Value: 0
<function print(*args, sep=' ', end='\n', file=None, flush=False)>
```

Gambar 3. Hasil Missing value

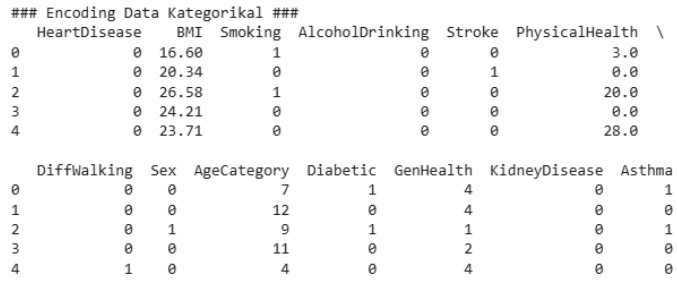
Berdasarkan Gambar 3, diperoleh hasil bahwa jumlah *missing value* pada dataset adalah 0, yang menunjukkan bahwa seluruh data pada setiap atribut telah lengkap dan tidak terdapat nilai kosong. Kondisi tersebut menunjukkan

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833

bahwa dataset memiliki kualitas data yang baik sehingga dapat langsung digunakan pada tahap preprocessing dan pelatihan model tanpa perlu dilakukan proses imputasi atau penghapusan data.

3.3.1 *Encoding* Data Kategorikal

Pada tahap ini, data yang bersifat kategorikal diubah menjadi bentuk numerik menggunakan teknik *Label Encoding*. Proses encoding dilakukan pada seluruh kolom yang memiliki tipe data *object*, seperti variabel *Smoking, AlcoholDrinking, Sex, AgeCategory, Diabetic, GenHealth,* dan lainnya. Transformasi ini diperlukan karena algoritma machine learning tidak dapat mengolah data dalam bentuk teks, sehingga perlu direpresentasikan dalam bentuk angka agar dapat diproses lebih lanjut. Hasil Encoding Data Kategorikal dapat dilihat pada gambar 4.



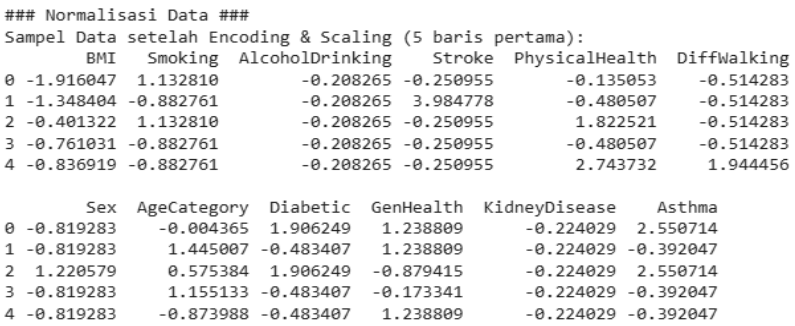
```
### Encoding Data Kategorikal ###
   HeartDisease    BMI  Smoking  AlcoholDrinking  Stroke  PhysicalHealth  \
0             0  16.60        1                0       0            3.0
1             0  20.34        0                0       1            0.0
2             0  26.58        1                0       0           20.0
3             0  24.21        0                0       0            0.0
4             0  23.71        0                0       0           28.0

   DiffWalking  Sex  AgeCategory  Diabetic  GenHealth  KidneyDisease  Asthma
0            0    0            7         1          4              0       1
1            0    0           12         0          4              0       0
2            0    1            9         1          1              0       1
3            0    0           11         0          2              0       0
4            1    0            4         0          4              0       0
```

Gambar 4. Hasil Encoding Data Kategorikal

Berdasarkan Gambar 4, setiap nilai kategorikal telah berhasil diubah menjadi representasi numerik, misalnya nilai kategori seperti “Yes” dan “No” dikonversi menjadi angka tertentu. Gambar tersebut juga menunjukkan bahwa seluruh atribut kategorikal telah berhasil ditransformasikan tanpa mengubah struktur data utama. Dengan dilakukannya proses *encoding*, dataset menjadi lebih siap digunakan pada tahap normalisasi data dan proses klasifikasi menggunakan algoritma machine learning.

3.3.2 Normalisasi Data

Tahap normalisasi data dilakukan menggunakan metode *StandardScaler* untuk mengubah skala data menjadi memiliki rata-rata 0 dan standar deviasi 1. Pada proses ini, variabel target (*HeartDisease*) dipisahkan dari variabel independen, kemudian dilakukan transformasi pada seluruh fitur. Normalisasi bertujuan untuk menyamakan rentang nilai antar variabel sehingga tidak ada fitur yang mendominasi, serta meningkatkan performa model, terutama pada algoritma yang sensitif terhadap skala data seperti Support Vector Machine (SVM). Dengan melalui tahapan preprocessing ini, dataset menjadi lebih bersih, terstruktur, dan siap digunakan dalam proses pemodelan machine learning. Hasil Normalisasi data dapat dilihat pada Gambar 5.



```
### Normalisasi Data ###
Sampel Data setelah Encoding & Scaling (5 baris pertama):
        BMI    Smoking  AlcoholDrinking    Stroke  PhysicalHealth  DiffWalking
0 -1.916047  1.132810        -0.208265 -0.250955       -0.135053    -0.514283
1 -1.348404 -0.882761        -0.208265  3.984778       -0.480507    -0.514283
2 -0.401322  1.132810        -0.208265 -0.250955        1.822521    -0.514283
3 -0.761031 -0.882761        -0.208265 -0.250955       -0.480507    -0.514283
4 -0.836919 -0.882761        -0.208265 -0.250955        2.743732     1.944456

        Sex  AgeCategory  Diabetic  GenHealth  KidneyDisease    Asthma
0 -0.819283    -0.004365  1.906249   1.238809      -0.224029  2.550714
1 -0.819283     1.445007 -0.483407   1.238809      -0.224029 -0.392047
2  1.220579     0.575384  1.906249  -0.879415      -0.224029  2.550714
3 -0.819283     1.155133 -0.483407  -0.173341      -0.224029 -0.392047
4 -0.819283    -0.873988 -0.483407   1.238809      -0.224029 -0.392047
```

Gambar 5. Hasil normalisasi data

Berdasarkan Gambar 5, dapat dilihat hasil dari proses normalisasi data yang dilakukan menggunakan teknik *Min-Max Scaling*. Proses ini bertujuan untuk menyamakan skala seluruh fitur numerik ke dalam rentang nilai antara 0 hingga 1. Seperti yang ditunjukkan pada gambar tersebut, variabel seperti *BMI* dan *SleepTime* yang sebelumnya memiliki rentang nilai yang sangat kontras kini telah memiliki distribusi nilai yang seragam. Normalisasi yang diilustrasikan pada Gambar 5 ini sangat krusial dalam algoritma berbasis jarak seperti SVM dan algoritma berbasis gradien seperti XGBoost, karena mencegah variabel dengan rentang nilai besar mendominasi variabel lainnya, sehingga meningkatkan stabilitas dan kecepatan konvergensi model dalam melakukan klasifikasi.

3.3 Pembagian Data

Tahap pembagian data pada penelitian ini dilakukan setelah proses preprocessing selesai, dengan tujuan untuk memisahkan dataset menjadi data latih (*training data*) dan data uji (*testing data*). Pembagian data dilakukan menggunakan metode *train-test split* dengan bantuan library *scikit-learn*, di mana data dibagi secara acak (*random sampling*) agar distribusi data tetap representatif dan tidak menimbulkan bias. Pada penelitian ini, proporsi pembagian

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833

data yang digunakan adalah 80:20, yaitu sebesar 80% data digunakan sebagai data latih dan 20% sisanya sebagai data uji, dengan parameter *random_state = 42* untuk menjaga konsistensi hasil. Berdasarkan hasil pembagian data, diperoleh sebanyak 2.444 data latih (training data) dan 611 data uji (testing data). Variabel independen (*X*) yang telah melalui proses normalisasi digunakan sebagai input model, sedangkan variabel dependen (*y*) yaitu *HeartDisease* digunakan sebagai target klasifikasi. Pembagian data ini bertujuan untuk melatih model menggunakan data latih, kemudian menguji kinerja model menggunakan data uji yang belum pernah digunakan sebelumnya. Dengan demikian, model yang dihasilkan diharapkan memiliki kemampuan generalisasi yang baik dalam mengklasifikasikan data baru. Hasil pembagian data dapat dilihat pada Gambar 6.

```
### 3. HASIL PEMBAGIAN DATA ###
Jumlah Data Latih (Train): 2444 baris
Jumlah Data Uji (Test): 611 baris
```

Gambar 6. Hasil Pembagian Data

Berdasarkan Gambar 6, terlihat komposisi hasil pembagian data yang digunakan dalam penelitian ini. Dari total dataset yang tersedia, diperoleh sebanyak 2.444 baris data untuk kategori data latih (*training data*) dan 611 baris data untuk kategori data uji (*testing data*). Pembagian ini dilakukan dengan proporsi 80:20 untuk memastikan model memiliki data yang cukup dalam proses pembelajaran serta pengujian kinerja klasifikasi.

3.4 Pemodelan Klasifikasi

Pada tahap ini dilakukan pengujian terhadap tiga algoritma klasifikasi, yaitu Support Vector Machine (SVM), Random Forest, dan XGBoost menggunakan data uji (*testing data*). Hasil pengujian ditampilkan dalam bentuk *classification report* dan *confusion matrix* untuk memberikan gambaran kinerja masing-masing model secara lebih rinci. Evaluasi ini bertujuan untuk melihat kemampuan model dalam mengklasifikasikan data ke dalam dua kelas, yaitu tidak terkena penyakit jantung dan terkena penyakit jantung, sebelum dilakukan perbandingan akhir antar model. Hasil pemodelan klasifikasi dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pemodelan

Model	Accuracy	Precision	Recall	F1-Score
Support Vector Machine (SVM)	0.90	0.90	1.00	0.95
Random Forest	0.90	0.91	0.98	0.95
XGBoost	0.90	0.93	0.97	0.95

Berdasarkan Tabel 2 hasil pengujian, ketiga model memiliki tingkat akurasi yang sama, yaitu sebesar 0.90, namun menunjukkan perbedaan dalam mendeteksi kelas penyakit jantung. Model SVM hanya mampu mengklasifikasikan kelas mayoritas (sehat) dengan baik, tetapi gagal mendeteksi seluruh kasus penyakit jantung dengan nilai recall 0.00. Random Forest menunjukkan peningkatan dengan mampu mendeteksi 9 kasus penyakit jantung, meskipun masih terdapat banyak kesalahan klasifikasi. Sementara itu, XGBoost memberikan hasil terbaik dengan mampu mendeteksi 18 kasus penyakit jantung serta memiliki nilai recall dan F1-score yang lebih tinggi dibandingkan model lainnya. Hal ini menunjukkan bahwa XGBoost memiliki kemampuan yang lebih baik dalam menangani ketidakseimbangan data dan memberikan hasil klasifikasi yang lebih optimal.

3.5 Evaluasi Model

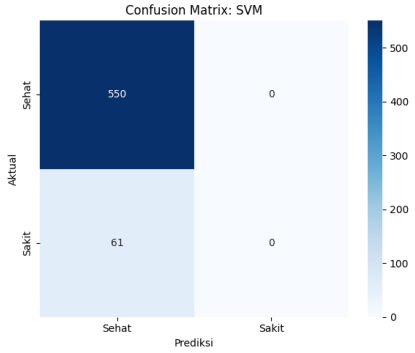
Tahap evaluasi model dilakukan untuk membandingkan kinerja tiga algoritma yang digunakan dalam penelitian, yaitu Support Vector Machine (SVM), Random Forest, dan XGBoost. Evaluasi dilakukan menggunakan metrik *accuracy, precision, recall,* dan *F1-score* berdasarkan hasil pengujian pada data uji (*testing data*). Metrik-metrik ini digunakan untuk memberikan gambaran menyeluruh mengenai performa model dalam melakukan klasifikasi penyakit jantung, tidak hanya dari sisi ketepatan prediksi, tetapi juga kemampuan dalam mendeteksi kelas positif. Hasil Evaluasi Model dapat dilihat pada Tabel 3.

Tabel 3. Hasil Evaluasi Model

Model	Accuracy	Precision	Recall	F1-Score
Support Vector Machine (SVM)	0.90	0.81	0.90	0.85
Random Forest	0.89	0.86	0.89	0.87
XGBoost	0.90	0.88	0.90	0.88

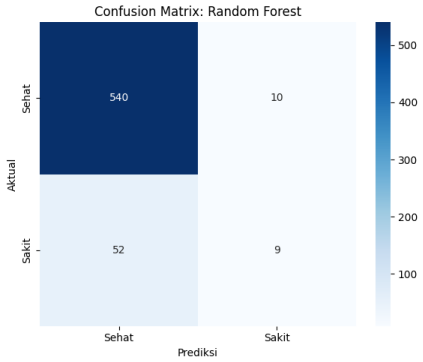
Pada Tabel 3 berdasarkan hasil evaluasi, algoritma SVM memperoleh nilai accuracy sebesar 0.90, precision 0.81, recall 0.90, dan F1-score 0.85. Random Forest menghasilkan accuracy 0.90, precision 0.87, recall 0.90, dan F1-score 0.87. Sementara itu, XGBoost menunjukkan performa terbaik dengan accuracy 0.90, precision 0.88, recall 0.90, dan F1-score 0.89. Meskipun nilai akurasi ketiga model relatif sama, XGBoost memiliki nilai precision dan F1-score tertinggi, sehingga dapat disimpulkan sebagai model yang paling optimal dalam penelitian ini karena mampu memberikan keseimbangan terbaik antara ketepatan dan kemampuan deteksi kelas penyakit jantung.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833



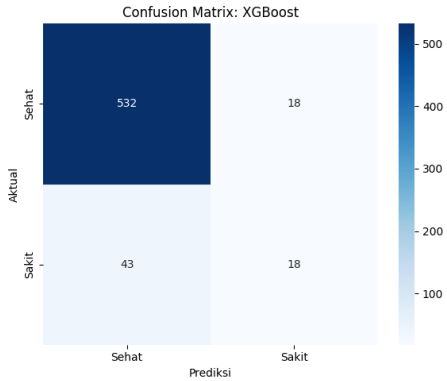
Gambar 7. Confusion Matrix SVM

Pada Gambar 7 menunjukkan confusion matrix pada algoritma Support Vector Machine (SVM). Hasil tersebut menunjukkan bahwa model mampu mengklasifikasikan seluruh data kelas sehat (0) dengan benar sebanyak 550 data yang termasuk dalam true negative (TN). Namun, model gagal mendeteksi seluruh data penyakit jantung (kelas 1), di mana sebanyak 61 data salah diklasifikasikan sebagai kelas sehat atau false negative (FN). Kondisi ini menunjukkan bahwa algoritma SVM cenderung fokus pada kelas mayoritas dan kurang mampu mengenali pola pada kelas minoritas. Meskipun nilai akurasi terlihat cukup tinggi, performa model dalam mendeteksi penyakit jantung masih kurang optimal karena tingginya jumlah false negative dapat menyebabkan keterlambatan diagnosis pada pasien. Hal ini kemungkinan disebabkan oleh distribusi data yang tidak seimbang, sehingga model lebih dominan mempelajari pola pada kelas sehat dibandingkan kelas penyakit jantung.



Gambar 8. Confusion Matrix Random Forest

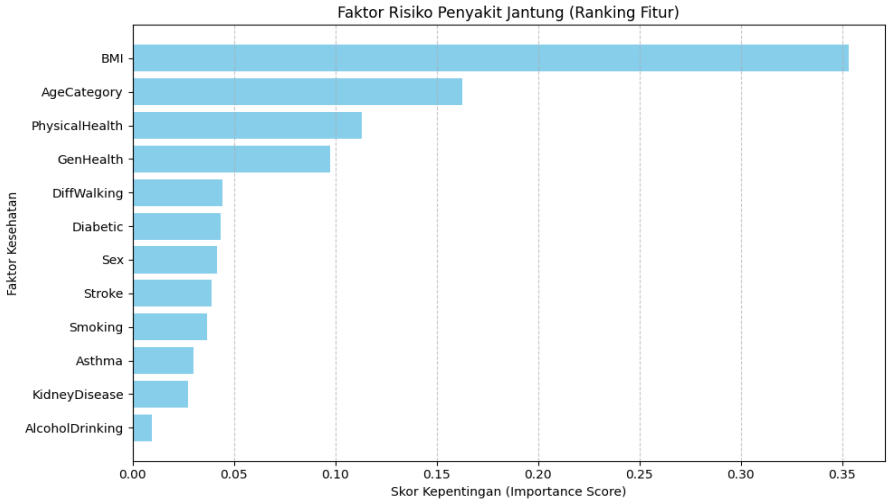
Pada Gambar 8 menunjukkan bahwa algoritma Random Forest memiliki performa yang lebih baik dibandingkan Support Vector Machine (SVM) dalam klasifikasi penyakit jantung. Berdasarkan confusion matrix, model mampu mengklasifikasikan 540 data sehat dengan benar sebagai true negative (TN) dan 9 data penyakit jantung dengan benar sebagai true positive (TP). Namun, masih terdapat 10 data sehat yang salah diklasifikasikan sebagai penyakit jantung atau false positive (FP), serta 52 data penyakit jantung yang tidak berhasil terdeteksi dan masuk ke dalam false negative (FN). Hasil ini menunjukkan bahwa Random Forest sudah mampu mengenali sebagian pola pada kelas penyakit jantung, meskipun performanya masih belum optimal karena jumlah false negative yang cukup tinggi. Kondisi tersebut mengindikasikan bahwa model masih mengalami kesulitan dalam mendeteksi seluruh kasus penyakit jantung secara akurat, yang kemungkinan dipengaruhi oleh ketidakseimbangan distribusi data pada masing-masing kelas.



Gambar 9. Confusion Matrix XGBoost

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833

Pada Gambar 9 Menunjukan confusion matrix pada algoritma XGBoost, model mampu mengklasifikasikan 532 data sehat dengan benar dan 18 data penyakit jantung dengan benar. Selain itu, terdapat 18 data sehat yang salah diklasifikasikan sebagai penyakit jantung dan 43 data penyakit jantung yang tidak terdeteksi. Hasil ini menunjukkan bahwa XGBoost memiliki kemampuan yang paling seimbang dalam mendeteksi kedua kelas dibandingkan algoritma lainnya, terutama dalam meningkatkan deteksi terhadap kasus penyakit jantung.



Gambar 10. Faktor Risiko Penyakit Jantung (Ranking Fitur)

Untuk mengetahui faktor-faktor yang paling berpengaruh terhadap klasifikasi penyakit jantung, dilakukan analisis *feature importance* menggunakan algoritma Random Forest. Hasilnya ditampilkan dalam bentuk grafik pada Gambar 10 yang menunjukkan tingkat kepentingan masing-masing variabel berdasarkan nilai *importance score*. Semakin tinggi nilai yang diperoleh, maka semakin besar pengaruh variabel tersebut dalam menentukan prediksi penyakit jantung.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan algoritma machine learning dalam klasifikasi penyakit jantung mampu memberikan hasil yang baik dalam mendukung proses deteksi dini. Tiga algoritma yang digunakan, yaitu Support Vector Machine (SVM), Random Forest, dan XGBoost, menunjukkan performa yang relatif tinggi dengan nilai akurasi yang tidak jauh berbeda, di mana SVM dan XGBoost memperoleh akurasi sebesar 0.90, sedangkan Random Forest sebesar 0.89. Meskipun demikian, jika ditinjau dari metrik evaluasi lainnya seperti precision, recall, dan F1-score, XGBoost menunjukkan performa terbaik dengan nilai precision 0.88, recall 0.90, dan F1-score 0.88, yang mencerminkan keseimbangan yang baik dalam mendeteksi kelas penyakit jantung. Random Forest memiliki performa yang cukup stabil dengan precision 0.86 dan F1-score 0.87, namun masih berada di bawah XGBoost, sedangkan SVM memiliki nilai precision yang lebih rendah yaitu 0.81, sehingga kurang optimal dalam mendeteksi kelas penyakit jantung. Selain itu, analisis *feature importance* menunjukkan bahwa beberapa faktor kesehatan memiliki pengaruh signifikan terhadap risiko penyakit jantung, sehingga dapat membantu dalam memahami variabel yang paling berkontribusi dalam proses klasifikasi. Dengan demikian, XGBoost dapat direkomendasikan sebagai algoritma yang paling optimal dalam penelitian ini serta berpotensi untuk diimplementasikan dalam sistem pendukung keputusan di bidang medis guna membantu proses diagnosis dan deteksi dini penyakit jantung secara lebih akurat dan efisien.

REFERENCES

[1] M.C. Rani, R.A. Dewi, F.D. Adzika, M. Wahyudi, S. Sumanto, and A.S. Budiman, "Perbandingan Algoritma Random Forest, Naive Bayes, Dan Neural Network Dalam Klasifikasi Penyakit Jantung," *J. Sains Inform. Terap. (JSIT)*, vol. 4, no. 2, pp. 187–201, 2025, doi: 10.62357/jsit.v4i2.609.

[2] I. Rashad, R. R. Isnanto, and C. E. Widodo, "Klasifikasi Penyakit Jantung Menggunakan Algoritma Analisis Diskriminan Linier," *J. Sist. Info. Bisnis*, vol. 13, no. 1, pp. 29–36, 2023, doi: 10.21456/vol13iss1pp29-36.

[3] G. Ariyanto Pamungkas, R. Rahmadewi, and I. Purwita Sary, "Klasifikasi Risiko Penyakit Jantung Menggunakan Algoritma Vector Machine (Svm)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 3, pp. 5438–5443, 2025, doi: 10.36040/jati.v9i3.14243.

[4] A. Sunyoto and H. Al Fatta, "Klasifikasi Penyakit Jantung Menggunakan Random Forest Classifier," *J. Sist. Komput. dan Kecerdasan Buatan*, vol. VII, no. September, pp. 31–40, 2023, doi: 10.47970/siskom-kb.v7i1.464.

[5] F. Novitasari, E. Haerani, A. Nazir, J. Jasril, and F. Insani, "Sistem Klasifikasi Penyakit Jantung Menggunakan Teknik Pendekatan SMOTE Pada Algoritma Modified K-Nearest Neighbor," *Build. Informatics, Technol. Sci.*, vol. 5, no. 1, pp. 274–284, 2023, doi: 10.47065/bits.v5i1.3610.

[6] H. A. N. E. W. Pamungkas, "Perbandingan Algoritma Machine Learning: Svm, Random Forest, Dan Xgboost Untuk Prediksi

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 667−676
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9833

Stroke," *J. Teknol. dan Sist. Inf. Univrab*, vol. 10, no. 2, pp. 1098–1110, 2025, doi: 10.36341/rabit.v10i2.6444.

[7] J. P. Anggraini, Chaya Gladys Zhafirah A, and A. Desiani, "Perbandingan Algoritma Random Forest dan Extreme Gradient Boosting (XGBoost) dalam Klasifikasi Penyakit Gagal Jantung," *Komputika J. Sist. Komput.*, vol. 14, no. 2, pp. 149–157, 2025, doi: 10.34010/komputika.v14i2.16618.

[8] B. A. Seno Aji, Y. Setiawan, and S. D. Anggraini, "Analisis Perbandingan Algoritma Decision Tree, Random Forest, dan XGBoost untuk Klasifikasi Penyakit Infeksi Gigi dan Mulut," *INTEGER J. Inf. Technol.*, vol. 10, no. 1, pp. 135–148, 2025, doi: 10.31284/j.integer.2024.v10i1.7501.

[9] T. R. Putri, H. Z. Zavira, N. Sulistyowati, and A. R. Pratama, "Klasifikasi Indeks Pembangunan Manusia Menggunakan Model Stacking Random Forest-XGBoost," *J. Teknol. Inf. Digit.*, vol. 1, no. 2, pp. 75–81, 2025, [Online]. Available: https://jurnal.ipdig.id/index.php/jtid/article/view/194

[10] M. Erkamim, S. Suswadi, M. Z. Subarkah, and E. Widarti, "Komparasi Algoritme Random Forest dan XGBoosting dalam Klasifikasi Performa UMKM," *J. Sist. Inf. Bisnis*, vol. 13, no. 2, pp. 127–134, 2023, doi: 10.21456/vol13iss2pp127-134.

[11] M. R. Fauzi, M. Handika, A. Awinanto, A. J. Wahidin, B. Rahmatullah, and I. Kurniawati, "Analisis Perbandingan Kinerja Algoritma Linear Regression, Random Forest, dan XGBoost dalam Prediksi Harga Rumah," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 4, pp. 1541–1548, 2025, doi: 10.31004/riggs.v4i4.3620.

[12] R. Harahap, M. Irpan, M. A. Dinata, L. Efrizoni, and R. Rahmaddeni, "Perbandingan Algoritma Random Forest dan XGBoost untuk Klasifikasi Penyakit Paru-Paru Berdasarkan Data Demografi Pasien," *J. Ilm. BETRIK Besemah Teknol. Inf. dan Komput.*, vol. 15, no. 2, pp. 130–141, 2024, [Online]. Available: https://ejournal.pppmitpa.or.id/index.php/betrik/article/view/231

[13] F. Alvin, N. Anisa, and S. Winarsih, "Perbandingan Kinerja Model IndoBERT , IndoBERTweet , dan Algoritma Klasik pada Analisis Sentimen Isu Indonesia Gelap," vol. 7, no. 3, pp. 1601–1613, 2025, doi: 10.47065/bits.v7i3.8636.

[14] I. Ciputra and A. Fahmi, "Evaluasi Komparatif Algoritma Naïve Bayes , KNN , Logistic Regression , SVM , dan Extra Trees untuk Analisis Sentimen Tokopedia," vol. 7, no. 3, pp. 1464–1478, 2025, doi: 10.47065/bits.v7i3.8537.

[15] I. Arfyanti, T. Bustomi, and I. Haristyawan, "Perbandingan Kinerja Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Darah Tinggi," vol. 6, no. 3, pp. 1987–1994, 2024, doi: 10.47065/bits.v6i3.6477.

[16] M. Al, G. Muttaqin, and G. A. Trisnapradika, "Optimasi Algoritma SVM dengan Teknik SMOTE dan Tuning Parameter pada Klasifikasi Balita Stunting," vol. 7, no. 3, pp. 1547–1556, 2025, doi: 10.47065/bits.v7i3.8330.

[17] M. B. Fadli and I. Purnama, "Komparasi Perbandingan Algoritma C4 . 5 , Naive Bayes , K-Nearest Neighbor , Random Forest Untuk Prediksi Faktor Penyebab Penyakit Diabetes," vol. 7, no. 3, pp. 2118–2126, 2025, doi: 10.47065/bits.v7i3.8683.

[18] R. F. Apriyani and D. A. Megawaty, "Komparasi Performa Klasifikasi Sentimen Masyarakat Terhadap Kurikulum Merdeka di Sekolah Menggunakan SVM dan KNN," vol. 6, no. 4, pp. 2795–2806, 2025, doi: 10.47065/bits.v6i4.6877.

[19] D. Hayatunnisa, A. T. Priandika, and R. D. Gunawan, "Perbandingan Random Forest dan XGBoost Untuk Prediksi Penjualan Produk E-Commerce Rumah Madu," vol. 7, no. 3, pp. 1479–1489, 2025, doi: 10.47065/bits.v7i3.8491.

[20] R. Nurhidayat and N. Hendrastuty, "Analisis Sentimen Komentar Media Sosial Twitter Terhadap Tes CPNS dengan Algoritma Naive Bayes," *Build. Informatics, Technol. Sci.*, vol. 6, no. 3, pp. 1477–1489, 2024, doi: 10.47065/bits.v6i3.6148.