Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

Komparasi Model Ensemble dan Algoritma Machine Learning Untuk Memprediksi Penyakit Jantung

Muhammad Syarief Albani, Dedy Kurniawan*, Ken Ditha Tania

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Sriwijaya, Palembang, Indonesia
Email: [1]muhammadsyariefalbani@gmail.com, [2,*]dedykurniawan@unsri.ac.id, [3]kenya.tania@gmail.com
Email Penulis Korespondensi: dedykurniawan@unsri.ac.id
Submitted: **09/09/2025**; Accepted: **19/03/2026**; Published**: 19/03/2026**

Abstrak−Studi ini membandingkan performa sembilan algoritma machine learning dalam memprediksi penyakit jantung dengan menggunakan dataset berasal dari tahun 1988 dan terdiri dari empat basis data: Cleveland, Hungaria, Swiss, dan Long Beach sebanyak 1025 data. Dataset yang digunakan mencakup fitur-fitur medis yang mencerminkan keadaan fisiologis, hasil pemeriksaan klinis, dan faktor risiko kardiovaskular, yaitu usia, jenis kelamin, jenis nyeri dada, tekanan darah saat istirahat, kadar kolesterol dalam serum, kadar gula darah puasa, hasil elektrokardiografi saat istirahat, denyut jantung maksimum, nyeri dada saat aktivitas fisik, depresi segmen ST, kemiringan segmen ST, jumlah pembuluh darah utama yang terlihat dengan fluoroskopi, serta keadaan thalassemia. Tahapan penelitian ini meliputi pembersihan data, transformasi data, dan evaluasi yang dilakukan dengan metode pembagian data untuk pelatihan dan pengujian serta K-fold cross-validation dengan metrik akurasi, presisi, recall, skor F1, dan AUC-ROC. Algoritma yang digunakan pada penelitian ini adalah Decision Tree, Random Forest, Support Vector Machine, MLP Classifier, Bagging Classifier, Gradient Boosting, CatBoost, XGBoost, dan LightGBM dengan model berbasis ensemble, seperti CatBoost, Random Forest, XGBoost, dan LightGBM, menunjukkan kinerja yang konsisten pada berbagai metrik evaluasi jika dibandingkan dengan model non-ensemble. Di antara semua model yang diuji, CatBoost menunjukkan kinerja terbaik, dengan akurasi mencapai 98%, F1-Score sebesar 0,980, dan Recall sebesar 0,9875 yang kemudian disusul dengan algortima ensemble lainnya seperti Random forest, XGboost dan LightGBM. Hasil penelitian ini menunjukkan bahwa model ensemble terbukti lebih efektif dalam memprediksi penyakit jantung. Penelitian ini berkontribusi untuk menyajikan studi komparatif yang mendalam mengenai performa algoritma ensemble dan machine learning moderen dalam prediksi penyakit jantung, serta memperkaya literatur yang berkaitan dengan penerapan Knowledge Discovery di bidang kesehatan dan memberikan dasar untuk pemilihan algoritma prediksi yang lebih handal dalam mendukung pengambilan keputusan klinis serta pengembangan sistem dukungan diagnosis penyakit jantung berbasis machine learning.

Kata Kunci: Machine Learning; Penyakit Jantung; Model Ensemble

Abstract−This study compared the performance of nine machine learning algorithms in predicting heart disease using a dataset dating back to 1988 and consisting of four databases: Cleveland, Hungary, Switzerland, and Long Beach totaling 1025 data. The dataset used includes medical features that reflect physiological states, clinical examination results, and cardiovascular risk factors, namely age, gender, type of chest pain, resting blood pressure, serum cholesterol levels, fasting blood sugar levels, resting electrocardiography results, maximum heart rate, chest pain during physical activity, ST segment depression, ST segment slope, number of major blood vessels visible by fluoroscopy, and thalassemia status. The stages of this study include data cleaning, data transformation, and evaluation carried out using the data splitting method for training and testing as well as K-fold cross-validation with metrics of accuracy, precision, recall, F1 score, and AUC-ROC. The algorithms used in this study are Decision Tree, Random Forest, Support Vector Machine, MLP Classifier, Bagging Classifier, Gradient Boosting, CatBoost, XGBoost, and LightGBM with ensemble-based models, such as CatBoost, Random Forest, XGBoost, and LightGBM, showing consistent performance on various evaluation metrics when compared to non-ensemble models. Among all models tested, CatBoost showed the best performance, with an accuracy reaching 98%, an F1-Score of 0.980, and a Recall of 0.9875 then followed by other ensemble algorithms such as Random Forest, XGBoost and LightGBM. The results of this study indicate that ensemble models are proven to be more effective in predicting heart disease. This study aims to present an in-depth comparative study of the performance of ensemble algorithms and modern machine learning in predicting heart disease, as well as enriching the literature related to the application of Knowledge Discovery in the health sector and providing a basis for selecting more reliable prediction algorithms to support clinical decision making and the development of machine learning-based heart disease diagnosis support systems.

Keywords: Machine Learning; Heart Disease; Ensemble Model

1. PENDAHULUAN

Survei terkini yang dilakukan oleh Organisasi Kesehatan Dunia menunjukkan bahwa 17,9 juta kematian global diakibatkan oleh penyakit kardiovaskular dan diperkirakan akan mencapai angka yang mencemaskan yaitu 23 juta pada tahun 2030 [1] Di antara Penyakit Kardiovaskular, penyakit jantung dianggap yang paling prevalen dan menyumbang sebagian besar beban kesehatan secara global. Penyakit jantung belakangan ini menjadi faktor utama kematian di kalangan orang muda dan paruh baya, yang sangat mempengaruhi ketahanan psikologis keluarga serta meningkatkan tanggungan finansial [2] . Di abad ke-21, Kecerdasan Buatan telah menjadi elemen krusial dalam industri kesehatan, salah satu aspek dari kemajuan Kecerdasan Buatan adalah *machine learning.* Tren pemanfaatan *Machine learning* untuk diagnosis penyakit berkembang pesat, terutama dalam penciptaan model prediksi guna mendeteksi penyakit kronis seperti kanker, diabetes, dan penyakit jantung [3] .

Kemajuan dalam *machine learning* dan analisis data telah mendorong lahirnya konsep Penemuan Pengetahuan, yaitu proses sistematik untuk mengidentifikasi pola, hubungan, dan informasi baru dari kumpulan data yang kompleks. Penemuan Pengetahuan telah menjadi fokus penelitian yang krusial untuk mendukung proses pengambilan keputusan [4]. Meski model *ML* telah diteliti secara mendalam dan dianggap sangat efektif, prediksi

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

penyakit jantung merupakan masalah kompleks dan masih terdapat banyak perbaikan yang diperlukan serta metode yang perlu dieksplorasi [5].

Penelitian terdahulu yang dilakukan oleh Dhita pada tahun 2025 [6] menerapkan beberapa algoritma klasifikasi seperti Logistic Regression, Random Forest, XGBoost, Support Vector Machine, dan Gradient Boosting untuk memprediksi customer churn pada sektor e-commerce, di mana XGBoost menunjukkan performa terbaik dengan akurasi mencapai 96% dan AUC sebesar 0,999. Namun demikian, penelitian tersebut berfokus pada domain bisnis dan perilaku pelanggan, sehingga belum mengeksplorasi penerapan metode tersebut dalam konteks prediksi penyakit berbasis data klinis. Penelitian lain yang dilakukan Alrayssa pada tahun 2025 yang berjudul Penerapan Metode Machine Learning dan Teknik SMOTE untuk Prediksi Diabetes [7] menggunakan beberapa algoritma machine learning seperti *Random Forest, XGBoost, Support Vector Machine,* dan *K-Nearest Neighbor* untuk memprediksi penyakit diabetes dengan bantuan teknik penyeimbangan data SMOTE. Hasil penelitian menunjukkan bahwa Random Forest dan XGBoost memberikan performa terbaik dengan nilai akurasi, precision, recall, dan *F1-score* mencapai 0,97. Meskipun menghasilkan performa yang sangat baik, penelitian ini hanya membandingkan empat algoritma sehingga ruang eksplorasi terhadap algoritma *machine learning* lain, terutama model *ensemble* moderen masih terbatas. Studi yang dilakukan oleh Cheisya pada tahun 2025 [8] menggunakan model *Random Forest, XGBoost, SVM,* dan *KNN* untuk mendeteksi diabetes. Hasil penelitian menunjukkan bahwa Random Forest dan XGBoost memberikan kinerja terbaik dengan nilai *Accuracy, Precision, Recall*, dan *F1-Score* mencapai 0,97. Selain itu, Penelitian yang dilakukan oleh Raviansyah pada tahun 2025 yang berjudul Komparasi algoritma machine learning (random forest, gradient boosting, dan ada boosting) untuk prediksi tingkat penyakit Alzheimer [9] membandingkan algoritma ensemble seperti *Random Forest, Gradient Boosting,* dan *AdaBoost* dalam memprediksi risiko penyakit Alzheimer, di mana *Gradient Boosting* memperoleh performa terbaik dengan akurasi sebesar 0,956 dan AUC sebesar 0,955. Meskipun penelitian ini menunjukkan potensi besar algoritma *ensemble* dalam bidang kesehatan, jumlah algoritma yang dibandingkan masih relatif terbatas sehingga belum memberikan gambaran komprehensif mengenai kinerja berbagai algoritma machine learning modern dalam kasus medis yang lebih kompleks. Penelitian yang dilakukan oleh Alfajr et al pada tahun 2025 [10] membandingkan algoritma *Random Forest Classifier* dan *Support Vector Machine* untuk meramalkan penyakit jantung dengan memanfaatkan 5432 data rekam medis dari RSUD Kabupaten Bekasi. Melalui proses KDD dan penanganan ketidakseimbangan menggunakan *SMOTE*, kedua model diuji dengan rasio pembagian data 60:40. Evaluasi membuktikan bahwa *SVM* memberikan hasil terbaik dengan akurasi 65%, presisi 70%, recall 68%, dan f1-score 64%, sementara *Random Forest* hanya mencapai sekitar 61% pada semua metrik. Nilai *AUC SVM* (0,67–0,68) juga lebih tinggi dibandingkan *RFC*, meskipun tetap tergolong dalam kategori klasifikasi yang buruk. Temuan ini mengindikasikan bahwa *SVM* lebih efektif dalam meramalkan penyakit jantung pada dataset tersebut. Walaupun penerapan *machine learning* untuk memprediksi sudah banyak diteliti, sebagian besar penelitian sebelumnya masih menghadapi keterbatasan.

Terdapat beberapa batasan dalam penelitian terdahulu. Pertama, sebagian besar penelitian hanya membandingkan jumlah algoritma yang relatif terbatas, sehingga belum memberikan gambaran menyeluruh tentang kinerja berbagai algoritma machine learning dalam mengatasi data kesehatan yang kompleks. Kedua, studi yang berfokus pada prediksi penyakit jantung masih didominasi oleh algoritma tradisional seperti Random Forest dan Support Vector Machine, sementara algoritma ensemble terbaru seperti CatBoost dan LightGBM, yang dirancang untuk meningkatkan kinerja klasifikasi pada data yang kompleks, masih jarang diujicobakan secara bersamaan. Ketiga, sejumlah penelitian yang telah dilaksanakan mengenai prediksi penyakit jantung masih menunjukkan hasil klasifikasi yang belum maksimal, sehingga diperlukan penjelajahan metode yang lebih luas untuk menemukan algoritma yang paling efisien dalam memodelkan ciri-ciri data medis. Dengan mempertimbangkan batasan tersebut, muncul kesenjangan penelitian (research gap) yang mengindikasikan bahwa masih sedikit studi yang membandingkan secara menyeluruh berbagai algoritma machine learning konvensional, neural network, dan ensemble modern dalam kerangka eksperimen yang serupa untuk memprediksi penyakit jantung. Namun, analisis komparatif yang lebih mendalam sangat krusial untuk memahami ciri-ciri kinerja setiap algoritma serta menemukan metode yang paling efisien dalam memodelkan data kesehatan.

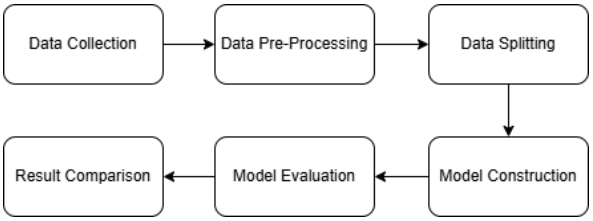
Untuk mengurangi kesenjangan tersebut, studi ini melakukan analisis perbandingan terhadap sembilan algoritma machine learning, yaitu *Random Forest, Decision Tree, Support Vector Machine, MLP Classifier, Bagging Classifier, Gradient Boosting, CatBoost, XGBoost,* dan *LightGBM* dalam memprediksi penyakit jantung. Model dievaluasi dengan memanfaatkan berbagai metrik kinerja seperti akurasi, presisi, *recall,* dan *F1-score*, serta divalidasi melalui teknik *K-Fold Cross Validation* guna menjamin kestabilan dan keandalan hasil model. Adapun kontribusi utama penelitian ini adalah untuk Menyediakan analisis komparatif yang lebih komprehensif dengan membandingkan sembilan algoritma machine learning yang mencakup model klasifikasi dasar, neural network, serta algoritma ensemble modern, Mengevaluasi kinerja algoritma ensemble modern seperti CatBoost, XGBoost, dan LightGBM dalam konteks prediksi penyakit jantung yang masih relatif terbatas dalam penelitian sebelumnya. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan metode prediksi penyakit jantung yang lebih akurat serta memperkaya literatur mengenai penerapan machine learning dan Knowledge Discovery dalam bidang Kesehatan. Selain itu, hasil penelitian ini juga diharapkan dapat memberikan dasar empiris bagi peneliti selanjutnya maupun praktisi di bidang kesehatan dalam memilih algoritma machine learning yang paling efektif dan andal untuk mendukung sistem prediksi penyakit jantung berbasis data klinis.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini melibatkan beberapa tahap yang dilakukan untuk mencapai tujuan yang telah ditetapkan



Gambar 1. Tahapan Penelitian

Gambar 1 ini mengilustrasikan keseluruhan tahapan penelitian dari awal hingga akhir Tahap pertama pada penelitian ini adalah Pengumpulan Data (*Data Collection)* untuk memahami karakteristik dan korelasi atribut pada dataset. Setelah itu dilakukan Pra-pemrosesan Data (Preprocessing Data) yang mencakup Pembersihan Data (Cleaning Data) dan Pengubahan Data (Transformasi data), termasuk proses transformasi untuk mempersiapkan format input yang sesuai bagi algoritma machine learning. kemudian pembagian data (split data) dengan rasio 80:20. Proses berikutnya adalah Data Mining, yaitu penerapan algoritma *Random Forest, Decision Tree, Support Vector Machine, MLP Classifier, Bagging Classifier, Gradient Boosting, CatBoost, XGBoost,* dan *LightGBM* untuk mengekstraksi pola. Tahap terakhir adalah Interpretasi dan Evaluasi.

2.2 Data Collection

Penelitian ini menggunakan dataset yang diperoleh dari *Kaggle* diupload oleh David Lapp pada tahun 2018 [11] dengan jumlah 1025 data. Kumpulan data ini berasal dari tahun 1988 dan terdiri dari empat basis data: Cleveland, Hungaria, Swiss, dan Long Beach. Atribut data dapat dilihat pada Tabel 1.

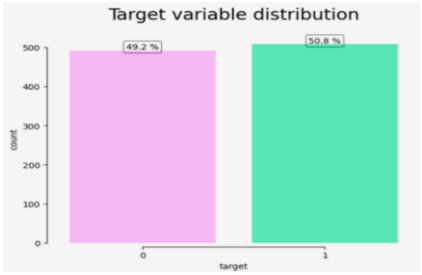
Tabel 1. Dataset Penyakit Jantung

Atribut	Keterangan	Nilai
Age	Usia (Pasien)	-
Sex	Jenis Kelamin Pasien	0 = Perempuan, 1 = Laki-laki
Cp	Jenis nyeri dada yang dialami pasien	0 = Typical angina, 1 = Atypical angina, 2 = Non-anginal pain, 3 = Asymptomatic
trestbps	Tekanan darah saat istirahat (mm Hg)	-
chol	Kadar kolesterol dalam darah (mg/dl)	-
fbs	Gula darah puasa lebih dari 120 mg/dl	0 = Tidak, 1 = Ya
restecg	Hasil elektrokardiografi saat istirahat	0 = Normal, 1 = Abnormalitas gelombang ST-T, 2 = Hipertrofi ventrikel kiri
thalach	Detak jantung maksimum yang dicapai saat latihan	-
exang	Apakah mengalami angina akibat olahraga	0 = Tidak, 1 = Ya
oldpeak	Depresi segmen ST akibat latihan dibanding kondisi istirahat	-
slope	Kemiringan segmen ST saat latihan	0 = Upsloping, 1 = Flat, 2 = Downsloping
ca	Jumlah pembuluh darah utama (0–3) yang terlihat melalui fluoroskopi	0, 1, 2, 3
thal	Jenis kelainan thalassemia	0 = Nilai error, 1 = Fixed defect, 2 = Normal, 3 = Reversible defect
Target	Status penyakit jantung	0 = Tidak memiliki penyakit jantung, 1 = Memiliki penyakit jantung

Berdasarkan Tabel 1, Dataset ini memiliki tipe data numerik dan kategorikal, terdiri dari 14 atribut yang mencakup informasi seperti usia, tekanan darah, kolesterol, serta detak jantung maksimal. Alasan penggabungan keempat dataset tersebut adalah untuk mendapatkan jumlah instans yang lebih besar sehingga model prediktif yang lebih stabil dapat diturunkan dengan teknik *ML*. Kumpulan data ini terdiri dari 14 atribut. Diantaranya yaitu usia, jenis kelamin, jenis nyeri dada yang dialami pasien, kadar kolesterol, gula darah, hasil eketrokardiografi saat istirahat, detak jantung maksimum, riwayat angina, depresi, jumlah pembuluh darah, jenis kelainan thalassemia, dan kolom "target" yang mengacu pada keberadaan penyakit jantung pada pasien. Kolom ini bernilai integer 0 = tidak ada penyakit dan 1 = penyakit.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

Penelitian ini tidak menerapkan teknik penanganan ketidakseimbangan data secara khusus seperti *oversampling (SMOTE), undersampling*, maupun pemberian bobot kelas (*class weighting*). Hal ini disebabkan oleh distribusi kelas pada variabel target yang relatif seimbang, sebagaimana ditunjukkan pada Gambar 2. Kondisi tersebut memungkinkan model machine learning dilatih secara langsung tanpa memerlukan metode penyeimbangan data tambahan serta dapat mengurangi potensi bias terhadap salah satu kelas. Selain itu, evaluasi performa model dilakukan menggunakan beberapa *metrik, yaitu accuracy, precision, recall, F1-score,* dan *Area Under the Curve (AUC).* Penggunaan berbagai metrik evaluasi ini bertujuan untuk memberikan penilaian performa model secara lebih komprehensif, khususnya dalam mengidentifikasi kesalahan klasifikasi yang relevan secara klinis.



Gambar 2. Distribusi Data

Gambar 2 menunjukkan hasil visualisasi distribusi kelas pada variabel target yang terdiri dari dua kategori, yaitu kelas 0 yang merepresentasikan pasien tanpa penyakit jantung dan kelas 1 yang merepresentasikan pasien dengan penyakit jantung. Berdasarkan visualisasi tersebut, dapat diketahui bahwa proporsi kedua kelas relatif seimbang, dengan 49,2% data berada pada kelas 0 dan 50,8% pada kelas 1.

2.3 Preprocessing

2.3.1 Pembersihan Data

Pembersihan data merupakan tahap penting dalam preprocessing yang bertujuan untuk menghilangkan noise, data yang tidak valid, serta inkonsistensi data sehingga dataset yang digunakan memiliki kualitas yang baik untuk proses analisis selanjutnya,

2.3.2. Transformasi Data

Transformasi data merupakan proses mengubah atau mengonversi data ke dalam format yang lebih sesuai untuk proses analisis atau pemodelan. Transformasi data dilakukan agar struktur data lebih konsisten dan dapat diproses oleh metode analisis atau algoritma tertentu.

2.3.3. Train-test-split

Train-Test Split merupakan proses pemisahan dataset menjadi dua subset, yaitu data latih (*training data*) dan data uji (*testing data*). Proses ini dilakukan dengan menggunakan fungsi t*rain_test_split()* dari *library scikit-learn*, yang berfungsi untuk membagi dataset ke dalam dua bagian, yaitu data fitur (x) dan label (y) dengan proporsi yang ditentukan.

2.4 Pembangunan Model

Pada tahap ini, dilakukan pembangunan model untuk mendeteksi penyakit jantung menggunakan berbagai algoritma *machine learning*. Algoritma yang digunakan meliputi *Random Forest, Decision Tree, Support Vector Machine, MLP Classifier, Bagging Classifier, Gradient Boost, CatBoost, XGBoost* dan *LightGBM*. Setiap algoritma dipilih karena memiliki keunggulan dan karakteristik yang berbeda dalam menangani masalah klasifikasi.

2.4.1 Model Ensemble

a. Random Forest
   Random Forest. Mengintegrasikan metodologi agregasi bootstrap dengan ide pohon keputusan, *Random Forest* adalah metode *Supervised Learning*. Ini merupakan bagian dari keluarga teknik *ensemble*, yang menggabungkan prediksi dari berbagai model dasar untuk menghasilkan estimasi akhir yang lebih tepat dan dapat diandalkan. Sekumpulan pohon keputusan mengembangkan metode ini. Subset acak dari fitur diterapkan oleh setiap pohon selama proses pelatihan dan dilatih pada sampel acak dari data pelatihan. Sejumlah pohon keputusan digabungkan dengan hutan acak guna membentuk model yang tangguh. Setiap pohon dibuat dengan menggunakan sampel bootstrap dari fitur yang dipilih secara acak. Prediksi untuk observasi baru dihasilkan dengan mengintegrasikan prediksi dari masing-masing pohon [12]
b. Bagging Classifier
   Bagging Classifier merupakan suatu teknik ensemble learning yang berakar dari gagasan *bootstrap aggregating*, di mana sekumpulan subset data (*bootstrap*) dipilih secara acak dari dataset asli untuk melatih beberapa model

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

dasar (*base learners*) secara terpisah, lalu hasil prediksi dari model-model tersebut digabungkan, umumnya dengan cara voting untuk klasifikasi atau dengan rata-rata untuk regresi. Metode ini berhasil dalam menurunkan varians model dan menghindari *overfitting*, khususnya pada algoritma yang peka terhadap variasi kecil dalam data pelatihan. *Bagging Classifier* dapat diterapkan untuk klasifikasi dan regresi, serta sangat sesuai untuk data yang memiliki tingkat variabilitas tinggi (noise), sehingga dapat meningkatkan stabilitas dan akurasi dalam prediksi [13]

c. Gradient Boost
Gradient Boosting merupakan teknik *ensemble learning* yang menciptakan model prediksi secara bertahap dengan mengintegrasikan beberapa model lemah (*weak learners*), umumnya *decision tree*, untuk menghasilkan model yang lebih kuat. Caranya dilakukan dengan membiarkan setiap model baru berupaya memperbaiki kesalahan (residual error) dari model sebelumnya melalui optimisasi fungsi loss dengan teknik gradient descent. Kelebihan *Gradient Boosting* terletak pada kemampuannya dalam menghasilkan model dengan presisi tinggi dan mengatasi data *non-linear* yang rumit, walaupun memerlukan penyesuaian parameter yang teliti agar tidak terjadi *overfitting* [14]

d. CatBoost
CatBoost merupakan algoritma *gradient boosting* yang dirancang untuk mengelola fitur kategorikal secara langsung tanpa memerlukan pengkodean manual. Melalui teknik *ordered boosting*, algoritma ini dapat menurunkan *overfitting* dan meningkatkan akurasi pada berbagai variasi data. *CatBoost* terbukti ampuh dalam penggunaan seperti prakiraan risiko kredit dan identifikasi penyakit kronis[15].

e. XGBoost
XGBoost adalah algoritma boosting yang sangat efektif dan efisien dalam mengatasi data dengan ketidakseimbangan kelas. *XGBoost* beroperasi dengan meningkatkan kinerja model secara bertahap melalui penambahan pohon keputusan baru yang menargetkan kesalahan prediksi dari pohon sebelumnya. Algoritma ini terkenal karena kecepatan dan kinerjanya yang sangat baik pada berbagai macam data [16].

f. Light GBM
LightGBM merupakan kerangka kerja gradient boosting sumber terbuka yang dikembangkan oleh Microsoft Research. Diciptakan untuk efisiensi maksimal dan penggunaan memori minimal pada dataset besar, *LightGBM* menggunakan strategi pertumbuhan pohon leaf-wise serta algoritma berbasis histogram, dan menerapkan dua teknik unggulan: *Gradient-based One-Side Sampling (GOSS)* dan *Exclusive Feature Bundling (EFB)* untuk mempercepat proses pelatihan tanpa mengorbankan akurasi prediksi [17].

2.4.2 Model Non-Ensemble

a. MLP Classifier
Pengklasifikasi *Multilayer Perceptron* (MLP) merupakan tipe jaringan saraf buatan yang terdiri dari lapisan input, satu atau lebih lapisan tersembunyi, dan lapisan output, di mana setiap lapisan saling terhubung sepenuhnya dengan bobot yang dapat dilatih. *MLP* beroperasi dengan menerapkan fungsi aktivasi non-linear seperti ReLU atau sigmoid di setiap neuron untuk merepresentasikan hubungan rumit antara input dan output. Algoritma ini menerapkan metode *backpropagation* untuk memperkecil kesalahan prediksi dengan melakukan pembaruan bobot secara iteratif. Sebagai salah satu algoritma pembelajaran terawasi, *MLP Classifier* dapat mengatasi permasalahan klasifikasi linear maupun non-linear dengan hasil yang baik pada data yang rumit [18].

b. Decision Tree
Decision Tree merupakan metode klasifikasi yang berlandaskan pada struktur pohon, di mana setiap lintasan dari akar mencerminkan rangkaian pemisahan data sampai mendapatkan hasil boolean di simpul daun. Pohon Keputusan adalah gambaran bertingkat dari hubungan pengetahuan yang terdiri dari node dan cabang [19].

c. Support Vector Machine
Support Vector Machine (SVM) merupakan model *machine learning* yang didasarkan pada teori statistik, digunakan untuk klasifikasi dan regresi, serta menunjukkan kinerja yang sangat baik baik untuk data linear maupun non-linear. Prinsip kerjanya adalah menemukan *hyperplane* terbaik yang memisahkan data ke dalam berbagai kelas dengan margin maksimum. *SVM* menggunakan trik kernel untuk mengubah data non-linear menjadi format yang bisa dipisahkan secara linear dalam ruang berdimensi lebih tinggi. Kelebihan utama *SVM* terletak pada kemampuannya untuk menangani data dengan dimensi tinggi serta pengendalian kompleksitas model, namun proses optimasinya membutuhkan penyelesaian Masalah Pemrograman Kuadratik (QPP) yang dapat mempengaruhi waktu komputasi [20].

2.5 Evaluasi Model

Evaluasi kinerja model dilaksanakan melalui teknik *K-fold cross-validation* dengan total fold sebanyak 5 pada data pelatihan. Dalam pendekatan ini, data latih dibagi menjadi lima segmen, di mana setiap iterasi memanfaatkan empat segmen sebagai data pelatihan dan satu segmen sebagai data validasi. Proses ini diulang sampai setiap komponen telah menjadi data validasi. Nilai akhir performa model didapatkan dari rata-rata hasil evaluasi di semua fold. Pendekatan ini dipilih karena dapat memberikan estimasi performa yang lebih konsisten dan mengurangi ketergantungan pada satu pembagian data tertentu, terutama pada dataset medis yang cenderung terbatas. Semua algoritma *machine learning* dilatih dengan parameter default yang disediakan oleh masing-masing pustaka. Pendekatan ini bertujuan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

untuk mendapatkan perbandingan kinerja yang adil di antara sembilan model yang diuji, terhindar dari bias yang disebabkan oleh perbedaan dalam pengoptimalan parameter. Penyesuaian parameter seperti *grid search* atau *random search* tidak dilakukan pada tahap ini, sehingga hasil yang didapatkan mencerminkan kinerja dasar (*baseline performance*) dari tiap algoritma. Pendekatan dasar ini juga mempermudah analisis perbandingan dan penafsiran hasil antar model. Semua eksperimen dalam penelitian ini dilaksanakan dengan menggunakan bahasa pemrograman *Python* serta memanfaatkan berbagai pustaka machine learning dan ilmu data, seperti *NumPy, Pandas, Scikit-learn, XGBoost, LightGBM*, dan *CatBoost*. Proses komputasi dilakukan di *Google Colab*, yang memungkinkan eksekusi kode secara interaktif dan dapat direproduksi. Selain itu, waktu komputasi untuk setiap model juga dicatat sebagai salah satu indikator efisiensi algoritma, sehingga tidak hanya performa prediksi yang dibandingkan, tetapi juga aspek kecepatan pemrosesan dalam lingkungan komputasi yang serupa

Setelah proses pelatihan model, tahap berikutnya adalah mengevaluasi untuk menilai kemampuan setiap model dalam mendeteksi penyakit jantung. Evaluasi dilakukan dengan memanfaatkan data uji yang tidak dikenal oleh model sebelumnya, sehingga dapat memberikan gambaran yang lebih tepat tentang kemampuan model dalam melakukan prediksi. Beberapa penilaian yang diterapkan dalam studi ini mencakup:

a. *Confusion matrix*: merupakan tabel yang memberikan gambaran menyeluruh tentang kinerja model klasifikasi. Tabel ini menampilkan data mengenai jumlah prediksi yang akurat dan tidak akurat yang dihasilkan oleh model, serta membandingkannya dengan label yang sebenarnya [21]. Dalam konteks ramalan penyakit jantung, matriks ini terdiri dari empat elemen utama:
   1. *True Positive* (TP): jumlah situasi di mana model dapat secara akurat memprediksi bahwa seseorang memiliki penyakit jantung dan itu terbukti benar.
   2. *True Negative* (TN): jumlah kejadian di mana model memperkirakan seseorang tidak menderita penyakit jantung dan prediksi itu akurat.
   3. *False Positive* (FP): jumlah situasi di mana model menilai seseorang menderita penyakit jantung sedangkan kenyataannya tidak (juga dikenal sebagai alarm palsu).
   4. *False Negative* (FN): total kasus di mana model memperkirakan seseorang tidak menderita penyakit jantung padahal orang tersebut sebenarnya mengalaminya
b. *Accuracy*: merupakan metrik dasar yang mengukur proporsi total taksiran yang benar (positif maupun negatif) dibandingkan dengan keseluruhan taksiran [22]. Dalam konteks prediksi penyakit jantung, akurasi mengindikasikan seberapa sering model mampu mengenali kondisi pasien dengan benar.
   Rumus akurasi adalah:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

   Walaupun akurasi memberikan gambaran umum tentang kinerja model, metrik ini bisa menyesatkan jika terdapat ketidakseimbangan kelas dalam data. Contohnya, jika mayoritas data terdiri dari pasien tanpa penyakit jantung, model dapat menunjukkan akurasi tinggi hanya dengan selalu memprediksi negatif, meskipun gagal mendeteksi pasien yang sebenarnya berisiko.
c. *Precision:*merupakan tolak ukur sejauh mana prediksi positif yang benar [22]. *Precision* digunakan untuk menilai berapa banyak dari prediksi positif (yaitu pasien yang diprediksi menderita penyakit jantung) yang sebenarnya positif. Presisi sangat krusial dalam bidang medis karena kita harus menghindari diagnosis *false positive*, yang dapat mengakibatkan stres, pemeriksaan tambahan yang tidak perlu, bahkan pengobatan yang tidak dibutuhkan.
   Rumus precision adalah:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

   Semakin besar nilai precision, semakin rendah kemungkinan model melakukan kesalahan dalam mendeteksi penyakit jantung pada individu yang sehat.
d. *Recall:* digunakan untuk mengukur kemampuan hasil analisis terhadap model saat mendeteksi seluruh kasus positif sebenarnya [21], Recall menilai sejauh mana model mampu mengidentifikasi seluruh kasus nyata penyakit jantung. Dalam bidang kedokteran, ukuran ini sangat penting karena kita tidak ingin mengabaikan pasien yang sebenarnya mengalami penyakit jantung.
   Rumus recall adalah:

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

   Recall yang tinggi menunjukkan bahwa model mampu mengenali hampir semua pasien yang sebenarnya menderita penyakit jantung, meskipun kemungkinan muncul lebih banyak false positive. Dalam situasi nyata, fokus utama adalah pada recall yang tinggi, karena hal ini mencegah kesalahan mengabaikan pasien yang sebenarnya memerlukan perhatian medis.
e. *F1-Score: F1-Score* merupakan rata-rata harmonis dari precision dan recall. Metrik ini diterapkan saat penting untuk mempertahankan keseimbangan antara kedua metrik itu, terutama dalam situasi di mana terdapat ketidakseimbangan antara kelas positif dan negatif [21].
   Rumus *F1-Score* adalah:

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

$$F1 = 2\ \frac{Precision\ .\ Recall}{Precision + Recall} \quad (4)$$

Dalam prediksi penyakit jantung, *F1-Score* sangat penting karena dapat menunjukkan kinerja model secara keseluruhan dalam mengatasi ketidakseimbangan data, serta mengurangi baik *false positives* maupun *false negatives.*

f. *ROC Curve*: merupakan grafik yang menunjukkan kinerja dari algoritma klasifikasi biner berdasarkan tingkat *true positive* dan *false positive.*

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pra-processing data

3.1.1 Pembersihan Data

Dalam proses pembersihan data, dilakukan penyaringan untuk menghilangkan nilai-nilai yang dianggap tidak valid pada dua atribut penting dalam dataset, yaitu *ca* (jumlah pembuluh darah utama yang tampak) dan *thal* (jenis kelainan thalassemia). Nilai atribut *ca* yang lebih besar atau sama dengan 4 dihapus karena tidak sesuai dengan dokumentasi resmi dataset Penyakit Jantung, di mana nilai yang valid seharusnya hanya berada antara 0 hingga 3. Sementara itu, pada atribut *thal,* data yang memiliki nilai 0 juga dihapus karena dianggap tidak mencerminkan kategori yang sah atau valid. Setelah proses pembersihan selesai, total data berkurang dari 1025 menjadi 1000, menyisakan hanya data yang valid dan siap untuk dianalisis lebih lanjut

3.1.2 Transformasi Data

Untuk meningkatkan keterbacaan dan mempermudah proses analisis data, dilakukan proses penyesuaian terhadap beberapa nama kolom serta nilai kategorikal pada dataset. Beberapa atribut awalnya menggunakan singkatan atau kode numerik yang kurang informatif sehingga berpotensi menyulitkan proses interpretasi data. Oleh karena itu, nama kolom diubah menjadi istilah yang lebih deskriptif, dan nilai kategorikal yang semula berbentuk kode angka dikonversi menjadi label teks yang lebih jelas. Rincian perubahan nama kolom dan nilai kategorikal pada dataset dapat dilihat pada Tabel 2.

Tabel 2. Transformasi Data

No	Nama Kolom Asli	Nama kolom baru	Contoh Perubahan Nilai
1	Cp	Chest_pain_type	0,1,2,3 → typical angina, atypical angina, dll
2	Fbs	Fasting_blood_sugar	0,1 → ≤ 120 mg/ml, > 120mg/ml
3	Thal	Thalassemia	3,6,7 → normal, fixed defect, reversible defect
4	Sex	Gender	0,1 → Wanita, pria
5	Restecg	Resting_electrocardiogram	0,1,2 → deskripsi hasil ECG
6	Exang	Exercise_indued_angina	0,1 → tidak, ya
7	Slope	St_slope	0,1,2 → unsloping, flat, downsloping

Hasil transformasi data dapat dilihat pada Tabel 2, menunjukkan bahwa dilakukan pengubahan nilai-nilai kategorikal yang semula berbentuk angka menjadi label teks yang lebih menjelaskan. Contohnya, pada kolom jenis kelamin, nilai 0 dan 1 diubah menjadi 'wanita' dan 'pria' untuk secara jelas menggambarkan gender. Dengan cara yang sama, nilai di kolom *chest_pain_type, fasting_blood_sugar, resting_electrocardiogram, exercise_induced_angina, st_slope,* dan *thalassemia* yang semula dalam bentuk kode numerik, diubah menjadi deskripsi yang lebih terang seperti *'typical angina'*, 'lebih dari 120mg/ml', atau '*fixed defect*'. Perubahan ini krusial untuk memperbaiki keterbacaan data serta mempermudah dalam analisis eksploratif dan penyampaian hasil kepada audiens non-teknis. Pada dataset ini, rata-rata usia pasien kira-kira 54,6 tahun, dengan usia paling muda 29 tahun dan paling tua 77 tahun. Untuk kadar kolesterol, rata-rata yang tercatat adalah 247 mg/dL, dengan nilai tertinggi mencapai 564 mg/dL dan terendah 126 mg/dL. Harus diperhatikan bahwa menurut standar medis, kadar kolesterol yang ideal adalah di bawah 200 mg/dL, dan kadar yang tinggi sering dihubungkan dengan risiko penyakit jantung. Tekanan darah pada saat istirahat (resting blood pressure) rata-ratanya adalah 131,6 mmHg, dengan nilai tertinggi 200 mmHg dan terendah 94 mmHg. Detak jantung maksimum yang dicapai pasien (max heart rate achieved) rata-ratanya 149 bpm, dengan nilai tertinggi 202 bpm dan terendah 71 bpm. Untuk fitur depresi ST (penurunan segmen ST saat stres), nilai rata-ratanya adalah 1,09, dengan nilai maksimum 6,2 dan minimum 0. Jumlah pembuluh darah utama (num_major_vessels) yang teramati melalui fluoroskopi berkisar antara 0 sampai 3, dengan rata-rata nilai 0,70.

3.1.3 Penanganan Data splitting

Dalam penelitian ini, dataset penyakit jantung dipisahkan menjadi data latih dan data uji dengan menggunakan metode *train–test split* yang membagi sebesar 80% untuk data latih dan 20% untuk data uji. Pembagian data dilaksanakan secara acak dengan nilai *random state* 42 untuk memastikan reprodusibilitas hasil percobaan. Data latih dipakai untuk

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

pelatihan model *machine learning* dan evaluasi dengan *cross-validation,* sementara data uji digunakan sebagai set pengujian *hold-out* untuk analisis lebih lanjut dengan menggunakan *confusion matrix.*

3.2 Perbandingan Model

Studi ini menganalisis kinerja sembilan model pembelajaran mesin dalam meramalkan penyakit jantung, yang meliputi model *ensemble* seperti *Random Forest, Bagging Classifier, XGBoost, CatBoost, Gradient Boosting, dan LightGBM,* dan model *non-ensemble* seperti *Decision Tree, Support Vector Machine (SVM),* serta *MLP Classifier*. Model dievaluasi dengan menggunakan berbagai metrik kinerja, yaitu Akurasi, Presisi, Recall, F1-Score, dan Luas Di Bawah Kurva (*AUC*). Untuk menjaga stabilitas dan konsistensi distribusi data selama pelatihan, penelitian ini menggunakan metode *K-Fold Cross-Validation*. Kumpulan data yang digunakan mencakup lebih dari 1000 data pasien yang selanjutnya dibagi menjadi data pelatihan dan data pengujian dengan metode pemisahan train–test yang memperhitungkan 80% untuk data pelatihan dan 20% untuk data pengujian.

Tabel 3. Komparasi Performa Model

	Model	Accuracy	Precision	Recall	F1-Score	AUC	Time(s)
0	Random Forest	0.97375	0.966001	0.9825	0.973951	0.997016	2.114225
1	Decision Tree	0.97125	0.965774	0.9775	0.971486	0.98820	0.114749
2	SVM	0.91250	0.902448	0.9250	0.913408	0.759906	0.727408
3	MLP Classifier	0.96625	0.960050	0.9750	0.966666	0.927969	13.617635
4	Bagging Classifier	0.96250	0.960301	0.9650	0.962427	0.988750	0.441632
5	Gradient Boosting	0.95625	0.941123	0.9750	0.957153	0.972062	1.639657
6	CatBoost	0.98000	0.973533	0.9875	0.980257	0.987844	6.710385
7	XGBoost	0.97625	0.968639	0.9850	0.976417	0.988656	0.324046
8	LightGBM	0.97500	0.963842	0.9875	0.975382	0.987156	0.374325

Hasil komparasi ke-9 model pada Tabel 3, menunjukkan secara keseluruhan bahwa semua model memperlihatkan kinerja yang cukup memuaskan dalam meramalkan penyakit jantung dengan tingkat akurasi melebihi 91%. Ini menunjukkan bahwa algoritma *machine learning* dapat secara efektif mempelajari pola hubungan antara fitur klinis pasien dan kemungkinan terjadinya penyakit jantung. Model berbasis ensemble, *seperti CatBoost, Random Forest, XGBoost,* dan *LightGBM*, menunjukkan kinerja yang lebih terjaga dan konsisten pada berbagai metrik evaluasi jika dibandingkan dengan model non-ensemble. Di antara semua model yang diuji, *CatBoost* menunjukkan kinerja terbaik, dengan akurasi mencapai 98%, *F1-Score* sebesar 0,980, dan *Recall* sebesar 0,9875. Nilai *recall* yang tinggi menandakan bahwa model ini sangat efektif dalam mendeteksi pasien yang benar-benar menderita penyakit jantung, sehingga dapat mengurangi kemungkinan kesalahan prediksi negatif dalam kasus medis.

Setelah melakukan klasifikasi dengan ke-9 model tersebut, evaluasi kinerja model dilakukan menggunakan *Confusion Matrix.* Evaluasi kinerja model dilakukan menggunakan Confusion Matrix untuk mengukur *True Positive(TP), True Negative(TN), False Positive(FP),* dan *False Negative(FN)* yang disajikan pada Tabel 4.

Tabel 4. Perbandingan Performa Confusion Matrix

	Random Forest	Decision Tree	SVM	MLP Classifier	Bagging Classifier	Gradient Boosting	Catboost	XGBoost	LightGBM
TP	92	92	64	70	89	89	92	92	92
TN	108	108	77	99	108	105	108	108	108
FP	0	0	28	22	3	3	0	0	0
FN	0	0	31	9	0	3	0	0	0

Dalam konteks prediksi penyakit jantung, nilai *False Negative (FN)* memiliki konsekuensi yang sangat penting secara klinis. *FN* mengindikasikan model salah memprediksi pasien yang tidak memiliki penyakit jantung, padahal faktanya pasien tersebut positif. Kesalahan ini berpotensi mengakibatkan keterlambatan dalam diagnosis, kurangnya terapi yang diperlukan, serta meningkatnya risiko komplikasi serius seperti serangan jantung atau kematian mendadak.

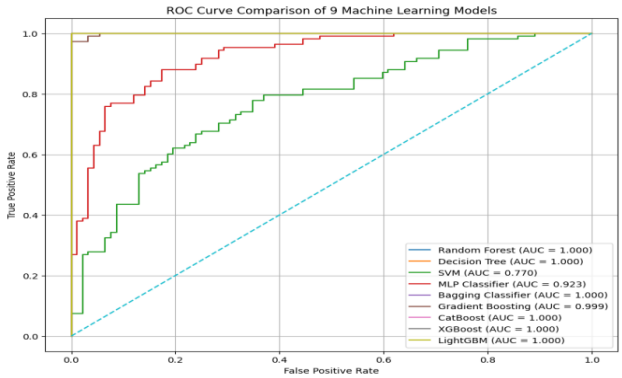
Tabel 4 menampilkan hasil dari confusion matrix untuk sembilan model, di mana model *ensemble* seperti *Random Forest, XGBoost*, dan *LightGBM* menunjukkan kinerja hampir sempurna, bahkan dalam *confusion matrix* menghasilkan nilai *FN* dan *FP* = 0. Sementara itu, model *SVM* dan *MLP* menunjukkan kinerja yang lebih rendah, terutama model *SVM* yang memiliki *FN* yang cukup tinggi. model dengan *FN* tinggi seperti *SVM* pada Tabel 4 memiliki tingkat risiko klinis lebih tinggi karena dapat berpotensi "mengabaikan" pasien yang membutuhkan penanganan segera. Sementara itu, *False Positive (FP)* muncul ketika model menganggap pasien memiliki penyakit jantung, padahal sebenarnya tidak. Dampaknya lebih bersifat psikologis dan ekonomi, seperti kecemasan pasien, pemeriksaan tambahan yang tidak diperlukan, peningkatan biaya medis, serta risiko *overtreatment*. Walaupun *FP* tidak seberbahaya *FN* dalam konteks penyakit berat, jumlah *FP* masih bisa membebani sistem layanan kesehatan.

Model *ensemble* seperti *Random Forest, XGBoost,* dan *LightGBM* yang menghasilkan *FP* = 0 dan *FN* = 0 menunjukkan performa klasifikasi yang sangat baik karena dapat menekan kedua jenis kesalahan tersebut secara bersamaan. Artinya, model ini tidak hanya valid secara statistik, tetapi juga lebih aman dalam konteks medis.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

Sebaliknya, model non-ensemble khususnya *SVM* dengan *FN* yang sangat tinggi tidak cocok digunakan dalam sistem dukungan keputusan medis karena dapat menimbulkan risiko *underdiagnosis.*

Analisis kemampuan diskriminasi model menunjukkan adanya variasi kinerja yang cukup jelas di antara sembilan algoritma machine learning yang diuji. Sebagian besar model berbasis ensemble, seperti *Random Forest, Decision Tree, Bagging Classifier, CatBoost, XGBoost,* dan *LightGBM,* menunjukkan performa yang sangat tinggi dengan nilai *AUC* mencapai 1,000. Kondisi tersebut menunjukkan bahwa model-model tersebut mampu membedakan secara hampir sempurna antara pasien yang memiliki dan yang tidak memiliki penyakit jantung pada berbagai nilai ambang keputusan, dengan sensitivitas yang sangat tinggi dan tingkat kesalahan klasifikasi yang sangat rendah. Dalam konteks klinis, karakteristik ini sangat penting karena memungkinkan identifikasi pasien berisiko tinggi secara lebih akurat tanpa meningkatkan jumlah false positive secara signifikan. Model Gradient Boosting juga memperlihatkan performa yang sangat baik dengan nilai *AUC* sebesar 0,999. Bentuk kurva *ROC* yang dihasilkan hampir serupa dengan model ensemble terbaik, meskipun masih terdapat sedikit penurunan sensitivitas pada beberapa rentang tingkat false positive. Hal ini menunjukkan bahwa kemampuan diskriminasi model tersebut tetap sangat kuat, namun masih terdapat kemungkinan kesalahan klasifikasi yang sedikit lebih tinggi dibandingkan model ensemble lain yang mencapai nilai *AUC* sempurna. Di sisi lain, model *MLP Classifier* dan *Support Vector Machine (SVM)* menunjukkan performa yang relatif lebih rendah dibandingkan model lainnya. *MLP Classifier* memperoleh nilai *AUC* sebesar 0,923, yang mengindikasikan kemampuan diskriminasi yang cukup baik tetapi belum mencapai tingkat optimal.

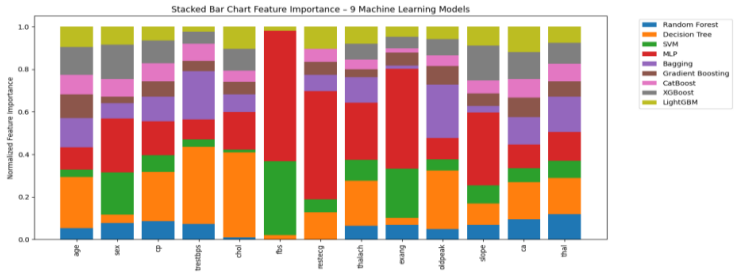


Gambar 3. Kurva ROC/AUC

Hasil komparasi ROC/AUC yang ditunjukkan pada Gambar 3 menunjukkan bahwa ROC pada model ini terlihat lebih landai pada bagian awal grafik, yang menunjukkan bahwa peningkatan sensitivitas cenderung diikuti oleh peningkatan tingkat false positive yang lebih besar. Sementara itu, *SVM* menjadi model dengan kinerja terendah dengan nilai *AUC* sebesar 0,770, di mana kurva *ROC* yang dihasilkan berada relatif dekat dengan garis diagonal. Kondisi tersebut menunjukkan bahwa kemampuan *SVM* dalam memisahkan kelas positif dan negatif pada dataset ini masih terbatas sehingga kurang ideal untuk diterapkan dalam konteks klinis yang menuntut tingkat sensitivitas yang tinggi.

Analisis terhadap sembilan model machine learning menunjukkan bahwa setiap algoritma memiliki pola yang berbeda dalam memberikan bobot atau penekanan terhadap fitur saat melakukan prediksi penyakit jantung. Secara umum, beberapa fitur klinis seperti cp (tipe nyeri dada*), oldpeak, thalach* (detak jantung maksimum), *exang* (angina yang dipicu oleh aktivitas fisik), dan *slope* muncul sebagai fitur yang paling konsisten memberikan kontribusi signifikan pada hampir seluruh model yang diuji. Hal ini menunjukkan bahwa indikator yang berhubungan langsung dengan respons jantung terhadap aktivitas fisik serta karakteristik nyeri dada memainkan peran penting dalam proses identifikasi atau klasifikasi penyakit jantung.

Selain itu, beberapa model memperlihatkan tingkat ketergantungan yang lebih tinggi pada fitur tertentu. Model *Decision Tree* dan *Bagging Classifier*, misalnya, cenderung menekankan sejumlah fitur spesifik seperti *cp, trestbps,* dan *chol,* yang terlihat dari dominasi segmen warna pada fitur-fitur tersebut dalam visualisasi. Pola ini mengindikasikan bahwa kedua model tersebut sangat memanfaatkan informasi dari fitur-fitur tersebut sebagai dasar utama dalam proses pengambilan keputusan prediksi, sebagaimana ditunjukkan pada Gambar 4.



Gambar 4. Stacked Bar Chart ke-9 model

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

Pada Gambar 4 terlihat jelas pola yang menunjukkan karakteristik *Decision Tree* yang cenderung memilih fitur dengan kemampuan pemisahan kelas terbaik pada tahap awal pembuatan pohon, sehingga kontribusi fitur menjadi lebih fokus. Sebaliknya, *Random Forest* menampilkan distribusi kontribusi fitur yang lebih merata, yang mencerminkan kemampuannya dalam mengurangi bias terhadap fitur tertentu melalui penggabungan dari berbagai pohon keputusan. Dalam kategori model boosting, seperti *Gradient Boosting, XGBoost,* dan *LightGBM,* terlihat fokus yang lebih jelas pada fitur-fitur kunci seperti *oldpeak, thalach, dan slope.* Ini mengindikasikan bahwa model boosting secara adaptif fokus pada pembelajaran dari fitur-fitur yang paling signifikan dalam mengurangi kesalahan klasifikasi secara iteratif. *XGBoost dan LightGBM* secara khusus memperlihatkan kontribusi signifikan terhadap fitur yang didasarkan pada ambang *(threshold-based features)*, yang menunjukkan kemampuan mereka untuk menangkap pola nonlinier dan interaksi kompleks antar fitur. Sementara itu, *CatBoost* menunjukkan distribusi kontribusi fitur yang relatif seimbang di antara fitur-fitur penting, tanpa adanya dominasi yang signifikan pada satu variabel tertentu. Pola ini menunjukkan kemampuan *CatBoost* dalam mengoptimalkan interaksi fitur secara lebih konsisten dan tangguh, khususnya pada data tabular yang memiliki karakteristik heterogen. Model *SVM dan MLP* juga menunjukkan kontribusi fitur yang tersebar, namun dengan pola y ang kurang tajam dibandingkan model ensemble, mencerminkan keterbatasan mereka dalam mengekstraksi struktur hierarkis dan interaksi fitur secara eksplisit. Secara keseluruhan, visualisas*i stacked bar chart* ini menegaskan bahwa model *ensemble* tidak hanya unggul dalam performa prediksi, tetapi juga lebih konsisten dalam mengidentifikasi fitur-fitur klinis yang relevan. Perbedaan distribusi feature importance antar model memberikan gambaran bahwa pendekatan *ensemble*, khususnya *boosting* dan *bagging,* lebih efektif dalam menangkap kompleksitas hubungan antar variabel klinis dibandingkan model tunggal atau berbasis margin seperti *SVM*.

4. KESIMPULAN

Penelitian ini melakukan analisis perbandingan sembilan algoritma machine learning untuk memprediksi penyakit jantung dengan memanfaatkan data klinis pasien. Temuan penelitian menunjukkan bahwa pendekatan *machine learning* memberikan performa klasifikasi yang sangat baik, di mana semua model meraih tingkat akurasi di atas 91%. Hasil penelitian ini menunjukkan bahwa hubungan antara karakteristik klinis dan kejadian penyakit jantung dapat dimodelkan dengan baik menggunakan algoritma *machine learning*. Secara keseluruhan, model *ensemble* menunjukkan kelebihan dibandingkan model *non-ensemble*. *CatBoost* menjadi model dengan kinerja paling baik dan konsisten berdasarkan akurasi, *recall, F1-score,* dan *AUC.* Model ensemble lainnya seperti *Random Forest, XGBoost,* dan *LightGBM* juga menunjukkan kinerja yang sangat baik, terutama dalam mengurangi kesalahan *false negative*. Penemuan ini krusial dari sudut pandang klinis, karena kesalahan *false negative* dapat mengakibatkan terlewatnya diagnosis pada pasien yang sebenarnya mengalami penyakit jantung. Perbandingan kontribusi fitur di antara model menunjukkan bahwa meskipun metode pembelajaran bervariasi, sebagian besar model *ensemble* menemukan pola klinis yang sama. *Random Forest* menitikberatkan pada stabilitas dan generalisasi, model *boosting* mengoptimalkan fitur-fitur paling diferensiatif, sementara *CatBoost* menawarkan keseimbangan terbaik antara kinerja, stabilitas, dan kemudahan interpretasi. Hasil ini menunjukkan bahwa keunggulan model tidak hanya bergantung pada kompleksitas algoritma, tetapi juga pada kemampuannya dalam merepresentasikan interaksi antara fitur yang relevan secara klinis. Berdasarkan hasil tersebut, studi ini merekomendasikan penerapan model *ensemble*, terutama *CatBoost*, sebagai metode yang paling potensial untuk pengembangan sistem pendukung keputusan medis dalam memprediksi penyakit jantung. Walau demikian, studi lebih lanjut sangat diperlukan untuk menguji penerapan model pada dataset yang lebih luas, beragam, dan berasal dari berbagai lembaga layanan kesehatan. Penelitian selanjutnya disarankan untuk menjelajahi teknik penanganan data yang tidak seimbang, melakukan penyesuaian parameter yang lebih mendalam, serta menggabungkan metode interpretabilitas canggih seperti *SHAP* secara menyeluruh untuk meningkatkan transparansi dan kepercayaan klinis terhadap hasil prediksi. Pendekatan ini diharapkan dapat memberikan kontribusi signifikan dalam peningkatan kualitas layanan kesehatan.

REFERENCES

[1] R. Luengo-Fernandez et al., “Cardiovascular disease burden due to productivity losses in European Society of Cardiology countries,” *Eur. Heart J. Qual. Care Clin. Outcomes*, vol. 10, no. 1, pp. 36–44, 2024, doi: 10.1093/ehjqcco/qcad031.

[2] G. E. Mandoli, L. Spaccaterra, E. Carluccio, and R. M. Inciardi, “Editorial: Methods in diagnosing heart failure,” *Front. Cardiovasc. Med.*, vol. 11, no. January, pp. 2–4, 2024, doi: 10.3389/fcvm.2024.1365006.

[3] L. Z. H. Jansen and K. E. Bennin, “A machine learning algorithm for personalized healthy and sustainable grocery product recommendations,” *International Journal of Information Management Data Insights,* vol. 5, no. 1, p. 100303, 2025, doi: 10.1016/j.jjimei.2024.100303.

[4] F. Alqurashi and I. Ahmad, “A data-driven multi-perspective approach to cybersecurity knowledge discovery through topic modelling,” *Alexandria Engineering Journal,* vol. 107, no. June, pp. 374–389, 2024, doi: 10.1016/j.aej.2024.07.044.

[5] H. Amadou Boubacar et al., “HeartPredict algorithm: Machine intelligence for the early detection of heart failure,” *Intell. Based. Med.,* vol. 5, p. 100044, 2021, doi: 10.1016/j.ibmed.2021.100044.

[6] D. Amanda Ardhani and K. D. Tania, “Knowledge Discovery on E-Commerce Customer Churn Using Interpretable Machine Learning: A Comparative Study of SHAP-Based Classifiers,” *Journal of Applied Informatics and Computing,* vol. 9, no. 5, pp. 2695–2702, 2025, doi: 10.30871/jaic.v9i5.10811.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2618−2628
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8346

[7] C. Andini Bahri and K. Ditha Tania, "Perbandingan Kinerja LSTM, Random Forest, dan SVR Berbasis Knowledge Discovery untuk Prediksi Harga Beras Sumatera Selatan," Jurnal Riset Komputer), vol. 12, no. 5, pp. 2407–389, 2025, doi: 10.30865/jurikom.v12i5.9140.

[8] A. Davinka, S. Depari, K. D. Tania, and P. E. Sevtiyuni, "Penerapan Metode Machine Learning Dan Teknik SMOTE untuk Prediksi Diabetes," vol. 7, pp. 436–447, 2025, doi: 10.30865/json.v7i2.9032.

[9] Muhammad Raviansyah, Andika Amansyah, Farhan Fadhilah, Sumanto Sumanto, Imam Budiawan, and Roida Pakpahan, "Komparasi Algoritma Machine Learning (Random Forest, Gradient Boosting, dan Ada Boosting) untuk Prediksi Tingkat Penyakit Alzheimer," *Jurnal Teknik Informatika dan Teknologi Informasi,* vol. 5, no. 3, pp. 131–145, 2025, doi: 10.55606/jutiti.v5i3.6227.

[10] N. H. Alfajr, G. Garno, and D. Yusup, "Studi Komparasi Algoritma Random Forest Classifier Dan Support Vector Machine Dalam Prediksi Penyakit Jantung," *Jurnal Informatika dan Teknik Elektro Terapan,* vol. 13, no. 3, pp. 22–30, 2025, doi: 10.23960/jitet.v13i3.6569.

[11] D. Lapp, "Heart Disease Dataset," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[12] E. Naroum et al., "Comparative analysis of deep learning and machine learning techniques for forecasting new malaria cases in Cameroon's Adamaoua region," *Intell. Based. Med.*, vol. 11, no. February, p. 100220, 2025, doi: 10.1016/j.ibmed.2025.100220.

[13] A. Alzayed, W. Almayyan, and A. Al-Hunaiyyan, "Diagnosis of Obesity Level based on Bagging Ensemble Classifier and Feature Selection Methods," *International Journal of Artificial Intelligence & Applications,* vol. 13, no. 02, pp. 37–54, 2022, doi: 10.5121/ijaia.2022.13203.

[14] Ö. Bezek Güre, "Comparison of the Performance of Gradient Boosting and Extreme Gradient Boosting Methods in Classifying Timms Science Achievement," *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 14, no. 2, pp. 1041–1059, 2025, doi: 10.17798/bitlisfen.1636812.

[15] H. Wang and L. Cheng, "CatBoost model with synthetic features in application to loan risk assessment of small businesses," 2021, [Online]. Available: http://arxiv.org/abs/2106.07954

[16] A. Izotova and A. Valiullin, "Comparison of Poisson process and machine learning algorithms approach for credit card fraud detection," *Procedia Comput.* Sci., vol. 186, pp. 721–726, 2021, doi: 10.1016/j.procs.2021.04.214.

[17] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.

[18] R. Atangana, D. Tchiotsop, G. Kenne, and L. C. DjoufackNkengfac k, "EEG Signal Classification using LDA and MLP Classifier," *Health Informatics - An International Journal,* vol. 9, no. 1, pp. 14–32, 2020, doi: 10.5121/hiij.2020.9102.

[19] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[20] S. Acharya, T. Kar, U. C. Samal, and P. K. Patra, "Performance Comparison between SVM and LS-SVM for Rice Leaf Disease detection," *EAI Endorsed Transactions on Scalable Information Systems,* vol. 10, no. 6, pp. 1–7, 2023, doi: 10.4108/eetsis.3940.

[21] X. Deng, H. Shao, L. Shi, X. Wang, and T. Xie, "A classification–detection approach of COVID-19 based on chest X-ray and CT by using keras pre-trained deep learning models," *CMES - Computer Modeling in Engineering and Sciences*, vol. 125, no. 2, pp. 579–596, 2020, doi: 10.32604/cmes.2020.011920.

[22] Muhamad Fadli and Rizal Adi Saputra, "Klasifikasi dan Evaluasi Performa Model Random Forest untuk Prediksi Stroke," JT: Jurnal Teknik, vol. 12, no. 2, pp. 72–80, 2023, [Online]. Available: http://jurnal.umt.ac.id/index.php/jt/index