Prosiding Annual Research Seminar 2025
Computer Science and ICT
ISBN :
Vol.5 No.1

KLASIFIKASI PENYAKIT JANTUNG BERBASIS SELEKSI FITUR *CHI-SQUARE* DAN *MACHINE LEARNING*

Fadillah Alfany Adhelia
Informatics Engineering
Faculty of Computer
Sriwijaya University
South Sumatera, Indonesia
adheliafadillahalfany@gmail.com

Annisa Darmawahyuni
Inteligent System Research Group
Faculty of Computer
Sriwijaya University
South Sumatera, Indonesia
riset.annisadarmawahyuni@gmail.com

Abstrak* — Penyakit jantung merupakan salah satu penyebab utama kematian di dunia, termasuk di Indonesia. Deteksi dini sangat penting untuk mencegah komplikasi yang lebih serius. Dalam penelitian ini, dilakukan klasifikasi penyakit jantung berdasarkan data klinis pasien menggunakan tiga model machine learning, yaitu K-Nearest Neighbors (KNN), Support Vector Machine (SVM), dan Random Forest. Dataset yang digunakan adalah heart.csv dari platform Kaggle, yang terdiri dari 918 data pasien dengan 11 fitur dan 1 target. Untuk mengatasi ketidakseimbangan data, digunakan metode Synthetic Minority Oversampling Technique (SMOTE), serta seleksi fitur dilakukan dengan metode SelectKBest berbasis chi-square. Evaluasi model dilakukan menggunakan metrik akurasi, precision, recall, f1-score, dan confusion matrix pada data validasi dan testing. Hasil penelitian menunjukkan bahwa pendekatan machine learning dapat digunakan secara efektif untuk membantu deteksi dini penyakit jantung dan mendukung pengambilan keputusan dalam bidang medis.

***Kata kunci* — Penyakit Jantung, Klasifikasi, Machine Learning, SMOTE, SVM, KNN, Random Forest.**

I. PENDAHULUAN

Jantung merupakan organ vital yang berperan penting dalam sirkulasi darah ke seluruh tubuh. Ketika kemampuan jantung dalam memompa darah yang mengandung oksigen menurun, tubuh tidak dapat memenuhi kebutuhan metaboliknya secara optimal, sehingga dapat menimbulkan kondisi medis yang dikenal sebagai gagal jantung [1]. Penyakit ini menjadi salah satu penyebab utama kematian di dunia dan berkontribusi signifikan terhadap peningkatan angka rawat inap, yang mencapai sekitar 6,5 juta kasus setiap tahun [2]. Kondisi tersebut menunjukkan perlunya upaya diagnostik yang lebih akurat dan efisien guna mendeteksi penyakit jantung sejak dini.

Data dari World Health Organization (WHO) dan World Heart Federation (WHF) memperkirakan bahwa pada tahun 2025, penyakit jantung akan menjadi penyebab utama kematian di negara-negara kawasan Asia. Sekitar 78% dari total kematian global akibat penyakit jantung terjadi pada kelompok masyarakat berpenghasilan rendah dan menengah. Selain itu, di negara berkembang, angka kematian akibat penyakit jantung menunjukkan peningkatan signifikan sebesar 120% pada perempuan dan 137% pada laki-laki selama periode 1990–2020. Sebaliknya, di negara maju, peningkatan tersebut relatif lebih rendah, yakni sebesar 29% pada perempuan dan 48% pada laki-laki [3]. Data ini mengindikasikan adanya kesenjangan dalam pencegahan dan penanganan penyakit jantung antara negara maju dan negara berkembang.

Hasil Riset Kesehatan Dasar (Riskesdas) yang diterbitkan oleh Kementerian Kesehatan Republik Indonesia tahun 2018 menunjukkan bahwa prevalensi gagal jantung di Indonesia yang terdiagnosis oleh tenaga medis mencapai 5%. Kasus gagal jantung lebih banyak dialami oleh laki-laki 66% dibandingkan perempuan 34%. Selain itu, tingkat kematian pasien gagal jantung selama masa perawatan di rumah sakit mencapai 17,2%, dengan angka kematian dalam satu tahun setelah perawatan sebesar 11,3%. Sementara itu, pasien yang mengalami rehospitalisasi atau perawatan ulang di rumah sakit tercatat sebesar 17% [4]. Angka tersebut menegaskan bahwa gagal jantung masih menjadi tantangan besar dalam sistem pelayanan kesehatan di Indonesia.

Kemajuan teknologi telah memberikan kontribusi signifikan dalam peningkatan kualitas layanan kesehatan, khususnya melalui penerapan metode berbasis data. Salah satu penerapan yang relevan adalah penggunaan pembelajaran mesin (*machine learning*) untuk melakukan klasifikasi dan deteksi penyakit. Pembelajaran mesin, yang merupakan cabang dari kecerdasan buatan (*artificial intelligence*), memungkinkan sistem komputer untuk belajar secara otomatis dari data guna mengenali pola dan menghasilkan prediksi yang akurat [15]. Pendekatan ini menawarkan potensi besar dalam membantu tenaga medis melakukan diagnosis lebih cepat dan efisien.

Penelitian ini menggunakan tiga model pembelajaran mesin, yaitu *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), dan *Random Forest*. Ketiga model tersebut dipilih karena mewakili pendekatan yang berbeda dalam pembelajaran mesin, mulai dari metode berbasis instance, berbasis margin, hingga metode ensemble yang

menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi. Dataset yang digunakan diperoleh dari platform Kaggle dengan nama berkas *heart.csv*, yang berisi 918 data pasien dan 11 atribut klinis, seperti usia, jenis kelamin, tekanan darah saat istirahat, kadar kolesterol, denyut jantung maksimum, *oldpeak*, jenis nyeri dada, serta satu atribut target berupa status penyakit jantung (positif atau negatif). Dataset tersebut menjadi dasar dalam proses pelatihan dan pengujian model klasifikasi.

Tujuan utama penelitian ini adalah memprediksi keberadaan penyakit jantung berdasarkan data klinis pasien serta mengevaluasi performa ketiga model klasifikasi yang digunakan. Untuk meningkatkan keakuratan model, diterapkan pula SMOTE *(Synthetic Minority Oversampling Technique)* guna mengatasi ketidakseimbangan data, serta seleksi fitur menggunakan metode *chi-square (SelectKBest)* untuk memperoleh atribut yang paling berpengaruh terhadap hasil klasifikasi. Evaluasi performa dilakukan menggunakan metrik akurasi, *precision, recall, f1-score*, serta visualisasi *confusion matrix* pada data validasi dan pengujian.

II. KAJIAN LITERATUR

Gagal jantung merupakan gangguan multisistem yang terjadi ketika jantung mengalami disfungsi, baik pada fungsi sistolik maupun diastolik. Disfungsi sistolik menyebabkan penurunan curah jantung pada ventrikel kiri, sedangkan disfungsi diastolik menurunkan komplians ventrikel kanan akibat gangguan relaksasi miokard [5][16].

Perkembangan teknologi telah memberikan kontribusi signifikan dalam penyelesaian berbagai permasalahan di berbagai bidang, terutama di bidang kesehatan. Salah satu penerapannya adalah dalam proses klasifikasi dan deteksi penyakit menggunakan metode *machine learning* [1]. *Machine learning* atau pembelajaran mesin merupakan cabang dari kecerdasan buatan (*artificial intelligence*) yang berfokus pada bagaimana komputer dapat belajar dari data untuk meningkatkan kemampuannya dalam membuat keputusan [6]. Dalam bidang medis, *machine learning* berperan penting dalam mendukung proses diagnostik dan prognostik, serta mampu menganalisis data medis untuk mendeteksi pola, menghapus data tidak valid, menjelaskan informasi dari perangkat medis, dan memantau kondisi pasien secara efektif. Selain itu, penerapan *machine learning* juga mampu meningkatkan efisiensi dalam manajemen pasien maupun kegiatan administrasi di sektor kesehatan [17].

Beberapa metode yang umum digunakan dalam *machine learning* adalah *K-Nearest Neighbors, Decision Tree, Random Forest, Naive Bayes, dan Support Vector Machine* [1]. Untuk mendeteksi awal kelangsungan penyakit gagal jantung, penelitian ini menggunakan beberapa model seperti *K-Nearest Neighbors, Random Forest dan Support Vector Machine.*

Support Vector Machine (SVM) bekerja dengan mencari fungsi pemisah terbaik (*optimal separating hyperplane*) untuk memisahkan data ke dalam dua kelas berbeda dan termasuk dalam pendekatan *supervised learning*. Tujuan utama SVM adalah menemukan *hyperplane* terbaik yang memaksimalkan margin antara dua kelas data. Berbeda dengan *Artificial Neural Network* (ANN) yang mempelajari seluruh data latih, SVM hanya menggunakan sebagian data yang disebut *support vectors* dalam proses pembentukan model klasifikasi [18]. Akurasi model SVM sangat bergantung pada fungsi *kernel* dan parameter yang digunakan [19].

Algoritma *K-Nearest Neighbors* (KNN) merupakan salah satu metode *instance-based learning* yang termasuk ke dalam kategori *lazy learning* karena tidak melakukan proses pelatihan eksplisit sebelum klasifikasi [20]. KNN banyak digunakan dalam tugas klasifikasi dan regresi, terutama untuk pengenalan pola dan konsistensi data [24]. Prinsip kerja KNN adalah mengklasifikasikan data baru berdasarkan jarak terhadap sejumlah *k* tetangga terdekat dari data pelatihan yang telah ada [7].

Random Forest (RF) merupakan metode *ensemble learning* yang menggabungkan sejumlah pohon keputusan (*decision trees*) untuk meningkatkan akurasi klasifikasi. Algoritma RF dikembangkan oleh Breiman pada tahun 2001 dan dapat digunakan untuk menyelesaikan masalah klasifikasi maupun regresi. RF dibangun dari metode *Classification and Regression Tree* (CART) dengan pendekatan *bagging* (*bootstrap aggregation*) dan *random feature selection* [5][8][9]. Keputusan akhir diambil melalui proses *voting* dari setiap pohon, di mana kelas dengan suara terbanyak menjadi hasil prediksi akhir. Teknik *bagging* ini mampu meningkatkan performa algoritma klasifikasi dengan mengurangi variansi model.

Selain itu, *Synthetic Minority Over-sampling Technique* (SMOTE) yang diperkenalkan oleh Nithes V. Chawla (2002) digunakan untuk mengatasi ketidakseimbangan data antara kelas mayoritas dan minoritas. SMOTE bekerja dengan menambahkan sampel sintetis pada kelas minoritas untuk menyeimbangkan distribusi data. Pendekatan ini bertujuan meningkatkan kinerja model klasifikasi, terutama dalam mendeteksi kelas dengan jumlah data yang sedikit [10].

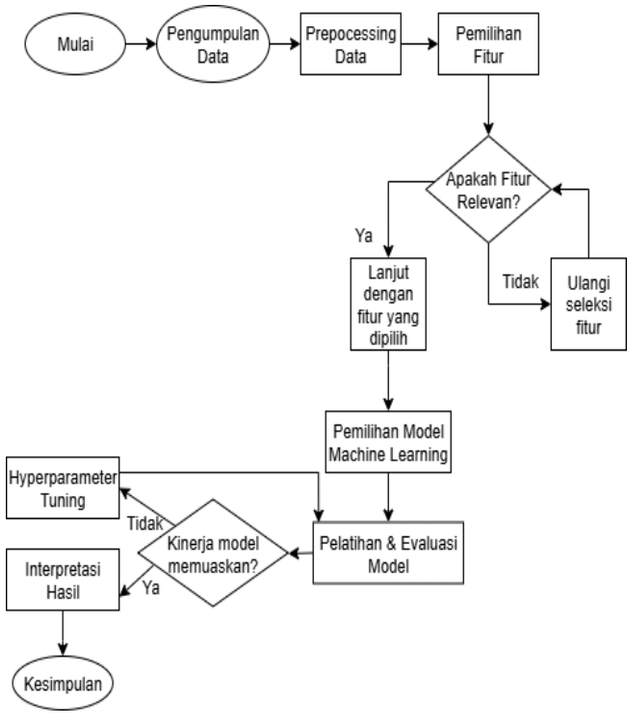
III. METODOLOGI PENELITIAN

A. PERSIAPAN DATASET

Penelitian ini menggunakan dataset yang diambil dari platform Kaggle dengan nama file heart.csv.

TABEL 1. Deskripsi dari atribut dalam dataset

No.	Atribut	Tipe	Peran
1.	*Age*	*int64*	numerik
2.	*Sex*	*object*	kategorikal
3.	*ChestPainType*	*object*	kategorikal
4.	*RestingBP*	*int64*	numerik
5.	*Cholesterol*	*int64*	numerik
6.	*FastingBP*	*int64*	biner
7.	*RestingECG*	*object*	kategorikal
8.	*MaxHR*	*int64*	numerik
9.	*ExerciseAngina*	*object*	biner
10.	*Oldpeak*	*float64*	numerik
11.	*ST_Slope*	*object*	kategorikal
12.	***HeartDisease***	***int64***	**label/target**



Gambar 1. Diagram Skema Metodologi Penelitian

B. PRA-PEMROSESAN DATA

Tahapan pra-pemrosesan data merupakan langkah awal yang krusial dalam proses klasifikasi, bertujuan untuk memastikan bahwa data dalam kondisi siap dan optimal sebelum digunakan pada pemodelan *machine learning*. Dataset yang digunakan dalam penelitian ini adalah *heart.csv* yang terdiri dari 918 data pasien dan 12 atribut, termasuk atribut numerik dan kategorik, serta satu kolom target yaitu *HeartDisease* 0 = tidak mengidap penyakit jantung, 1 = mengidap penyakit jantung.

Langkah awal dimulai dengan pembersihan data yang mencakup penghapusan duplikasi dan pengisian nilai kosong dengan median dari kolom numerik. Data kategorik seperti *Sex, ChestPainType, RestingECG, ExerciseAngina,* dan *ST_Slope* dikonversi menjadi bentuk numerik menggunakan *Label Encoding* agar dapat diproses oleh algoritma pembelajaran mesin.

Selanjutnya, penanganan outlier dilakukan dengan metode *Interquartile Range* (IQR). Setiap fitur numerik seperti *Age, RestingBP, Cholesterol, MaxHR, dan Oldpeak* dicapping dengan batas bawah Q1 − 1.5× IQR dan batas atas Q3 + 1.5× IQR agar nilai ekstrim tidak mendistorsi hasil pelatihan model.

Setelah itu, dilakukan normalisasi pada fitur numerik menggunakan metode *Min-Max Scaling*. Tujuan dari normalisasi ini adalah untuk menyamakan skala seluruh fitur ke dalam rentang antara 0 hingga 1. Dengan skala yang seragam, tidak ada fitur yang memiliki nilai terlalu besar yang dapat mendominasi fitur lainnya, terutama pada algoritma yang sensitif terhadap jarak seperti *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM). Proses normalisasi ini membantu meningkatkan stabilitas dan keakuratan model dalam melakukan klasifikasi. Dengan menggunakan rumus normalisasi *Min-Max* sebagai berikut:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Di tahap selanjutnya, untuk mengatasi masalah ketidakseimbangan kelas dalam data pelatihan, digunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). *Synthetic Minority Oversampling Technique* (SMOTE) adalah salah satu teknik *oversampling*, digunakan untuk menambah data minoritas pada dataset yang tidak seimbang [22]. SMOTE bekerja untuk menghasilkan sampel sintesis baru pada data minoritas. Langkah kerja SMOTE dimulai dengan mengelompokkan data berdasarkan tetangga terdekat kemudian mencari jarak terdekat antara dua data yaitu data yang akan dievaluasi dengan banyaknya tetangga terdekat (k) pada data training (latih) menggunakan jarak *Euclidean* [11]. Dengan menggunakan rumus sebagai berikut:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2} \quad (2)$$

Keterangan :
d (x,y) = jarak *Euclidean*
xi dan *yi* = koordinat x dan y pada dimensi ke-i
n = jumlah dimensi *Euclidean*

C. SELEKSI FITUR

Seleksi fitur merupakan tahap penting untuk memastikan bahwa hanya atribut yang paling relevan digunakan dalam proses klasifikasi. Penelitian ini menggunakan metode SelectKBest dengan uji Chi-Square **(χ^2)** untuk menilai tingkat keterkaitan masing-masing fitur terhadap variabel target (*HeartDisease).*

Berdasarkan hasil perhitungan skor chi-square, diperoleh 9 fitur terbaik yang dipilih untuk pelatihan model, yaitu: *Age*, *Sex*, *ChestPainType*, *Cholesterol*, *RestingECG*, *MaxHR*, *ExerciseAngina*, *Oldpeak*, *dan ST_Slope.* Fitur-fitur ini memiliki relevansi klinis sebagai berikut:

1. *Age*: usia merupakan faktor risiko utama penyakit jantung, semakin tua usia semakin tinggi risiko.
2. *Sex*: pria lebih berisiko mengalami penyakit jantung dibandingkan wanita pada usia lebih muda, sedangkan wanita meningkat risikonya setelah menopause.
3. *ChestPainType*: jenis nyeri dada adalah gejala klasik dari penyakit arteri koroner.
4. *Cholesterol*: kadar kolesterol tinggi dapat menyebabkan penyumbatan pembuluh darah (aterosklerosis).
5. *RestingECG*: hasil EKG saat istirahat dapat menunjukkan kelainan irama jantung atau tanda iskemia.
6. *MaxHR:* denyut jantung maksimum saat aktivitas menunjukkan kapasitas fungsional jantung.

7. *ExerciseAngina*: angina (nyeri dada) yang muncul saat olahraga adalah tanda klasik adanya gangguan aliran darah ke jantung.
8. *Oldpeak:* depresi segmen ST pada EKG yang menunjukkan tingkat keparahan iskemia.
9. *ST_Slope*: pola kemiringan segmen ST yang sangat penting dalam diagnosis penyakit jantung iskemik.

Dengan menggunakan fitur-fitur terpilih ini, model diharapkan menjadi lebih efisien, mengurangi risiko *overfitting*, dan fokus pada atribut yang secara klinis memang berpengaruh terhadap diagnosis penyakit jantung.

D. ALGORITMA KLASIFIKASI

Setelah tahapan pra-pemrosesan data dilakukan, tahap selanjutnya dalam penelitian ini adalah membangun model klasifikasi. Penelitian ini menerapkan proses klasifikasi menggunakan empat algoritma pembelajaran mesin yang diterapkan dalam analisis data kesehatan, yakni *Random Forest* (RF), *Support Vector Machine* (SVM), dan *K-Nearest Neighbor* (KNN).

1. *Random Forest*

Pada proses perancangan model algoritma *Random Forest*, model algoritma *Random Forest* yang dibangun dengan menggunakan 'n' pohon *Decision Tree* dengan mencari nilai 'n' terbaik [12]. RF adalah metode untuk mengklasifikasikan sejumlah besar data. Algoritma ini memiliki kemampuan untuk menangani dua jenis masalah, yakni klasifikasi dan regresi [23]. Klasifikasi *random forest* dilakukan dengan menggabungkan pohon dengan pelatihan pada data sampel. Cara menghitung entropy untuk menentukan tingkat polusi atribut dan nilai perolehan informasi adalah dengan dimulainya pohon keputusan [5].

$$Entropy\ (Y) = -\sum_i p\ (c|Y) log^2 p\ (c|Y) \quad (3)$$

Keterangan:
Y = Himpunan kasus
$p(c|Y)$ = Proporsi nilai Y terhadap kelas c

2. *Support Vector Machine*

SVM adalah metode pembelajaran terawasi untuk menganalisis data dan mengidentifikasi pola yang digunakan dalam klasifikasi. SVM memiliki keuntungan untuk dapat mendefinisikan superclass yang berbeda yang memaksimalkan margin antara dua kelas yang berbeda [19] [25] [26]. Metode ini menggunakan pemetaan *nonlinier* untuk mengubah data pelatihan asli ke dimensi yang lebih tinggi. Dalam dimensi baru ini akan dicari optimasi *linier* yang memisahkan dua kelas target dengan *hyperplane. Hyperplane* adalah batas keputusan yang memisahkan tipe dari satu kelas dengan kelas yang lain. SVM menemukan *hyperplane* menggunakan vektor pendukung dan margin [5] [13].

$$Ef\ (x) = w^t \Phi\ (x) + b \quad (4)$$

Keterangan:
b = Bias
$x = (x_1, x_2, \dots, x_D)^T$ = Variabel input
$w = (w_0, w_1, \dots, w_D)^T$ = Parameter bobot
$\Phi\ (x)$ = Fungsi transformasi fitur

3. *K-Nearest Neighbors* (KNN)

KNN adalah metode yang menggunakan klasifikasi tetangga (neighbors) sebagai prediktor untuk instance query yang baru. Proses klasifikasi pada metode ini akan terjadi didasarkan pada kesamaan antara data yang baru dan data sampel yang meningkat [5] [8]. Jarak antara data terbaru dan data yang sudah lama dihitung dengan rumus sebagai berikut:

$$d = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \quad (5)$$

Keterangan:
$X1$ = Data sampel
$X2$ = Data *training* atau *testing*
i = Variabel data
d = Jarak
p = Dimensi data

E. METRIK EVALUASI

Untuk mengukur performa model klasifikasi penyakit hepatitis yang dikembangkan, digunakan beberapa metrik evaluasi yang umum digunakan dalam studi klasifikasi medis, yaitu *akurasi, presisi, recall, Spesifitas*, dan *F1-score*:

- Akurasi

Akurasi Berfungsi untuk mengukur proporsi prediksi yang benar dari total prediksi [2] [8].

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

- Presisi

Presisi Berguna dalam mengukur seberapa tepat model dalam memprediksi kelas positif [6][8].

$$Presisi = \frac{TP}{TP + FP} \quad (7)$$

- Recall

Recall digunakan untuk mengukur kemampuan model dalam mendeteksi data yang benar-benar positif [5] [6].

$$Sensitivitas = \frac{TP}{TP + FN} \quad (8)$$

- F1-score

F1-score merupakan rata-rata harmonic dari presisi dan recall, sangat berguna ketika data tidak seimbang [2][13].

$$F1 - Score = 2 \times \frac{Presisi \times Sensitivitas}{Presisi + Sensitivitas} \quad (9)$$

IV. HASIL PENELITIAN

A. HASIL PENELITIAN

Penelitian ini bertujuan untuk mengembangkan model klasifikasi penyakit jantung secara otomatis dan dini menggunakan pendekatan *machine learning* berbasis data klinis, dengan menerapkan teknik pengolahan data yang tepat dan pemilihan algoritma yang sesuai. Proses pelatihan dimulai dengan tahap pra-pemrosesan, yaitu penghapusan duplikat, penanganan nilai hilang, *encoding* data kategorikal menggunakan *LabelEncoder, capping outlier* dengan metode IQR, dan normalisasi fitur numerik menggunakan *Min-Max Scaling*. Selain itu, ketidakseimbangan kelas dalam data pelatihan diatasi menggunakan metode SMOTE *(Synthetic Minority Over-sampling Technique)*.

Data dibagi menjadi tiga bagian menggunakan metode *stratified split*: 80% untuk pelatihan, 10% untuk validasi, dan 10% untuk pengujian. Seleksi fitur dilakukan menggunakan metode SelectKBest dengan skor chi-square, yang menghasilkan 9 fitur terbaik yaitu *Age, Sex, ChestPainType, Cholesterol, RestingECG, MaxHR, ExerciseAngina, Oldpeak,* dan *ST_Slope*. Fitur-fitur ini dianggap penting karena memiliki keterkaitan klinis dengan risiko penyakit jantung. Model pelatihan dilakukan terhadap dua kondisi data: dengan SMOTE dan tanpa SMOTE. Tiga algoritma klasifikasi yang digunakan yaitu *Random Forest* (RF), *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN).

B. HASIL EVALUASI

Evaluasi model dilakukan untuk mengukur kinerja model pada data validasi dan data uji menggunakan metrik evaluasi: *accuracy*, *precision, recall, dan f1-score.* Evaluasi dilakukan secara terpisah untuk model yang dilatih dengan SMOTE.

Model	Parameter	Precision 0	Precision 1	Akurasi
SVM	C = 10 gamma=”auto” kernel=’rbf’ random_state = 42	0.8571	0.9000	88.04%
Random Forest	N = 300	0.9487	0.9245	93.48%
KNN	k = 5	0.7755	0.9302	84.78%

TABEL 2. Hasil Evaluasi Akurasi dan *Precision* Model *Machine Learning* pada Data Uji

Model	Parameter	Recall 0	Recall 1	F1-score 0	F1-score 1
SVM	C = 10 gamma=” auto” kernel=’rbf’ random_state = 42	0.8780	0.8824	0.8675	0.8911
Random Forest	N = 300	0.9024	0.9608	0.9250	0.9423
KNN	k = 5	0.9268	0.7843	0.8444	0.8511

TABEL 3. Hasil Evaluasi *Recall* dan *F1-score* Model *Machine Learning* pada Data Uji

Pada Tabel 2, ditampilkan nilai akurasi dan precision dari 3 model prediksi penyakit jantung, dengan:

- Kelas 0 = tidak terdeteksi penyakit jantung
- Kelas 1 = terdeteksi penyakit jantung

Random Forest menghasilkan akurasi tertinggi yaitu 93,48% dengan nilai precision yang seimbang pada kedua kelas. Sementara itu, SVM berada di posisi kedua dengan akurasi 88,04%, sedangkan KNN menunjukkan akurasi terendah 84,78% meskipun memiliki *precision* cukup baik pada kelas positif.

Pada Tabel 3, *Random Forest* kembali unggul dengan nilai *recall* dan *F1-score* yang konsisten tinggi dibandingkan dua model lainnya. SVM tetap stabil dengan performa menengah, sedangkan KNN cenderung lebih lemah dalam mendeteksi kelas positif. Hal ini menegaskan bahwa Random Forest merupakan model yang paling andal untuk klasifikasi penyakit jantung dalam penelitian ini.

Kesimpulannya, *Random Forest* adalah model paling andal untuk prediksi penyakit jantung karena memiliki akurasi tinggi serta performa *precision, recall,* dan *F1-score* yang seimbang di kedua kelas. Visualisasi hasil prediksi masing-masing model dapat dilihat melalui *confusion matrix* pada gambar berikut:

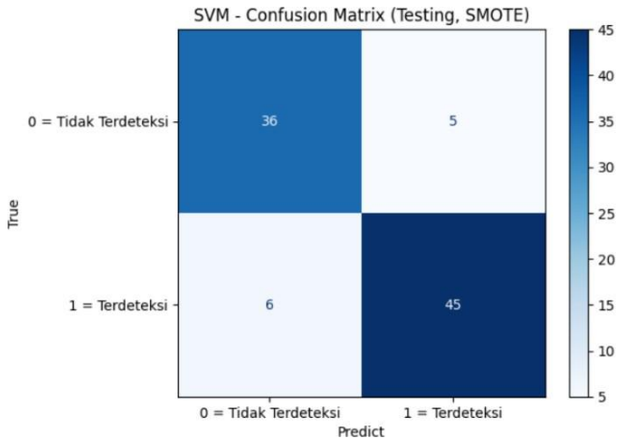


FIG.*1. Confusion Matrix* SVM

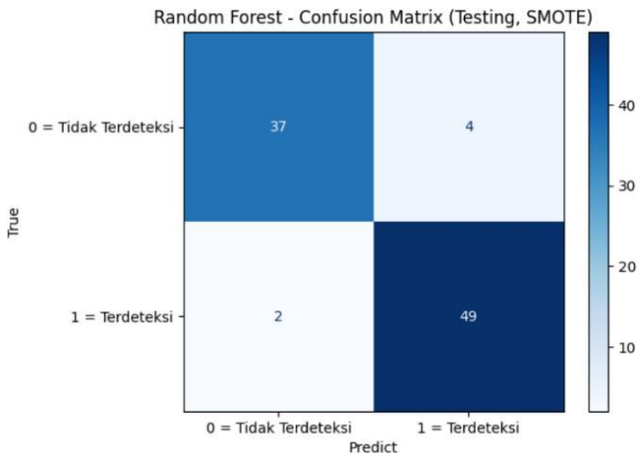


FIG. 2. *Confusion Matrix Random Forest*

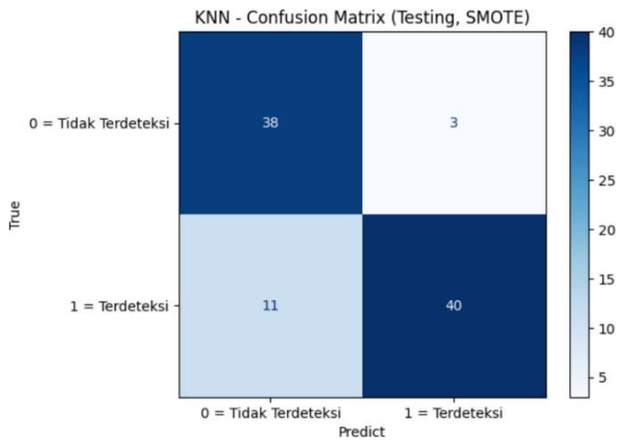


FIG. *3. Confusion Matrix* KNN

Berdasarakan hasil *confusion matrix* dari masing-masing model, dapat dilihat bahwa kemampuan klasifikasi berbeda beda tergantung pada karakteristik algoritma yang digunakan.

1. *Support Vector Machine* (SVM)*:*

Model SVM memberikan hasil prediksi yang cukup seimbang, di mana 36 pasien sehat dan 45 pasien gagal jantung berhasil diklasifikasikan dengan benar. Namun, masih terdapat 6 pasien gagal jantung yang tidak terdeteksi dan 5 pasien sehat yang salah diprediksi sebagai sakit. Hal ini menunjukkan bahwa SVM bekerja cukup baik, tetapi masih menyisakan risiko kesalahan terutama pada kasus gagal jantung yang tidak terdeteksi.

2. *Random Forest:*

Random Forest menunjukkan performa terbaik dibandingkan model lainnya. Model ini mampu mengenali 37 pasien sehat dan 49 pasien gagal jantung secara tepat, dengan hanya 2 kasus gagal jantung yang tidak terdeteksi serta 4 pasien sehat yang salah diklasifikasikan. Dengan jumlah kesalahan yang sangat sedikit, Random Forest dinilai paling andal untuk mendeteksi gagal jantung.

3. *K-Nearest Neighbors* (KNN)*:*

Model KNN dapat mengenali pasien sehat dengan cukup baik, terbukti dari 38 kasus sehat yang terdeteksi benar. Namun, kelemahan utamanya adalah tingginya jumlah pasien gagal jantung yang tidak terdeteksi, yaitu sebanyak 11 orang. Kondisi ini menunjukkan bahwa KNN kurang efektif dalam konteks medis karena masih terlalu sering meloloskan pasien yang seharusnya terdiagnosis gagal jantung.

V. KESIMPULAN

Penelitian ini mengembangkan model klasifikasi penyakit jantung berbasis data klinis dengan pendekatan machine learning, mengintegrasikan seleksi fitur Chi-Square dan teknik penyeimbangan data SMOTE. Dari dataset heart.csv Kaggle yang berisi 918 pasien dengan 11 atribut klinis, dipilih 9 fitur utama yang relevan secara medis *(Age, Sex, ChestPainType, Cholesterol, RestingECG, MaxHR, ExerciseAngina, Oldpeak, dan ST_Slope)*. Tahapan pra-pemrosesan meliputi pembersihan data, encoding variabel kategorikal, penanganan *outlier* dengan IQR, serta normalisasi untuk memastikan kualitas data sebelum pemodelan.

Hasil penelitian menunjukkan bahwa pemilihan fitur terbaik diperoleh pada nilai k = 9 dengan metode chi-square. Model yang diuji terdiri dari tiga algoritma, yaitu Support Vector Machine (SVM) dengan parameter C = 10, gamma = "auto", kernel = "rbf", dan random_state = 42; Random Forest dengan jumlah estimator n = 300; serta K-Nearest Neighbor (KNN) dengan nilai k = 5. Uji coba dilakukan untuk mengevaluasi performa model menggunakan metrik akurasi, precision, recall, dan f1-score.

Berdasarkan hasil pengujian, Random Forest memberikan performa terbaik dengan akurasi sebesar 93,48% serta nilai precision, recall, dan f1-score yang seimbang pada kedua kelas. SVM berada di posisi kedua dengan akurasi 88,04% dan performa yang stabil, sedangkan KNN menghasilkan akurasi 84,78% namun masih lemah dalam mendeteksi pasien dengan penyakit jantung. Random Forest menunjukkan kinerja terbaik sebagai model paling andal untuk mendukung deteksi dini penyakit jantung, sementara SVM dan KNN tetap memberikan kontribusi sebagai model pembanding dalam analisis ini.

REFERENSI

[1] Sitanggang, B. F., & Sitompul, P. (2024). Deteksi awal kelangsungan hidup pasien gagal jantung menggunakan machine learning metode Random Forest. *Innovative: Journal of Social Science Research, 4*(2), 3347–3357.

[2] Puspita, D., & Fadil, M. (2020). Penggunaan ventilasi mekanik pada gagal jantung akut. *Jurnal Kesehatan Andalas, 9*(1S).

[3] World Health Organization. (2019). *Heart disease.* WHO.

[4] Hasibuan, L. (2022, September 29). *Ini 4 penyebab utama penyakit jantung di Indonesia.* CNBC Indonesia. https://www.cnbcindonesia.com

[5] Adi, S., & Wintarti, A. (2022). Komparasi metode Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Random Forest (RF) untuk prediksi penyakit gagal jantung. *MATHunesa: Jurnal Ilmiah Matematika, 10*(2), 258–268.

[6] Wahyono, T. (2018). *Fundamental of Python for Machine Learning (Dasar-dasar Pemrograman Python untuk Machine Learning dan Kecerdasan Buatan)*. Yogyakarta: Gava Media.

[7] Andiani, L., Sukemi, S., & Rini, D. P. (2020, February). Analisis penyakit jantung menggunakan metode KNN dan Random Forest. *Annual Research Seminar (ARS, 5*(1), 165–169).

[8] Rahayu, S., Purnama, J. J., Pohan, A. B., Nugraha, F. S., Nurdiani, S., & Hadianti, S. (2020). Prediction of survival of heart failure patients using Random Forest. *Jurnal Pilar Nusa Mandiri, 16*(2), 255–260.

[9] Yulianti, A., Fitri, F., Amalita, N., & Vionanda, D. (2023). The SMOTE application of CART methods for coping imbalanced data in classifying status work on labor force in the city of Padang. *UNP Journal of Statistics and Data Science, 1*(3), 172–179.

[10] Nurussakinah, N. (2024). *Implementasi algoritma Random Forest dan teknik SMOTE untuk klasifikasi penyakit diabetes* (Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim).

[11] Wibisono, A. B., & Fahrurozi, A. (2020). Perbandingan algoritma klasifikasi dalam pengklasifikasian data penyakit jantung koroner. *Jurnal Ilmiah Teknologi dan Rekayasa, 24*(3), 161–170.

[12] Permana, D. S., & Silvanie, A. (2021). Prediksi penyakit jantung menggunakan Support Vector Machine dan Python pada basis data pasien di Cleveland. *Jurnal Nasional Informatika (JUNIF), 2*(1), 29–34.

[13] Aryanofadh, A. R. (2021). *Analisis sentimen terhadap kebijakan penggunaan sertifikat vaksin menggunakan metode Multi-Layer Perceptron (MLP) dengan ekstraksi fitur fast text* (Bachelor's thesis, Universitas Islam Negeri Syarif Hidayatullah Jakarta).

[14] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

[15] Ananto, D. T., Mahardewantoro, D. D., Mustafa, F., Ardianto, M. G., Rafi, M. M., Zein, R. A., ... & Adharani, Y. (2023, October). Edukasi dan pelatihan pengenalan machine learning dan computer vision untuk mengeksplorasi potensi visual. *Prosiding Seminar Nasional Pengabdian Masyarakat LPPM UMJ.*

[16] Suryajaya, P. I. (2021). *Peranan parameter fungsi diastolik ventrikel kiri berbasis Doppler dan parameter Echo Heart Failure Score sebagai prediktor readmisi terkait gagal jantung sistolik dengan ejeksi fraksi rendah* (Tesis, Universitas Udayana).

[17] Raup, A., Ridwan, W., Khoeriyah, Y., Supiana, S., & Zaqiah, Q. Y. (2022). Deep learning dan penerapannya dalam pembelajaran. *JIIP: Jurnal Ilmiah Ilmu Pendidikan, 5*(9), 3258–3267.

[18] Dharmawan, W. S. (2022). Komparasi algoritma klasifikasi SVM-PSO dan C4.5-PSO dalam prediksi penyakit jantung. *Informatika, 13*(2), 31–41.

[19] Rahayu, D. S., Nursafika, N., Afifah, J., & Intan, S. (2023, August). Klasifikasi penyakit diabetes melitus menggunakan algoritma C4.5, Support Vector Machine (SVM), dan regresi linear. *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat,* 56–63.

[20] Puspita, R., & Widodo, A. (2021). Perbandingan metode KNN, Decision Tree, dan Naïve Bayes terhadap analisis sentimen pengguna layanan BPJS. *Jurnal Informatika Universitas Pamulang, 5*(4), 646–654.

[21] Hidayat, H., Sunyoto, A., & Al Fatta, H. (2023). Klasifikasi penyakit jantung menggunakan Random Forest classifier. *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan), 7*(1), 31–40.

[22] Nugraha, N. P., Azim, R., Daffa, S. Z., & Ningayu, P. S. (2023). Perbandingan akurasi metode Naïve Bayes dan metode KNN untuk memprediksi gagal ginjal kronis. *Jurnal Rekayasa Elektro Sriwijaya, 5*(1), 1–10.

[23] Hidayati, T., & Putra, B. F. (2024). Implementasi deep learning untuk image classification menggunakan Convolutional Neural Network pada citra wayang: Studi kasus SDN Leuwibatu 03. *Scientia Sacra: Jurnal Sains, Teknologi dan Masyarakat, 4*(1), 1–7.

[24] Grandis, G. F., Sari, Y. A., & Indriati, I. (2021). Seleksi fitur Gain Ratio pada analisis sentimen kebijakan pemerintah mengenai pembelajaran jarak jauh dengan K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 5*(8), 3507–3514.

[25] Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering, 53,* 453–462.

[26] Chou, J.-S., Cheng, M.-Y., Wu, Y.-W., & Pham, A.-D. (2014). Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification. *Expert Systems with Applications, 41*(8), 3955–3964.