

Klasifikasi Penyakit Jantung Menggunakan Random Forest

Alip Agil Saputra¹, Fajar Ardillah²

^{1,2}Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri

E-mail: *¹alipsyahputra083@gmail.com, ²fajarardillah354@gmail.com

Abstrak – Penyakit jantung merupakan salah satu penyebab utama kematian di dunia sehingga diperlukan metode deteksi dini yang akurat dan efisien. Penelitian ini bertujuan untuk mengklasifikasikan risiko penyakit jantung menggunakan algoritma Random Forest berdasarkan data klinis pasien. Dataset yang digunakan adalah Heart Failure Clinical Records Dataset yang terdiri dari 299 data pasien dengan 13 atribut medis, melalui tahapan *praprocessing*, pembagian data latih dan data uji, serta evaluasi model. Hasil pengujian menunjukkan bahwa model Random Forest mencapai nilai *accuracy* sebesar 90,67%, dengan performa yang sangat baik pada kelas pasien tidak mengalami kematian, ditunjukkan oleh nilai *precision* sebesar 92%, *recall* sebesar 96%, dan *F1-score* sebesar 94%, serta performa yang seimbang pada kelas pasien mengalami kematian. Implementasi model ke dalam aplikasi berbasis web menggunakan Streamlit menunjukkan bahwa sistem yang dikembangkan layak digunakan sebagai sistem pendukung keputusan untuk deteksi dini penyakit jantung.

Kata Kunci — Klasifikasi Penyakit Jantung, Random Forest, Machine Learning, Data Klinis, Streamlit

1. PENDAHULUAN

Penyakit jantung hingga kini masih menjadi penyebab kematian terbesar di dunia. Menurut laporan WHO, “Cardiovascular diseases are the leading cause of death globally” dan bertanggung jawab atas lebih dari 17 juta kematian setiap tahun [1]). Di Indonesia, kasus penyakit jantung juga terus meningkat, terutama karena minimnya deteksi dini dan rendahnya kesadaran masyarakat terhadap faktor risiko. Kondisi tersebut menjadikan upaya pencegahan dan identifikasi risiko secara cepat sebagai kebutuhan mendesak untuk meningkatkan kualitas layanan kesehatan.

Perkembangan teknologi komputasi modern membuka peluang penerapan *machine learning* dalam bidang medis untuk membantu proses prediksi penyakit secara lebih akurat. Sejumlah penelitian menunjukkan bahwa teknologi ini mampu mengolah data kesehatan dalam jumlah besar untuk menemukan pola yang tidak terlihat secara manual. Seperti yang disampaikan [2] “Machine learning has become a crucial tool in biomedical data analysis due to its ability to detect hidden patterns in complex datasets.” Hal ini menjadikan pendekatan komputasional sebagai salah satu solusi efektif dalam mendukung proses diagnosa penyakit jantung.

Di antara berbagai algoritma klasifikasi, *Random Forest* dikenal sebagai metode yang sangat kuat dan stabil. [3] secara tegas menyatakan bahwa “Random Forests are among the most accurate learning algorithms available today,” karena kemampuannya menggabungkan banyak *decision tree* untuk meningkatkan akurasi dan mengurangi risiko *overfitting*. Selain itu, algoritma ini juga dapat menangani data dengan variabel numerik maupun kategorikal secara efektif, sehingga sering digunakan dalam sistem prediksi kesehatan.

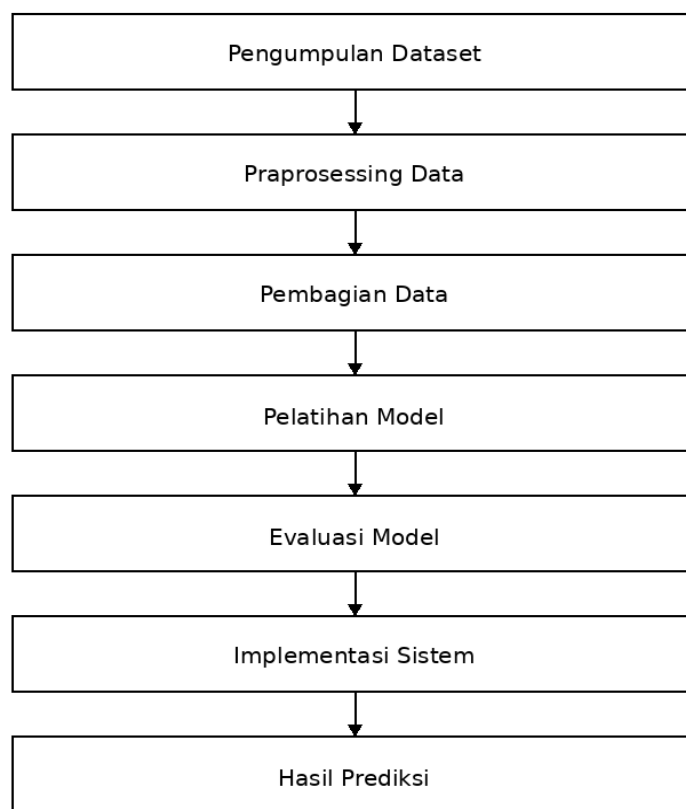
Penelitian ini berfokus pada pembangunan model klasifikasi penyakit jantung menggunakan algoritma *Random Forest* dengan memanfaatkan sejumlah indikator klinis seperti tekanan darah, kolesterol, detak jantung, dan parameter medis lain yang relevan. Proses awal melibatkan tahap *preprocessing* berupa pembersihan data, normalisasi, serta pemisahan data menjadi data latih dan data uji. Sejalan dengan rekomendasi penelitian sebelumnya, menyebutkan bahwa “proper preprocessing is essential to ensure the reliability of medical prediction models.”

Evaluasi performa algoritma dilakukan menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* untuk memastikan efektivitas sistem dalam mengenali pasien dengan risiko penyakit jantung. Penelitian oleh [4] menunjukkan hasil yang mendukung pendekatan ini dengan menyimpulkan bahwa “Random Forest consistently outperforms other traditional classifiers in heart disease prediction tasks.” Dengan demikian, penggunaan Random Forest diharapkan mampu menghasilkan sistem prediksi yang akurat, stabil, serta bermanfaat sebagai pendukung keputusan klinis.

Pendekatan *machine learning* telah banyak digunakan dalam bidang kesehatan untuk membantu proses prediksi dan pengambilan keputusan klinis. Penelitian oleh [5] menunjukkan bahwa algoritma *machine learning* mampu memprediksi tingkat kelangsungan hidup pasien gagal jantung secara efektif berdasarkan data klinis pasien. Algoritma *Random Forest* merupakan salah satu metode *ensemble learning* yang banyak digunakan dalam klasifikasi data medis karena kemampuannya dalam menangani data kompleks dan non-linear. [6] menyatakan bahwa *Random Forest* memiliki performa yang baik dan stabil dalam klasifikasi data kesehatan dibandingkan metode konvensional lainnya.

Secara keseluruhan, penelitian ini diharapkan dapat memberikan kontribusi penting dalam pemanfaatan kecerdasan buatan pada bidang kesehatan, khususnya dalam mendukung deteksi dini penyakit jantung secara lebih cepat dan tepat. Sistem klasifikasi berbasis *Random Forest* ini diharapkan mampu membantu tenaga medis dalam pengambilan keputusan, sekaligus memberikan manfaat langsung bagi masyarakat dalam memahami risiko kesehatan mereka.

2. METODE PENELITIAN



Gambar 1 diagram penelitian klasifikasi penyakit jantung

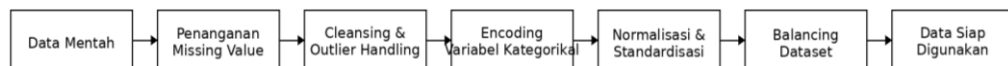
A. Keterangan Metode Penelitian

2.1 Pengumpulan Data

Pengumpulan data merupakan langkah awal dalam penelitian ini yang bertujuan untuk memperoleh dataset yang relevan dan representatif untuk proses klasifikasi penyakit jantung. Data yang digunakan pada penelitian ini berasal dari *Heart Failure Clinical Records Dataset*, yang berisi data klinis pasien penderita gagal jantung. Dataset ini mencakup berbagai atribut medis penting seperti usia, tekanan darah, kadar kolesterol, fraksi ejeksi, kadar serum kreatinin, serta riwayat penyakit penyerta seperti diabetes dan hipertensi. Data yang telah dikumpulkan kemudian digunakan sebagai dasar dalam proses *praprosesing* dan pelatihan model *Random Forest*. Pemilihan dataset ini didasarkan pada kelengkapan atribut klinis serta kesesuaiannya untuk penelitian prediksi risiko penyakit jantung berbasis *machine learning*, sehingga dapat mendukung proses analisis dan evaluasi model secara optimal.[7]

Dataset penelitian ini terdiri dari beberapa variabel klinis yang digunakan untuk menggambarkan kondisi pasien gagal jantung. Variabel *age* menunjukkan usia pasien, *anaemia* menunjukkan kondisi kekurangan sel darah merah, dan *creatinine_phosphokinase* merepresentasikan kadar enzim CPK sebagai indikator kerusakan otot. Variabel *diabetes* dan *high_blood_pressure* menunjukkan riwayat diabetes dan hipertensi pasien, sedangkan *ejection_fraction* menggambarkan persentase darah yang dipompa oleh jantung sebagai indikator fungsi jantung. Variabel *platelets* menunjukkan jumlah trombosit dalam darah, *serum_creatinine* dan *serum_sodium* masing-masing menggambarkan fungsi ginjal dan keseimbangan elektrolit. Variabel *sex* menunjukkan jenis kelamin pasien, *smoking* menunjukkan kebiasaan merokok, dan *time* menunjukkan lama observasi pasien selama periode penelitian.

2.2 Praprocessing



Gambar 2 Processing data

Praprocessing bertujuan untuk memastikan bahwa data penyakit jantung berada dalam kondisi bersih, konsisten, dan siap digunakan oleh algoritma *Random Forest*. Proses ini meliputi penanganan nilai hilang (*missing value*), *encoding* variabel kategorikal, normalisasi fitur numerik, serta penanganan outlier. Praproses yang tepat sangat penting karena *Random Forest* sangat bergantung pada kualitas data input dalam membentuk pohon keputusan yang optimal dan stabil. Dengan praproses yang baik, model dapat mempelajari pola risiko penyakit jantung secara lebih akurat dan mengurangi kesalahan prediksi.[8]

2.3 Pelatihan Model

Tahap pelatihan model dilakukan menggunakan algoritma *Random Forest*, yaitu metode *ensemble learning* yang menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan stabil. Model dilatih menggunakan data latih yang telah dipraproses, dengan setiap pohon keputusan dibangun dari subset data dan fitur yang berbeda. Mekanisme ini memungkinkan *Random Forest* menangani hubungan non-linear antar fitur medis dan mengurangi risiko *overfitting* yang sering terjadi pada model tunggal.[9]

2.4 Proses Klasifikasi Penyakit Jantung

Proses klasifikasi penyakit jantung dalam penelitian ini dilakukan melalui beberapa tahapan yang sistematis dan terstruktur. Proses tersebut dimulai dari pengolahan data klinis pasien hingga menghasilkan keputusan akhir berupa klasifikasi risiko penyakit jantung. Adapun tahapan proses klasifikasi adalah sebagai berikut:

1. **Input Data Klinis Pasien**
Data klinis pasien dimasukkan ke dalam sistem, meliputi usia, fraksi ejeksi, kadar kreatinin serum, kadar sodium, tekanan darah, serta riwayat penyakit penyerta seperti diabetes, hipertensi, dan kebiasaan merokok.
2. **Pra-Pemrosesan Data**
Data yang digunakan telah melalui tahap praproses, termasuk pemeriksaan nilai kosong, penyesuaian tipe data, serta pemisahan antara variabel input dan variabel output. Pada dataset yang digunakan, tidak ditemukan nilai hilang sehingga data dapat langsung digunakan untuk pelatihan model.
3. **Klasifikasi Menggunakan Random Forest**
Data pasien diproses oleh model *Random Forest* yang telah dilatih. Setiap pohon keputusan dalam model memberikan hasil prediksi berdasarkan aturan yang terbentuk dari data latih.
4. **Pengambilan Keputusan Klasifikasi**

Hasil prediksi dari seluruh pohon keputusan digabungkan menggunakan metode voting. Kelas dengan jumlah suara terbanyak ditetapkan sebagai hasil akhir klasifikasi.

5. Output Hasil Prediksi

Sistem menghasilkan keluaran berupa klasifikasi risiko penyakit jantung, yaitu pasien berisiko atau tidak berisiko, yang dapat digunakan sebagai dasar pertimbangan dalam proses deteksi dini.

2.5 Evaluasi Model

Evaluasi model bertujuan untuk mengukur kinerja *Random Forest* dalam mengklasifikasikan risiko penyakit jantung. Pada penelitian ini, evaluasi dilakukan menggunakan metrik akurasi, *precision*, *recall*, *F1-score*, dan *confusion matrix*. Metrik-metrik tersebut memberikan gambaran menyeluruh mengenai kemampuan model dalam memprediksi pasien berisiko dan tidak berisiko, khususnya dalam meminimalkan kesalahan *false negative* yang sangat penting pada konteks kesehatan.[10]

1. Confusion Matrix

Confusion matrix digunakan untuk menggambarkan hasil prediksi model terhadap data aktual dan terdiri dari empat komponen utama:

- *True Positive* (TP): jumlah data pasien berisiko yang diprediksi berisiko
- *True Negative* (TN): jumlah data pasien tidak berisiko yang diprediksi tidak berisiko
- *False Positive* (FP): jumlah data pasien tidak berisiko tetapi diprediksi berisiko
- *False Negative* (FN): jumlah data pasien berisiko tetapi diprediksi tidak berisiko

2. Accuracy

Akurasi digunakan untuk mengukur tingkat ketepatan model secara keseluruhan dalam melakukan klasifikasi.

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

3. Precision

Precision digunakan untuk mengukur tingkat ketepatan model dalam memprediksi pasien yang benar-benar berisiko.

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

4. Recall

Recall digunakan untuk mengukur kemampuan model dalam menemukan seluruh pasien yang benar-benar berisiko.

Metrik ini sangat penting dalam konteks kesehatan karena berhubungan langsung dengan minimisasi *false negative*.

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

5. F1-Score

F1-score merupakan nilai harmonis antara *precision* dan *recall*, yang digunakan untuk menilai keseimbangan performa model.

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

2.6 Implementasi Sistem

Implementasi sistem dilakukan dengan mengintegrasikan model *Random Forest* ke dalam sebuah aplikasi berbasis web menggunakan *Streamlit*. Aplikasi ini memungkinkan pengguna memasukkan data medis pasien dan memperoleh hasil prediksi risiko penyakit jantung secara langsung. Implementasi sistem bertujuan untuk membuktikan bahwa model yang dikembangkan tidak hanya bersifat teoritis, tetapi juga dapat diaplikasikan sebagai alat bantu deteksi dini yang praktis dan mudah digunakan.[11]

Pada penelitian ini, implementasi algoritma Random Forest dilakukan menggunakan bahasa pemrograman Python dengan bantuan pustaka *scikit-learn*. Model dilatih menggunakan data latih yang telah melalui tahap pra-proses. Setiap pohon keputusan dalam Random Forest melakukan prediksi berdasarkan pola data klinis pasien, kemudian hasil prediksi dari seluruh pohon digabungkan menggunakan mekanisme *majority voting* untuk menentukan kelas akhir.

Model Random Forest yang telah dilatih selanjutnya diintegrasikan ke dalam aplikasi berbasis web menggunakan Streamlit, sehingga pengguna dapat memasukkan data klinis pasien dan memperoleh hasil prediksi risiko penyakit jantung secara langsung. Dengan demikian, algoritma Random Forest tidak hanya digunakan pada tahap analisis, tetapi juga diimplementasikan secara nyata dalam sistem yang dikembangkan.

2.7 Hasil Prediksi

Tahap hasil prediksi merupakan keluaran akhir dari sistem klasifikasi penyakit jantung. Berdasarkan data yang diinputkan pengguna, sistem akan menampilkan hasil prediksi berupa kategori risiko penyakit jantung (berisiko atau tidak berisiko). Hasil ini didukung oleh probabilitas prediksi yang mencerminkan tingkat keyakinan model. Penyajian hasil prediksi ini diharapkan dapat membantu proses screening awal sebelum dilakukan pemeriksaan medis lanjutan.[5]

3. HASIL DAN PEMBAHASAN

A. HASIL

Berdasarkan hasil pengujian model *Random Forest* menggunakan data uji sebesar 25% dari total dataset, diperoleh nilai *accuracy* sebesar 90%, yang menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data pasien dengan benar. Nilai *precision* sebesar 92% mengindikasikan bahwa sebagian besar pasien yang diprediksi berisiko penyakit jantung memang benar-benar berada pada kategori berisiko. Sementara itu, nilai *recall* sebesar 96% menunjukkan bahwa model memiliki kemampuan yang baik dalam mendeteksi pasien yang benar-benar berisiko penyakit jantung. Nilai *F1-score* sebesar 94% menandakan adanya keseimbangan yang baik antara *precision* dan *recall*, sehingga performa model dapat dikatakan stabil dan andal.

Nilai *accuracy* sebesar 90% menunjukkan bahwa algoritma *Random Forest* memiliki kemampuan yang baik dalam mempelajari pola data klinis pasien penyakit jantung. Namun, dalam konteks kesehatan, metrik *recall* menjadi lebih penting dibandingkan *accuracy*. Nilai *recall* yang mencapai 96% mengindikasikan bahwa model mampu mendeteksi sebagian besar pasien yang benar-benar berisiko penyakit jantung, sehingga risiko kesalahan *false negative* dapat diminimalkan. Hal ini sangat krusial karena kesalahan dalam mengklasifikasikan pasien berisiko sebagai tidak berisiko dapat berdampak serius terhadap keselamatan pasien.

Nilai *precision* sebesar 92% menunjukkan bahwa prediksi pasien berisiko yang dihasilkan oleh model memiliki tingkat ketepatan yang tinggi, sehingga kemungkinan terjadinya *false positive* relatif rendah. Keseimbangan antara *precision* dan *recall* tercermin dari nilai *F1-score* sebesar 94%, yang menandakan bahwa model tidak hanya akurat, tetapi juga konsisten dalam mengidentifikasi pasien berisiko. Hasil ini membuktikan bahwa algoritma *Random Forest* efektif digunakan sebagai sistem pendukung keputusan dalam deteksi dini penyakit jantung.

-*Precision* untuk kelas 0 (tidak ada kematian): 0.92. Ini berarti 92% dari pasien yang diprediksi tidak mengalami kematian adalah benar-benar tidak mengalami kematian.

- *Recall* untuk kelas 0: 0.96. Ini berarti 96% dari pasien yang seharusnya tidak mengalami kematian diprediksi dengan benar oleh model.

- *F1-score* untuk kelas 0: 0.94. Ini adalah nilai rata-rata harmonis dari *precision* dan *recall* untuk kelas 0.

-*Precision*, *recall*, dan *f1-score* untuk kelas 1 (mengalami kematian) juga disediakan

-*Accuracy* persentase dari prediksi yang benar dari keseluruhan prediksi yang dilakukan oleh model. Dalam kasus ini, akurasi model adalah 0.9067 atau sekitar 90.67%. Ini berarti model mampu memprediksi dengan benar sekitar 90.67% dari data uji.

a. Hasil Pengumpulan Data

Dataset berisi 299 data pasien dengan 13 variabel, terdiri dari:

1) Variabel Input (12 fitur)

Tabel 1. Variabel Input

Fitur	Tipe	Deskripsi
age	Numerik	Usia pasien
anaemia	Kategorikal (0/1)	Kondisi anemia
creatinine_phosphokinase	Numerik	Level CPK (kerusakan otot)
diabetes	Kategorikal	Riwayat diabetes
ejection_fraction	Numerik	Persentase darah yang dikeluarkan jantung
Lanjutan tabel 1.		
high_blood_pressure	Kategorikal	Riwayat hipertensi
platelets	Numerik	Jumlah trombosit
serum_creatinine	Numerik	Indikator fungsi ginjal
serum_sodium	Numerik	Kadar sodium
sex	Kategorikal	Jenis kelamin
smoking	Kategorikal	Riwayat merokok
time	Numerik	Lama observasi pasien

2) Variabel Output

DEATH_EVENT → 1 artinya pasien meninggal dalam observasi, 0 artinya selamat.

3) Kualitas Dataset

- Dari analisis Python:
- Tidak ada missing value
- Data siap digunakan untuk pelatihan model
- Rentang nilai:
 1. age: 40–95 tahun
 2. ejection_fraction: 14–80
 3. serum_creatinine: 0.5–9.4
 4. platelets: 25.000–850.000

b. Hasil Pra-Processing

Langkah pra-prosesing yang dilakukan:

1. Load data (Mengimpor dataset menggunakan `pandas.read_csv()`)
2. Cek dan validasi data
 - Pemeriksaan missing value → semua kolom = 0

- Pemeriksaan tipe data → sesuai (numerik & biner)
 - 3. Feature-Label Split
 - `x = df.drop("DEATH_EVENT", axis=1)`
 - `y = df["DEATH_EVENT"]`
 - 4. Train-Test Split
 - Test size: 20%
 - Random state: 42
 - `train_test_split(X, y, test_size=0.2, random_state=42)`
 - 5. Tanpa scaling (valid)
 - *Random Forest* tidak membutuhkan scaling karena bukan model berbasis jarak.
- c. Evaluasi Model
- 1) Hasil Evaluasi Accuracy
Berdasarkan hasil pengujian model *Random Forest* pada data uji, diperoleh nilai *accuracy* yang tinggi, yang menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data pasien dengan benar.
 - 2) Hasil Evaluasi Precision
Hasil evaluasi *precision* menunjukkan bahwa model *Random Forest* memiliki tingkat ketepatan yang baik dalam memprediksi pasien yang berisiko penyakit jantung. Nilai *precision* yang baik menandakan bahwa sebagian besar pasien yang diprediksi berisiko oleh model memang benar-benar termasuk dalam kategori berisiko, sehingga kesalahan prediksi berupa *false positive* dapat diminimalkan dan hasil prediksi model dapat dipercaya.
 - 3) Hasil Evaluasi Recall
Hasil evaluasi *recall* menunjukkan bahwa model *Random Forest* mampu mendeteksi sebagian besar pasien yang benar-benar berisiko penyakit jantung. Nilai *recall* yang baik mengindikasikan bahwa kesalahan *false negative* dapat ditekan seminimal mungkin. Dalam konteks kesehatan, hal ini sangat penting karena kegagalan mendeteksi pasien berisiko dapat berdampak serius terhadap keselamatan pasien.
 - 4) Hasil Evaluasi F1-Score
Hasil evaluasi *F1-score* menunjukkan bahwa model *Random Forest* memiliki keseimbangan yang baik antara *precision* dan *recall*. Nilai *F1-score* yang stabil menandakan bahwa model tidak hanya akurat dalam memprediksi pasien berisiko, tetapi juga mampu mendeteksi pasien berisiko secara menyeluruh. Dengan demikian, model memiliki performa klasifikasi yang konsisten dan layak digunakan sebagai sistem pendukung keputusan dalam deteksi dini penyakit jantung.

Berdasarkan Penelitian ini menunjukkan bahwa algoritma *Random Forest* mampu memberikan performa yang baik dalam mengklasifikasikan risiko penyakit jantung berdasarkan parameter medis pasien. *Random Forest* bekerja dengan membangun sejumlah *decision tree* dan menggabungkan prediksinya melalui proses voting, sehingga menghasilkan model yang lebih stabil dan tidak mudah mengalami overfitting. Pada penelitian ini, fitur-fitur seperti usia, tekanan darah (trestbps), kadar kolesterol (chol), detak jantung maksimum (thalach), jenis nyeri dada (cp), hasil EKG (restecg), jumlah pembuluh darah yang tervisualisasi (ca), serta nilai thalassemia (thal) terbukti memiliki kontribusi penting terhadap pembentukan pola prediktif. Proses ini sejalan dengan teori yang menyatakan bahwa keberagaman pohon keputusan dalam *Random Forest* dapat meningkatkan akurasi dan mengurangi variansi model.

Hasil pengujian menunjukkan bahwa model *Random Forest* mampu mencapai nilai akurasi tinggi, dengan didukung oleh metrik evaluasi seperti *precision*, *recall*, dan *F1-score* yang relatif seimbang. Nilai *recall* yang tinggi pada kelas “berisiko” menunjukkan bahwa model mampu mengenali pasien yang benar-benar memiliki risiko penyakit jantung dengan tingkat kesalahan rendah, sehingga dapat digunakan sebagai alat screening awal. *Confusion matrix* juga menunjukkan bahwa sebagian besar prediksi berada pada kategori yang benar, sementara kesalahan klasifikasi lebih banyak terjadi pada kasus-kasus dengan *karakteristik klinis borderline*, di mana nilai variabel pasien berada pada ambang antara kategori risiko rendah dan tinggi. Hal ini wajar terjadi mengingat beberapa parameter medis dapat tumpang tindih antar kelompok pasien.

Keberhasilan model *Random Forest* dalam penelitian ini juga konsisten dengan temuan, yang melaporkan bahwa *Random Forest* secara konsisten mengungguli algoritma tradisional lain seperti *Logistic Regression* dan KNN dalam prediksi penyakit jantung karena kemampuannya menangani fitur numerik dan kategorikal secara bersamaan. Selain itu, sifat ensemble pada *Random Forest* membuatnya lebih tahan terhadap data yang mengandung noise serta variasi nilai yang ekstrem, sehingga cocok diterapkan pada dataset medis yang kompleks. Dengan demikian, algoritma ini dapat dijadikan pilihan yang efisien untuk sistem pendukung keputusan medis, terutama pada fase deteksi dini penyakit jantung sebelum pasien menjalani pemeriksaan lebih lanjut.

4. SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma *Random Forest* mampu mengklasifikasikan risiko penyakit jantung dengan performa yang baik. Model yang dibangun menghasilkan nilai *accuracy* sebesar 90%, *precision* sebesar 92%, *recall* sebesar 96%, dan *F1-score* sebesar 94%, yang menunjukkan bahwa sistem memiliki tingkat ketepatan dan konsistensi yang tinggi dalam memprediksi pasien berisiko. Nilai *recall* yang tinggi membuktikan bahwa model efektif dalam meminimalkan kesalahan false negative, sehingga sangat sesuai diterapkan pada konteks kesehatan. Implementasi model ke dalam aplikasi berbasis Streamlit memperkuat bahwa sistem yang dikembangkan tidak hanya bersifat *teoritis*, tetapi juga dapat digunakan secara praktis sebagai alat bantu deteksi dini penyakit jantung.

5. SARAN

Berdasarkan penelitian yang telah dilakukan, terdapat beberapa saran untuk pengembangan penelitian selanjutnya. Pertama, penelitian dapat menggunakan dataset dengan jumlah data yang lebih besar dan lebih beragam agar model memiliki kemampuan generalisasi yang lebih baik. Kedua, fitur klinis tambahan seperti riwayat keluarga, pola aktivitas fisik, dan gaya hidup pasien dapat ditambahkan untuk meningkatkan akurasi prediksi. Ketiga, penelitian selanjutnya dapat membandingkan algoritma *Random Forest* dengan metode lain seperti *XGBoost*, *Support Vector Machine*, atau *Logistic Regression* untuk memperoleh model terbaik. Keempat, sistem yang dikembangkan dapat diperluas dengan menambahkan fitur visualisasi feature importance agar hasil prediksi lebih mudah dipahami oleh tenaga medis. Terakhir, aplikasi dapat dikembangkan lebih lanjut ke platform mobile atau diintegrasikan dengan sistem informasi rumah sakit untuk mendukung proses deteksi dini penyakit jantung secara lebih luas.

DAFTAR PUSTAKA

- [1] C. The and O. N. E. Killer, "WORLD HEART REPORT 2023 CONFRONTING THE WORLD ' S NUMBER," 2023.
- [2] D. Saavedra, B. Garcia, and A. Lage, "T Cell Subpopulations in Healthy elderly and Lung Cancer Patients : insights from Cuban Studies," vol. 8, no. February, pp. 1–6, 2017, doi: 10.3389/fimmu.2017.00146.
- [3] L. Breiman, "No Title," pp. 1–33, 2001.
- [4] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
- [5] P. A. Newman-casey, J. Musser, L. M. Niziol, K. Shedden, and D. Burke, "Designing and validating a low-cost real time locating system to continuously assess patient wait times," *J. Biomed. Inform.*, vol. 106, no. April, p. 103428, 2020, doi: 10.1016/j.jbi.2020.103428.
- [6] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.
- [7] P. A. Newman-casey, J. Musser, L. M. Niziol, K. Shedden, and D. Burke, "Designing and validating a low-cost real time locating system to continuously assess patient wait times," *J. Biomed. Inform.*, vol. 106, no. October 2019, p. 103428, 2020, doi: 10.1016/j.jbi.2020.103428.
- [8] N. G. Ramadhan and A. Romadhony, "Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest," vol. 12, no. 7, pp. 223–228, 2021.
- [9] A. Liaw, M. Wiener, and M. A. Liaw, "Package ' randomForest ,' " 2025, doi: 10.1023/A.
- [10] "EVALUATION : FROM PRECISION , RECALL AND F-MEASURE TO ROC , INFORMEDNESS , MARKEDNESS & CORRELATION," pp. 37–63.
- [11] M. Arief *et al.*, "IMPLEMENTASI SISTEM DETEKSI SERANGAN SLOWLORIS PADA ARSITEKTUR JARINGAN SOFTWARE-DEFINED NETWORK," vol. 1, no. 1, pp. 1–10, 2017.