

Prediksi Risiko Penyakit Jantung Menggunakan Algoritma Random Forest Berdasarkan data Kesehatan Pasien

Salsa Julia Iriani¹, Arda Try Nabila²

^{1,2}Program Studi Sistem Informasi, Universitas Darwan Ali

Email : salsairn05@gmail.com¹, atrynabila@gmail.com²

ABSTRACT— Cardiovascular disease remains the leading cause of death worldwide, so early and data-driven detection of heart disease risk is needed to support timely clinical decisions. This study applies the Random Forest algorithm to classify heart disease status using a public heart_disease.csv dataset consisting of 10,000 patient records with 21 demographic, clinical, and lifestyle attributes. Missing values were handled through mean imputation for numerical attributes and mode imputation for categorical attributes, and all categorical variables were transformed into numerical form using Label Encoding. The dataset was then split into 80% training data and 20% testing data, and a Random Forest Classifier with 100 estimators and the entropy criterion was trained to predict heart disease status. The resulting model achieved an overall accuracy of 80.65%; however, evaluation using precision, recall, and F1-score revealed that the model completely failed to identify patients who actually have heart disease (class 1), obtaining a recall of 0% for that class. This indicates that the model is strongly biased toward the majority class due to class imbalance in the dataset, and that overall accuracy alone is a misleading metric for this problem. Feature importance analysis further showed that Sleep Hours, Homocysteine Level, CRP Level, BMI, and Triglyceride Level are the most influential predictors of heart disease status in this dataset. These findings suggest that class-balancing techniques such as oversampling, undersampling, SMOTE, or class weighting are necessary in future work in order to build a clinically meaningful heart disease prediction model.

Keywords— Heart disease, random forest, classification, machine learning, class imbalance, feature importance.

ABSTRAK— Penyakit jantung merupakan penyebab kematian tertinggi di dunia, sehingga deteksi dini risiko penyakit jantung berbasis data sangat diperlukan untuk mendukung pengambilan keputusan klinis yang tepat waktu. Penelitian ini menerapkan algoritma Random Forest untuk mengklasifikasikan status penyakit jantung dengan menggunakan dataset heart_disease.csv yang terdiri atas 10.000 data pasien dengan 21 atribut demografi, klinis, dan gaya hidup. Nilai kosong pada dataset ditangani melalui imputasi rata-rata (mean) untuk atribut numerik dan modus untuk atribut kategorikal, sedangkan seluruh variabel kategorikal diubah menjadi bentuk numerik menggunakan Label Encoding. Dataset kemudian dibagi menjadi 80% data latih dan 20% data uji, lalu model Random Forest Classifier dengan 100 estimator dan kriteria entropy dilatih untuk memprediksi status penyakit jantung. Model yang dihasilkan memperoleh akurasi keseluruhan sebesar 80,65%; namun evaluasi menggunakan precision, recall, dan F1-score menunjukkan bahwa model sama sekali gagal mengenali pasien yang benar-benar menderita penyakit jantung (kelas 1), dengan nilai recall 0% untuk kelas tersebut. Hal ini mengindikasikan model bias kuat terhadap kelas mayoritas akibat ketidakseimbangan kelas (class imbalance) pada dataset, dan bahwa akurasi keseluruhan saja merupakan metrik yang menyesatkan untuk permasalahan ini. Analisis feature importance lebih lanjut menunjukkan bahwa Sleep Hours, Homocysteine Level, CRP Level, BMI, dan Triglyceride Level merupakan atribut yang paling berpengaruh terhadap status penyakit jantung pada dataset ini. Temuan ini menunjukkan bahwa teknik penyeimbangan kelas seperti oversampling, undersampling, SMOTE, atau pembobotan kelas diperlukan pada penelitian selanjutnya agar diperoleh model prediksi penyakit jantung yang lebih bermakna secara klinis.

Kata kunci— Penyakit jantung, random forest, klasifikasi, machine learning, ketidakseimbangan kelas, feature importance.

I. PENDAHULUAN

Penyakit jantung atau penyakit kardiovaskular merupakan sekelompok gangguan pada jantung dan pembuluh darah yang hingga saat ini masih menjadi penyebab kematian tertinggi di dunia. Organisasi Kesehatan Dunia (WHO) mencatat bahwa sekitar 19,8 juta orang meninggal akibat penyakit kardiovaskular pada tahun 2022, atau setara dengan kurang lebih 32% dari seluruh kematian secara global, dan sebagian besar di antaranya disebabkan oleh serangan jantung dan stroke [1]. Tingginya angka kematian akibat penyakit jantung ini menunjukkan pentingnya upaya deteksi dini agar

penanganan medis dapat dilakukan sebelum kondisi pasien menjadi semakin parah.

Perkembangan teknologi kecerdasan buatan, khususnya machine learning, membuka peluang besar untuk membantu proses deteksi dini penyakit jantung secara lebih cepat, murah, dan objektif dibandingkan dengan pemeriksaan manual oleh tenaga medis. Penelitian ini diarahkan untuk memanfaatkan data kesehatan pasien yang bersifat historis guna membangun model klasifikasi yang mampu memprediksi status penyakit jantung seseorang. Algoritma yang digunakan dalam penelitian ini adalah Random Forest, sebuah algoritma ensemble learning yang dikenal memiliki performa baik dan tahan

terhadap overfitting pada berbagai kasus klasifikasi data medis [2].

Data yang digunakan dalam penelitian ini adalah dataset `heart_disease.csv` yang terdiri atas 10.000 data pasien dengan 21 atribut, meliputi data demografi seperti usia dan jenis kelamin, data klinis seperti tekanan darah, kadar kolesterol, kadar gula darah puasa, kadar CRP (C-Reactive Protein), kadar homosistein, dan kadar trigliserida, serta data gaya hidup seperti kebiasaan olahraga, kebiasaan merokok, konsumsi alkohol, konsumsi gula, tingkat stres, dan jam tidur. Sebagian atribut pada dataset ini memiliki nilai kosong (missing value), dengan jumlah baris yang mengandung setidaknya satu nilai kosong mencapai 2.933 baris dari total 10.000 baris data, sehingga diperlukan tahap pembersihan data sebelum data dapat digunakan untuk pemodelan.

Permasalahan utama yang ingin dijawab dalam penelitian ini adalah sejauh mana algoritma Random Forest mampu memprediksi status penyakit jantung pasien secara akurat berdasarkan atribut kesehatan yang tersedia, serta atribut apa yang paling berpengaruh terhadap prediksi tersebut. Selain itu, penelitian ini juga ingin mengevaluasi apakah metrik akurasi saja sudah cukup untuk menilai kualitas model klasifikasi penyakit, mengingat pada kasus medis distribusi kelas antara pasien sehat dan pasien sakit seringkali tidak seimbang (class imbalance), sehingga model yang terlihat akurat secara keseluruhan belum tentu mampu mengenali kelas pasien yang benar-benar sakit.

Beberapa penelitian terdahulu telah menunjukkan potensi algoritma Random Forest dalam prediksi penyakit jantung. Charan, Mahendranath, dan Thirunavukkarasu [3] melaporkan bahwa Random Forest mampu mencapai akurasi hingga 92,16% pada dataset UCI Heart Disease. Suleiman, Luka, dan Ibrahim [7] serta Sumwiza dkk. [8] juga menunjukkan bahwa Random Forest secara konsisten unggul dibandingkan algoritma klasifikasi lainnya seperti Support Vector Machine, Naive Bayes, dan Logistic Regression pada berbagai dataset penyakit jantung. Sementara itu, Pradini dkk. [5] membandingkan Random Forest dengan tiga algoritma lain dan menyimpulkan bahwa Random Forest memberikan akurasi tertinggi sebesar 86,8% dengan kemampuan membedakan kelas yang baik berdasarkan kurva ROC.

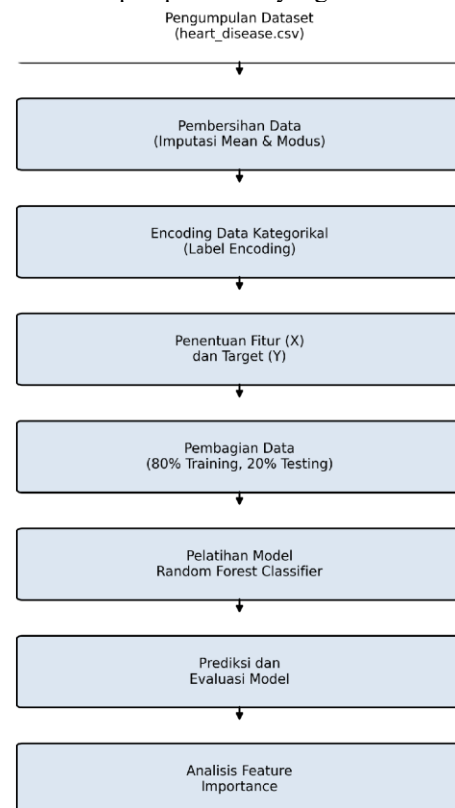
Meskipun berbagai penelitian terdahulu melaporkan akurasi Random Forest yang tinggi, sebagian besar penelitian tersebut belum secara khusus menyoroti permasalahan ketidakseimbangan kelas yang umum terjadi pada dataset medis. Khalilia, Chakraborty, dan Popescu [4] menegaskan bahwa prediksi risiko penyakit dari data yang sangat tidak seimbang memerlukan penanganan khusus, karena model yang dilatih tanpa mempertimbangkan proporsi kelas cenderung bias terhadap kelas mayoritas meskipun akurasi keseluruhannya tetap terlihat tinggi. Jamadar dkk. [10] serta Singh dan Sharma [9] juga menekankan pentingnya tahap pra-pemrosesan data yang baik, termasuk penanganan nilai kosong dan pemilihan fitur, agar model yang dibangun benar-benar mampu mengenali pola

penyakit jantung secara andal, bukan sekadar mengejar angka akurasi.

Berdasarkan uraian tersebut, penelitian ini mengangkat tema penerapan algoritma Random Forest untuk prediksi penyakit jantung dengan tidak hanya berhenti pada pelaporan nilai akurasi, tetapi juga melakukan evaluasi yang lebih mendalam menggunakan precision, recall, F1-score, confusion matrix, dan analisis feature importance. Dengan pendekatan ini, penelitian diharapkan dapat memberikan gambaran yang lebih jujur mengenai kemampuan dan keterbatasan model Random Forest pada dataset `heart_disease.csv`, sekaligus memberikan rekomendasi perbaikan bagi penelitian selanjutnya, khususnya terkait penanganan ketidakseimbangan kelas.

II. METODOLOGI PENELITIAN

Penelitian ini dilaksanakan melalui delapan tahapan utama, mulai dari pengumpulan dataset hingga analisis feature importance. Seluruh tahapan diimplementasikan menggunakan bahasa pemrograman Python dengan memanfaatkan pustaka `pandas`, `numpy`, `scikit-learn`, `matplotlib`, dan `seaborn` [6]. Gambar 1 menunjukkan diagram alur tahapan penelitian yang dilakukan.



Gambar 1. Diagram Alur tahapan penelitian

A. Sumber Data

Data yang digunakan pada penelitian ini adalah dataset `heart_disease.csv` yang berisi 10.000 baris data pasien dengan 21 kolom atribut. Atribut tersebut terdiri atas 9 atribut numerik (Age, Blood Pressure, Cholesterol Level, BMI, Sleep Hours, Triglyceride Level, Fasting

Blood Sugar, CRP Level, dan Homocysteine Level) serta 12 atribut kategorikal (Gender, Exercise Habits, Smoking, Family Heart Disease, Diabetes, High Blood Pressure, Low HDL Cholesterol, High LDL Cholesterol, Alcohol Consumption, Stress Level, Sugar Consumption, dan Heart Disease Status sebagai variabel target). Dataset dibaca menggunakan fungsi `pandas.read_csv()` ke dalam sebuah DataFrame bernama data.

B. Pembersihan dan Pra-pemrosesan Data

Pada tahap ini dilakukan pemeriksaan nilai kosong `sss(missing value)` menggunakan fungsi `data.info()` dan `data.isnull()`. Alih-alih menghapus baris yang mengandung nilai kosong, penelitian ini menerapkan strategi imputasi agar jumlah data tetap terjaga sebanyak 10.000 baris. Nilai kosong pada kolom numerik diisi menggunakan nilai rata-rata (mean) dari kolom yang bersangkutan, sedangkan nilai kosong pada kolom kategorikal diisi menggunakan nilai modus (nilai yang paling sering muncul) dari kolom tersebut, masing-masing menggunakan perulangan `for` pada kolom bertipe `float64/int64` dan kolom bertipe `object`.

C. Transformasi Data Kategorikal

Karena algoritma Random Forest pada `scikit-learn` memerlukan input berupa nilai numerik, seluruh kolom bertipe kategorikal (`object`) diubah menjadi representasi numerik menggunakan `LabelEncoder` dari pustaka `sklearn.preprocessing`. Proses encoding dilakukan terhadap setiap kolom kategorikal secara berurutan menggunakan perulangan `for`, termasuk kolom target Heart Disease Status yang diubah menjadi label biner 0 (Tidak Sakit) dan 1 (Sakit).

D. Penentuan Fitur dan Pembagian Data

Variabel fitur (X) ditentukan dengan menghapus kolom Heart Disease Status dari DataFrame, sedangkan variabel target (y) diambil dari kolom Heart Disease Status. Selanjutnya, dataset dibagi menjadi data latih (training set) dan data uji (testing set) menggunakan fungsi `train_test_split` dengan proporsi 80% data latih dan 20% data uji serta parameter `random_state=42` agar hasil pembagian data dapat direproduksi.

E. Pembangunan Model Random Forest

Model klasifikasi dibangun menggunakan `RandomForestClassifier` dari pustaka `sklearn.ensemble` dengan parameter `n_estimators=100` (jumlah pohon keputusan sebanyak 100 buah), `criterion='entropy'` (ukuran kualitas pemisahan berdasarkan information gain), dan `random_state=42`. Model kemudian dilatih menggunakan data latih (X_{train} , y_{train}) melalui pemanggilan fungsi `model.fit()`.

F. Evaluasi Model

Model yang telah dilatih digunakan untuk melakukan prediksi terhadap data uji (X_{test}) menggunakan fungsi `model.predict()`. Hasil prediksi (y_{pred}) kemudian dibandingkan dengan label sebenarnya (y_{test}) untuk menghitung nilai akurasi menggunakan `accuracy_score`, serta laporan klasifikasi yang meliputi `precision`, `recall`, dan `F1-score` menggunakan `classification_report`. Selain itu, `confusion matrix` dihitung menggunakan `confusion_matrix` dan divisualisasikan dalam bentuk `heatmap` menggunakan

pustaka `seaborn`. Sebagai tahap akhir, tingkat kepentingan setiap fitur (`feature importance`) diambil dari atribut `feature_importances_` pada model Random Forest yang telah dilatih, kemudian diurutkan dan divisualisasikan dalam bentuk grafik batang.

III. DESAIN, HASIL DAN PEMBAHASAN

A. Deskripsi Dataset

Dataset `heart_disease.csv` terdiri atas 10.000 baris data pasien dengan 21 atribut. Tabel I menunjukkan deskripsi masing-masing atribut beserta tipe data dan jumlah data yang terisi (non-null) sebelum proses pembersihan data dilakukan. Terlihat bahwa hampir seluruh atribut memiliki sejumlah kecil nilai kosong, dengan atribut Alcohol Consumption memiliki jumlah nilai kosong paling banyak, yaitu hanya 7.414 dari 10.000 data yang terisi.

TABEL I
 DESKRIPSI ATRIBUT DATASET HEART_DISEASE.CSV

No	Atribut	Tipe Data	Non-Null
1	Age	float64	9971
2	Gender	object	9981
3	Blood Pressure	float64	9981
4	Cholesterol Level	float64	9970
5	Exercise Habits	object	9975
6	Smoking	object	9975
7	Family Heart Disease	object	9979
8	Diabetes	object	9970
9	BMI	float64	9978
10	High Blood Pressure	object	9974
11	Low HDL Cholesterol	object	9975
12	High LDL Cholesterol	object	9974
13	Alcohol Consumption	object	7414
14	Stress Level	object	9978
15	Sleep Hours	float64	9975
16	Sugar Consumption	object	9970
17	Triglyceride Level	float64	9974
18	Fasting Blood Sugar	float64	9978
19	CRP Level	float64	9974
20	Homocysteine Level	float64	9980
21	Heart Disease Status (Target)	object	10000

B. Hasil Pra-pemrosesan Data

Hasil pemeriksaan menunjukkan bahwa terdapat 2.933 baris dari total 10.000 baris data (sekitar 29,3%) yang mengandung setidaknya satu nilai kosong pada salah satu kolomnya. Alih-alih menghapus baris-baris tersebut yang akan mengurangi jumlah data secara signifikan, penelitian ini menerapkan imputasi mean untuk 9 kolom numerik dan imputasi modus untuk 12 kolom kategorikal. Setelah proses imputasi, dataset tetap memiliki 10.000 baris dan 21 kolom tanpa nilai kosong sama sekali. Seluruh kolom kategorikal selanjutnya berhasil ditransformasi menjadi nilai numerik menggunakan `LabelEncoder`, sehingga dataset siap digunakan untuk tahap pemodelan.

C. Hasil Pembagian Data Latih dan Data Uji

Dataset yang telah bersih dan terencode kemudian dibagi menjadi data latih dan data uji dengan proporsi 80:20. Hasil pembagian menghasilkan 8.000 data latih dan 2.000 data uji. Proporsi ini dipilih agar model memiliki data yang cukup banyak untuk belajar pola pada tahap pelatihan, sekaligus tetap menyisakan data uji yang representatif untuk mengevaluasi kemampuan generalisasi model terhadap data baru yang belum pernah dilihat sebelumnya.

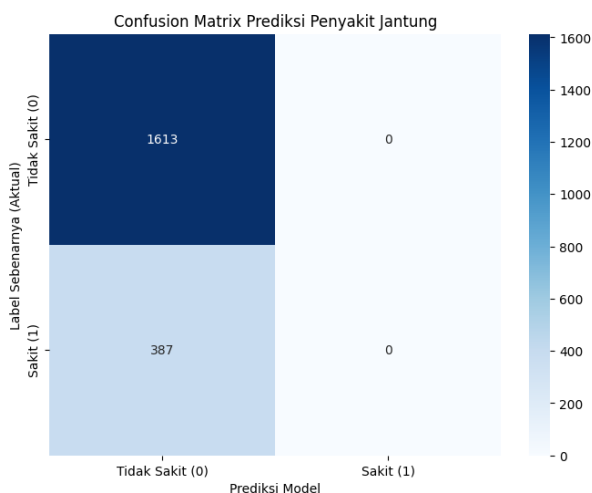
D. Hasil Pelatihan dan Evaluasi Model

Model RandomForestClassifier dengan parameter $n_estimators=100$, $criterion='entropy'$, dan $random_state=42$ berhasil dilatih menggunakan 8.000 data latih. Ketika diuji terhadap 2.000 data uji, model menghasilkan akurasi keseluruhan (overall accuracy) sebesar 80,65%. Meskipun angka ini terlihat cukup tinggi, hasil laporan klasifikasi lengkap pada Tabel II menunjukkan gambaran yang jauh berbeda ketika ditinjau per kelas.

TABEL II
HASIL CLASSIFICATION REPORT MODEL RANDOM FOREST

Kelas	Precisio n	Recall	F1- Score	Suppor t
0 (Tidak Sakit)	0,81	1,00	0,89	1613
1 (Sakit)	0,00	0,00	0,00	387
Accuracy			0,81	2000
Macro avg	0,40	0,50	0,45	2000
Weighted avg	0,65	0,81	0,72	2000

Tabel II memperlihatkan bahwa untuk kelas 0 (Tidak Sakit) model memperoleh precision 0,81 dan recall 1,00, yang berarti model memprediksi seluruh 1.613 data uji berlabel Tidak Sakit dengan benar. Namun untuk kelas 1 (Sakit), model memperoleh precision, recall, dan F1-score sebesar 0,00, yang berarti model sama sekali tidak berhasil mengenali satu pun dari 387 pasien yang sebenarnya menderita penyakit jantung pada data uji.



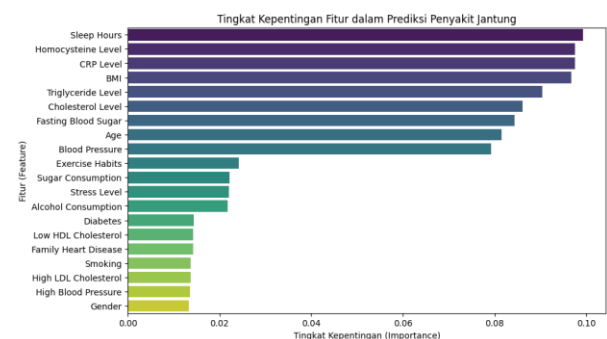
Gambar 2. Confusion matrix hasil prediksi model Random Forest pada data uji

Hal ini dikonfirmasi oleh confusion matrix pada Gambar 2, yang menunjukkan bahwa dari 1.613 data aktual kelas Tidak Sakit, seluruhnya diprediksi benar sebagai Tidak Sakit (True Negative = 1613, False Positive = 0). Sebaliknya, dari 387 data aktual kelas Sakit, seluruhnya justru diprediksi sebagai Tidak Sakit (False Negative = 387, True Positive = 0). Dengan kata lain, model yang dihasilkan pada dasarnya hanya memprediksi kelas mayoritas (Tidak Sakit) untuk seluruh data uji, sehingga akurasi 80,65% yang terlihat tinggi sesungguhnya menyesatkan karena model tidak memiliki kemampuan apa pun dalam mendeteksi pasien yang benar-benar berisiko terkena penyakit jantung.

Fenomena ini merupakan indikasi kuat dari permasalahan class imbalance, yaitu ketidakseimbangan jumlah data antar kelas pada dataset. Pada data uji, proporsi kelas Tidak Sakit (1.613 data atau 80,65%) jauh lebih besar dibandingkan kelas Sakit (387 data atau 19,35%), sehingga model Random Forest yang dilatih tanpa penanganan khusus terhadap ketidakseimbangan kelas cenderung bias dan lebih menguntungkan untuk selalu memprediksi kelas mayoritas guna meminimalkan kesalahan secara keseluruhan. Temuan ini sejalan dengan peringatan yang disampaikan oleh Khalilia dkk. [4] bahwa prediksi risiko penyakit dari data yang sangat tidak seimbang memerlukan penanganan khusus agar model tidak hanya mengoptimalkan akurasi keseluruhan semata.

E. Analisis Feature Importance

Untuk memahami atribut apa saja yang paling berpengaruh terhadap prediksi model, dilakukan analisis feature importance menggunakan atribut `feature importances` dari model Random Forest yang telah dilatih. Gambar 3 menampilkan visualisasi tingkat kepentingan seluruh 20 fitur yang digunakan dalam pemodelan, diurutkan dari yang paling berpengaruh hingga yang paling kecil pengaruhnya.



Gambar 3. Tingkat kepentingan fitur (feature importance) dalam prediksi penyakit jantung

TABEL III
SEPULUH FITUR TERPENTING BERDASARKAN FEATURE IMPORTANCE

No	Fitur	Importanc e
1	Sleep Hours	0,0993
2	Homocysteine Level	0,0976
3	CRP Level	0,0975

4	BMI	0,0967
5	Triglyceride Level	0,0904
6	Cholesterol Level	0,0861
7	Fasting Blood Sugar	0,0844
8	Age	0,0815
9	Blood Pressure	0,0792
10	Exercise Habits	0,0242

Tabel III menunjukkan bahwa lima atribut dengan pengaruh terbesar terhadap prediksi penyakit jantung pada model ini secara berurutan adalah Sleep Hours (0,0993), Homocysteine Level (0,0976), CRP Level (0,0975), BMI (0,0967), dan Triglyceride Level (0,0904). Menariknya, seluruh lima atribut numerik teratas ini berkaitan dengan kondisi klinis dan gaya hidup pasien, sedangkan atribut kategorikal seperti Exercise Habits, Smoking, dan Gender memiliki tingkat kepentingan yang jauh lebih rendah, masing-masing di bawah 0,025. Terdapat kesenjangan importance yang cukup besar antara sembilan atribut numerik teratas (Sleep Hours hingga Blood Pressure, dengan nilai importance antara 0,079 hingga 0,099) dengan atribut-atribut kategorikal sisanya. Hal ini mengindikasikan bahwa Random Forest cenderung lebih mengandalkan atribut numerik yang memiliki rentang nilai kontinu untuk melakukan pemisahan (split) pada setiap pohon keputusan.

F. Pembahasan

Hasil penelitian ini memperlihatkan sebuah temuan penting bahwa akurasi tinggi tidak selalu mencerminkan kualitas model klasifikasi yang baik, khususnya pada kasus medis dengan distribusi kelas yang tidak seimbang. Meskipun akurasi 80,65% yang diperoleh sekilas sebanding dengan hasil penelitian sebelumnya, seperti Suleiman dkk. [7] dan Pradini dkk. [5] yang melaporkan akurasi Random Forest di atas 85% pada dataset heart disease lainnya, evaluasi lebih mendalam pada penelitian ini menunjukkan bahwa model gagal total dalam mendeteksi kelas Sakit (recall 0%). Perbandingan ini menegaskan bahwa dataset dan tingkat ketidakseimbangan kelas sangat memengaruhi performa aktual model, sehingga metrik akurasi saja tidak cukup untuk menilai kelayakan sebuah model prediksi penyakit sebagaimana ditekankan pada beberapa penelitian rujukan [4], [9].

Dari sisi praktis, model yang hanya memprediksi kelas mayoritas jelas tidak dapat digunakan sebagai alat bantu diagnosis, karena tujuan utama sistem prediksi penyakit jantung justru adalah mengenali pasien yang berisiko sedini mungkin. Oleh karena itu, penelitian selanjutnya perlu mempertimbangkan teknik penanganan class imbalance seperti oversampling pada kelas minoritas menggunakan SMOTE (Synthetic Minority Over-sampling Technique), undersampling pada kelas mayoritas, maupun pemberian bobot kelas (class_weight) pada saat pelatihan model Random Forest. Selain itu, hasil analisis feature importance pada penelitian ini, yang menunjukkan bahwa Sleep Hours, Homocysteine Level, CRP Level, BMI, dan Triglyceride Level merupakan atribut paling berpengaruh, dapat menjadi acuan bagi tenaga medis maupun peneliti selanjutnya untuk

memprioritaskan pemeriksaan atribut-atribut tersebut dalam upaya deteksi dini risiko penyakit jantung.

IV. KESIMPULAN

Penelitian ini menerapkan algoritma Random Forest untuk memprediksi status penyakit jantung menggunakan dataset heart_disease.csv yang terdiri atas 10.000 data pasien dengan 21 atribut. Setelah melalui tahap pembersihan data dengan imputasi mean dan modus, transformasi data kategorikal menggunakan Label Encoding, serta pembagian data latih dan data uji dengan proporsi 80:20, model RandomForestClassifier dengan 100 estimator dan kriteria entropy berhasil dilatih dan menghasilkan akurasi keseluruhan sebesar 80,65% pada data uji.

Namun demikian, evaluasi yang lebih mendalam menggunakan precision, recall, F1-score, dan confusion matrix mengungkap bahwa model yang dihasilkan sama sekali tidak mampu mengenali pasien yang benar-benar menderita penyakit jantung, dengan nilai recall 0% pada kelas Sakit. Hal ini menunjukkan bahwa model mengalami bias terhadap kelas mayoritas akibat ketidakseimbangan kelas pada dataset, sehingga akurasi yang tinggi tidak dapat dijadikan satu-satunya tolok ukur keberhasilan model prediksi penyakit. Hasil ini menegaskan pentingnya evaluasi model klasifikasi medis secara menyeluruh, tidak hanya berdasarkan akurasi.

Analisis feature importance menunjukkan bahwa Sleep Hours, Homocysteine Level, CRP Level, BMI, dan Triglyceride Level merupakan lima atribut yang paling berpengaruh terhadap prediksi penyakit jantung pada dataset ini. Temuan ini dapat menjadi masukan bagi penelitian selanjutnya maupun tenaga medis dalam menentukan prioritas pemeriksaan kesehatan. Sebagai rekomendasi, penelitian selanjutnya disarankan untuk menerapkan teknik penanganan class imbalance seperti SMOTE, undersampling, atau pembobotan kelas (class_weight) pada algoritma Random Forest agar diperoleh model prediksi penyakit jantung yang benar-benar andal dalam mengenali pasien berisiko, bukan hanya unggul pada metrik akurasi semata.

V. REFERENSI

- [1] S. S. Rawat, A. K. Mishra, and D. Motwani, "Forecasting the Survival and Mortality of Patients by Machine Learning Trained on Heart Failure Clinical Imbalanced Data," *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2024*, 2024, doi:org
- [2] A. Rahman, S. Zaman, S. M. K. Parvej, and H. M. Abdul Fattah, "Heart Disease Prediction Using Ensemble Solutions and Exploring the Effect of Target Engineering with Pearson Correlation Based Feature Selection," *Proceedings - 6th International Conference on Electrical Engineering and Information and Communication Technology*,

- ICEEICT 2024, pp. 1407–1412, 2024, doi.org: 10.1109/ICEEICT62016.2024.10534517.
- [3] G. M. Kumar, L. A. Abdulaziz Alharbi, A. A. Imam, S. Bin Basri, and M. H. B. Hassan, *Cardiovascular Diseases Prediction Using Advanced Machine Learning Algorithms on Cleveland Dataset*, vol. 1417 LNEE. 2025. doi: 10.1007/978-981-96-5848-0_33.
- [4] K. Elangovan and K. N. S. S. K. Aditya, “Heart Disease Detection Using Modified Random Classifier,” *2025 International Conference on Information, Implementation, and Innovation in Technology, I2ITCON 2025*, 2025, doi: 10.1109/I2ITCON65200.2025.11210647.
- [5] Z. Shahbazi and S. Nowaczyk, “Towards personalized cardiometabolic risk prediction: A fusion of exposome and AI,” *Heliyon*, vol. 11, no. 1, 2025, doi: 10.1016/j.heliyon.2024.e40859.
- [6] J. Kukreja and R. C. Poonia, *Machine learning for healthcare*, vol. 143. 2026. doi: 10.1016/bs.adcom.2026.01.005.
- [7] S. R. Holagh and B. Xu, “A Meta-Learning Model for Mortality Prediction in Patients with Chronic Cardiovascular Disease,” *CMES - Computer Modeling in Engineering and Sciences*, 2025, doi: 10.32604/cmes.2025.072259.
- [8] M. I. Hossain, M. H. Maruf, M. A. R. Khan, M. S. Sabbir Ejaz, and M. A. S. Khan, “Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison,” *Iran Journal of Computer Science*, vol. 6, no. 4, pp. 397–417, 2023, doi: 10.1007/s42044-023-00148-7.
- [9] R. B. Z. Auvée, M. F. I. Fuad, Y. Hossain, S. R. Ahmed, and H. A. Karim, “Cardiovascular diagnostic detection using multilayer perceptron approach,” *Systems and Soft Computing*, vol. 8, 2026, doi: 10.1016/j.sasc.2026.200490.
- [10] M. Rahmathullah, S. Nagakishore Bhavanam, and M. V. Devi, “Predictive modeling of Cardiovascular Disorder: A comparison of Regression model and Random Forest,” doi: 10.1117/12.3070257.
- [11] S. Suzuki *et al.*, “Comparison of risk models for mortality and cardiovascular events between machine learning and conventional logistic regression analysis,” *PLoS One*, 2019, doi: 10.1371/journal.pone.0221911.
- [12] N. Komal Kumar, R. Govindarajan, S. Sathya, J. Senthil Murugan, and K. V Bindu, “Comparative analysis of explainable machine learning models for cardiovascular risk stratification using clinical data and shapley additive explanations,” *Intell. Based. Med.*, vol. 12, 2025, doi: 10.1016/j.ibmed.2025.100286.
- [13] K. Ng, S. R. Steinhubl, C. R. DeFilippi, S. Dey, and W. F. Stewart, “Early Detection of Heart Failure Using Electronic Health Records: Practical Implications for Time before Diagnosis, Data Diversity, Data Quantity, and Data Density,” *Circ. Cardiovasc. Qual. Outcomes*, vol. 9, no. 6, pp. 649–658, 2016, doi: 10.1161/CIRCOUTCOMES.116.002797.
- [14] B. M. Yankam, “Bridging the gap: methodological challenges and innovations in systematic reviews of machine learning models in healthcare,” *Front. Artif. Intell.*, 2026, doi: 10.3389/fraci.2026.1739192.
- [15] N. S. Alghamdi, M. Zakariah, A. Shankar, and W. G. Viriyasitavat, “Heart disease prediction using autoencoder and DenseNet architecture,” *Egyptian Informatics Journal*, vol. 28, 2024, doi: 10.1016/j.eij.2024.100559.
- [16] K. S. Divya, M. Mohebbanaaz, and G. Fayaz Hussain, “Machine Learning in Cardiology: Assessing Model Reliability for Real world Deployment,” *2025 IEEE 7th International Conference on Computing, Communication and Automation, ICCCA 2025*, 2025, doi: 10.1109/ICCCA66364.2025.11325131.
- [17] N. Habashneh, M. Alwars, M. R. AlShehhi, and M. A. Al Betar, “Comparative Analysis of Machine Learning Models for Heart Disease Prediction Using the Cleveland Dataset,” *2025 10th International Conference on Information Technology Trends, ITT 2025*, pp. 158–163, 2025, doi: 10.1109/ITT69610.2025.11352907.
- [18] M. F. H. Fahim, M. A. M. Hassan, S. M. Karpagavalli, R. Jha, and T. R. Mahesh, “Advancing Cardiovascular Health Prediction: Machine Learning Algorithm Analysis,” *Proceedings - 2024 13th IEEE International Conference on Communication Systems and Network Technologies, CSNT 2024*, pp. 588–593, 2024, doi: 10.1109/CSNT60213.2024.10546094.