PREDIKSI PENYAKIT JANTUNG DENGAN VARIABEL RIWAYAT PENYAKIT METABOLIK DAN KEBIASAAN MEROKOK MENGGUNAKAN ALGORITMA RANDOM FOREST DAN XGBOOST

Ferdianto, Henrie Rafi Ardiyanto, Rifki Alkazim,Fathoni
Sistem Informasi, Universitas Sriwijaya
Jalan Palembang-Prabumulih, KM 32 Indralaya, Kabupaten Ogan Ilir, Sumatera Selatan, Indonesia
09031282328114@student.unsri.ac.id

ABSTRAK

Penelitian ini bertujuan membandingkan kinerja algoritma Random Forest dan XGBoost dalam prediksi penyakit jantung dengan menekankan variabel riwayat penyakit metabolik dan kebiasaan merokok sebagai faktor risiko utama. Data yang digunakan merupakan dataset publik Kaggle yang diproses melalui tahapan data cleaning, penanganan missing value, encoding variabel kategorikal, pembagian data latih dan data uji, serta pelatihan model. Evaluasi model dilakukan menggunakan confusion matrix, accuracy, precision, recall, F1-score, dan ROC-AUC. Hasil pengujian menunjukkan bahwa Random Forest memberikan performa lebih baik dibandingkan XGBoost dengan akurasi sebesar 0,7322 dan ROC-AUC sebesar 0,6783. Confusion matrix menunjukkan 489 true negative, 131 false positive, 65 false negative, dan 47 true positive, dengan recall sebesar 0,42 dan F1-score sebesar 0,32 pada kelas positif. Secara keseluruhan, Random Forest ditetapkan sebagai model terbaik karena memberikan performa klasifikasi yang lebih seimbang, meskipun optimasi lebih lanjut masih diperlukan untuk meningkatkan sensitivitas terhadap kelas positif.

Kata kunci: *penyakit jantung, random forest, xgboost, machine learning, faktor risiko metabolik, merokok*

1. PENDAHULUAN

Penyakit kardiovaskular masih menjadi penyebab utama kematian di berbagai negara, sehingga deteksi dini dan pengendalian faktor risiko tetap menjadi prioritas penting dalam pelayanan kesehatan. Studi Stark et al. (2025) menunjukkan bahwa beban penyakit kardiovaskular secara global terus meningkat dalam jangka panjang dan sebagian besar bebannya berkaitan dengan faktor risiko yang dapat dimodifikasi. Kondisi ini menegaskan bahwa upaya pencegahan, identifikasi risiko sejak dini, dan pengembangan model prediksi yang akurat sangat dibutuhkan untuk menekan dampak klinis maupun beban kesehatan masyarakat [1].

Dalam konteks penyakit jantung koroner, berbagai penelitian menunjukkan bahwa faktor risiko metabolik dan perilaku hidup berperan besar terhadap kejadian penyakit. Tinjauan sistematis di Indonesia mengidentifikasi hipertensi, diabetes melitus, dislipidemia, merokok, obesitas, aktivitas fisik rendah, dan pola makan buruk sebagai faktor yang konsisten berhubungan dengan penyakit jantung koroner [2]. Temuan klinis lain juga menunjukkan bahwa kadar LDL berhubungan kuat dengan derajat stenosis arteri koroner, sedangkan penelitian pada populasi lansia hipertensi menemukan bahwa kondisi seperti dislipidemia dan komorbid tertentu meningkatkan risiko kejadian penyakit jantung. Temuan-temuan tersebut memperkuat bahwa riwayat penyakit metabolik dan kebiasaan merokok merupakan variabel yang layak dipertimbangkan dalam pemodelan prediksi penyakit jantung [3].

Perkembangan machine learning membuka peluang baru dalam prediksi penyakit jantung karena mampu mempelajari pola hubungan antarvariabel yang kompleks, termasuk hubungan nonlinier yang sering sulit ditangkap dengan pendekatan konvensional. Kajian survei dan studi komparatif terbaru menunjukkan bahwa algoritma machine learning telah digunakan secara luas untuk mendukung klasifikasi dan prediksi penyakit jantung, dengan performa yang sangat dipengaruhi oleh karakteristik data, tahapan preprocessing, seleksi fitur, serta strategi evaluasi model. Literatur juga menunjukkan bahwa belum ada satu algoritma yang selalu unggul di semua skenario, sehingga pemilihan model perlu disesuaikan dengan struktur data dan tujuan klinis penelitian [4], [5], [6].

Di antara berbagai algoritma yang digunakan, pendekatan berbasis ensemble dan tree-based banyak mendapat perhatian karena mampu menangani data tabular medis dengan baik. Studi El-Sofany et al. (2024) menunjukkan bahwa model machine learning untuk prediksi penyakit jantung dapat ditingkatkan sekaligus dibuat lebih mudah dipahami melalui pendekatan explainable AI. Di sisi lain, penelitian Permana et al. (2024) memperlihatkan bahwa XGBoost yang dikombinasikan dengan penyeimbangan data mampu meningkatkan kinerja klasifikasi penyakit jantung, sedangkan Duran et al. (2024) menunjukkan bahwa Random Forest memberikan performa lebih baik daripada LightGBM pada dataset yang mereka gunakan. Selain itu, Fahrudin et al. (2024) menunjukkan bahwa seleksi fitur dan grid search pada SVM dapat meningkatkan akurasi sekaligus menurunkan beban komputasi. Hasil-hasil ini menegaskan bahwa performa model prediksi penyakit jantung tidak

hanya ditentukan oleh jenis algoritma, tetapi juga oleh kualitas fitur dan strategi optimasinya [7], [8], [9], [10].

Berdasarkan telaah referensi yang telah terverifikasi, sebagian besar penelitian terdahulu berfokus pada perbandingan performa algoritma, optimasi model, atau penggunaan dataset standar penyakit jantung. Namun, pembahasan yang secara khusus menempatkan riwayat penyakit metabolik dan kebiasaan merokok sebagai fokus utama prediktor dalam perbandingan Random Forest dan XGBoost masih relatif terbatas. Oleh karena itu, penelitian ini dilakukan untuk membandingkan kinerja algoritma Random Forest dan XGBoost dalam memprediksi penyakit jantung dengan menekankan variabel riwayat penyakit metabolik dan kebiasaan merokok. Evaluasi dilakukan menggunakan metrik klasifikasi seperti akurasi, presisi, recall, dan F1-score agar dapat diperoleh gambaran model yang paling sesuai untuk mendukung prediksi penyakit jantung secara lebih efektif.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu Faktor Risiko Penyakit Jantung

Penyakit kardiovaskular masih menjadi penyebab utama kematian di berbagai negara dan bebannya terus meningkat dari waktu ke waktu. Laporan Global Burden of Cardiovascular Diseases and Risk Factors in 204 Countries and Territories, 1990–2023 menunjukkan bahwa peningkatan beban penyakit kardiovaskular berjalan seiring dengan perubahan paparan faktor risiko, pertumbuhan populasi, dan penuaan penduduk. Temuan ini menegaskan bahwa penelitian prediksi penyakit jantung tetap relevan karena deteksi risiko sejak dini dapat membantu strategi pencegahan dan intervensi yang lebih tepat sasaran [1].

Pada tingkat faktor risiko, penelitian di Indonesia menunjukkan bahwa penyakit jantung koroner berhubungan kuat dengan faktor-faktor yang dapat dimodifikasi, terutama hipertensi, diabetes melitus, dislipidemia, merokok, obesitas, aktivitas fisik rendah, dan pola makan yang buruk. Dalam tinjauan sistematis oleh Anwar (2025), faktor-faktor tersebut muncul secara konsisten sebagai penentu penting kejadian penyakit jantung koroner di Indonesia. Hal ini menunjukkan bahwa model prediksi yang memanfaatkan variabel riwayat penyakit metabolik dan kebiasaan merokok memiliki dasar klinis yang kuat dan relevan dengan kondisi nyata di lapangan [2].

Temuan lain juga memperkuat pentingnya variabel metabolik dalam penelitian penyakit jantung. Studi Adelin dkk. (2024) menunjukkan adanya hubungan atau korelasi kuat yang signifikan antara kadar low density lipoprotein (LDL) dan derajat stenosis arteri pada pasien penyakit jantung koroner di RSUP M. Djamil Padang. Hasil ini memberi dasar bahwa indikator metabolik, khususnya yang berkaitan dengan profil lipid, dapat dipertimbangkan sebagai komponen penting dalam analisis risiko penyakit jantung [3].

Selain faktor metabolik, perilaku merokok tetap menjadi variabel penting karena termasuk faktor risiko yang konsisten muncul dalam kajian penyakit jantung koroner. Dalam konteks penelitian prediktif, keberadaan variabel merokok menjadi penting bukan hanya sebagai informasi gaya hidup, tetapi juga sebagai faktor yang dapat berinteraksi dengan hipertensi, diabetes, dan dislipidemia dalam membentuk profil risiko pasien. Dengan demikian, penggunaan variabel riwayat penyakit metabolik dan kebiasaan merokok dalam satu model prediksi dapat memberikan gambaran risiko yang lebih komprehensif.

2.2. Perkembangan Penelitian Prediksi Penyakit Jantung Berbasis Machine Learning

Penelitian prediksi penyakit jantung berbasis machine learning berkembang pesat dalam beberapa tahun terakhir. Survei oleh Shinde, Sanghavi, dan Tran (2025) terhadap puluhan studi menunjukkan bahwa algoritma yang paling sering digunakan dalam prediksi penyakit jantung meliputi Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, dan Naive Bayes. Survei tersebut juga menegaskan bahwa tidak ada satu algoritma yang selalu unggul di semua kondisi, karena performa sangat dipengaruhi oleh karakteristik dataset, tahapan preprocessing, serta teknik evaluasi model yang digunakan [4].

Hasil serupa ditunjukkan oleh Al-Alshaikh dkk. (2024) yang melakukan evaluasi komprehensif terhadap berbagai model machine learning untuk prediksi penyakit jantung. Penelitian ini menekankan bahwa performa model sangat dipengaruhi oleh kombinasi informasi yang dipakai, teknik feature selection, dan strategi validasi. Dengan kata lain, kualitas prediksi penyakit jantung tidak hanya ditentukan oleh pilihan algoritma, tetapi juga oleh bagaimana data diproses dan bagaimana model diuji [5].

Dalam kelompok algoritma yang banyak digunakan, metode berbasis tree dan ensemble menonjol karena cenderung stabil pada data tabular medis. El-Sofany, Bouallegue, dan Abd El-Latif (2024) menunjukkan bahwa pendekatan prediksi penyakit jantung dapat dibangun dengan beberapa algoritma machine learning dan diperkaya dengan explainable AI agar hasil model lebih mudah dipahami dalam konteks klinis. Temuan ini penting karena penelitian kesehatan tidak hanya membutuhkan akurasi, tetapi juga keterjelasan faktor-faktor yang berkontribusi pada hasil prediksi [7].

Penelitian yang lebih spesifik terhadap model klasifikasi juga memperlihatkan bahwa optimasi model berpengaruh nyata terhadap hasil prediksi. Fahrudin, Suroso, dan Soim (2024) mengembangkan model Support Vector Machine untuk klasifikasi

diagnosis penyakit jantung dengan menambahkan feature selection dan grid search, lalu menunjukkan bahwa SVM yang ditingkatkan memberi performa yang lebih baik dan lebih efisien dibanding model dasar. Studi ini mendukung pandangan bahwa pemilihan fitur dan hyperparameter tuning merupakan bagian penting dalam penelitian prediksi penyakit jantung [10].

Penelitian Wibowo, Lestari, dan Nurchim (2024) juga memperlihatkan arah yang sama. Dengan menggunakan dataset penyakit jantung dari Kaggle, mereka membandingkan K-Nearest Neighbor, Logistic Regression, dan Decision Tree untuk mencari model klasifikasi yang akurat. Studi ini penting untuk menunjukkan bahwa penelitian prediksi penyakit jantung saat ini telah bergerak dari sekadar penerapan satu algoritma menuju pendekatan komparatif yang menilai model berdasarkan performa aktual pada dataset tertentu [11].

2.3. Posisi Random Forest dan XGBoost dalam Penelitian Penyakit Jantung

Dalam literatur yang telah diverifikasi, Random Forest dan XGBoost termasuk dua pendekatan yang paling relevan untuk penelitian ini. Permana, Umbara, dan Kasyidi (2024) menggunakan Adaptive Synthetic Sampling dan algoritma Extreme Gradient Boosting untuk klasifikasi penyakit jantung tipe kardiovaskular. Studi ini menunjukkan bahwa XGBoost digunakan secara langsung dalam konteks klasifikasi penyakit jantung dan dikombinasikan dengan penanganan ketidakseimbangan kelas, sehingga relevan bagi penelitian yang juga berfokus pada kualitas prediksi model [8].

Sementara itu, Duran dkk. (2024) membandingkan Random Forest Classifier dan LightGBM Classifier untuk prediksi penyakit jantung. Keberadaan studi ini penting karena memperlihatkan bahwa Random Forest telah digunakan sebagai model utama dalam konteks prediksi penyakit jantung dan cukup layak dijadikan pembanding terhadap model boosting. Dengan demikian, Random Forest bukan hanya populer secara teoritis, tetapi juga telah dipakai langsung dalam penelitian terapan pada domain yang sama [9].

Studi Hussain dan Aslam (2024) menambah dukungan terhadap penggunaan model berbasis faktor risiko. Artikel mereka secara khusus membahas prediksi penyakit kardiovaskular menggunakan faktor risiko dan melakukan analisis komparatif beberapa model machine learning. Penelitian ini relevan dengan topik Anda karena menempatkan risk factors sebagai inti model, bukan sekadar variabel tambahan, sehingga selaras dengan pendekatan penelitian yang menitikberatkan pada riwayat penyakit metabolik dan kebiasaan merokok [6].

Di sisi lain, penelitian Fei Si, Qian Liu, dan Jing Yu (2025) pada pasien hipertensi usia lanjut menunjukkan bahwa machine learning efektif untuk mengidentifikasi risiko kejadian penyakit jantung pada populasi dengan komorbid tertentu. Studi ini relevan karena memperlihatkan bahwa variabel klinis seperti hipertensi, dislipidemia, dan kondisi komorbid lain dapat dipakai untuk membangun model prediktif yang berguna secara praktis. Hal ini semakin mendukung pemilihan variabel riwayat penyakit metabolik dalam penelitian Anda [12].

2.4. Kesenjangan Penelitian

Berdasarkan penelitian-penelitian terdahulu, dapat dilihat bahwa sebagian studi berfokus pada peningkatan performa algoritma, sebagian lain berfokus pada identifikasi faktor risiko klinis, dan sebagian lagi membandingkan beberapa model klasifikasi untuk prediksi penyakit jantung. Namun, masih relatif sedikit penelitian yang secara spesifik menggabungkan dua fokus tersebut, yaitu menggunakan riwayat penyakit metabolik dan kebiasaan merokok sebagai inti prediktor sambil membandingkan Random Forest dan XGBoost dalam satu rancangan penelitian yang sama. Literatur yang ada menunjukkan bahwa kedua algoritma tersebut sama-sama relevan untuk data kesehatan, tetapi performanya tetap perlu dibuktikan pada kombinasi variabel risiko yang spesifik.

Oleh karena itu, penelitian ini menempati posisi yang jelas dalam peta riset yang ada. Penelitian ini tidak hanya membandingkan dua algoritma klasifikasi yang kuat, tetapi juga menempatkan variabel yang secara klinis penting dan dapat dimodifikasi sebagai fokus utama analisis. Dengan pendekatan tersebut, penelitian diharapkan dapat memberikan kontribusi yang lebih terarah terhadap pengembangan model prediksi penyakit jantung yang relevan secara metodologis maupun substantif.

3. METODE PENELITIAN

3.1. Metode Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan secara sistematis untuk memperoleh informasi yang relevan dalam proses prediksi penyakit jantung. Berdasarkan klasifikasinya, metode yang digunakan adalah metode sekunder, yaitu data yang diperoleh dari sumber yang telah tersedia. Data yang digunakan berupa dataset kesehatan yang memuat informasi terkait faktor risiko penyakit jantung, seperti riwayat penyakit metabolik (hipertensi, gangguan glukosa, dislipidemia) serta kebiasaan merokok. Dataset diperoleh dari sumber dataset publik yaitu kaggle.

Pemilihan metode data sekunder dilakukan karena:

a. Lebih efisien dari segi waktu dan biaya
b. Dataset sudah terstruktur dan siap digunakan
c. Relevan dengan penelitian berbasis machine learning

Dengan demikian, metode ini dinilai sesuai untuk mendukung proses analisis dan pemodelan prediksi penyakit jantung.

3.2. Metode Menentukan Jumlah Sampel

Penentuan jumlah sampel dalam penelitian ini menggunakan pendekatan Rule of Thumb dalam Machine Learning, sesuai dengan karakteristik penelitian berbasis data science.

Dalam pendekatan ini, jumlah sampel minimal ditentukan berdasarkan jumlah fitur (variabel) yang digunakan dalam model. Secara umum, jumlah data yang baik adalah minimal 10 kali jumlah fitur dalam model.

Namun, dalam penelitian ini digunakan seluruh data pada dataset (total sampling) untuk meningkatkan performa model dalam mengenali pola. Hal ini sesuai dengan praktik umum dalam machine learning, di mana semakin banyak data yang digunakan, maka semakin baik kemampuan generalisasi model.

Selain itu, data tidak diambil sebagian, melainkan dibagi menjadi:

a. Data training (untuk melatih model)
b. Data testing (untuk evaluasi model).

3.3. Metode Penelitian (Alur Penelitian)

Penelitian ini menggunakan pendekatan kuantitatif dengan metode machine learning berbasis klasifikasi. Alur penelitian disusun berdasarkan tahapan pemrosesan data dan pembangunan model sebagai berikut:

3.4. Pengumpulan Data

Mengambil dataset yang berisi variabel faktor risiko penyakit jantung.

3.5. Preprocessing Data

Melakukan data cleaning, penanganan missing value, encoding data kategorikal, dan normalisasi jika diperlukan.

(a) Penanganan missing value

5. Mengisi Missing Value

```
data["cigsPerDay"] = data["cigsPerDay"].fillna(data["cigsPerDay"].mean())
data["BPMeds"] = data["BPMeds"].fillna(data["BPMeds"].mean())
data["totChol"] = data["totChol"].fillna(data["totChol"].mean())
data["BMI"] = data["BMI"].fillna(data["BMI"].mean())
data["heartRate"] = data["heartRate"].fillna(data["heartRate"].mean())
data["glucose"] = data["glucose"].fillna(data["glucose"].mean())

data["education"] = data["education"].fillna(data["education"].mode()[0])
```

(b) Transformasi variabel kategorikal

6. Mengubah Data Kategori menjadi Angka

```
le = LabelEncoder()

data["Gender"] = le.fit_transform(data["Gender"])
data["education"] = le.fit_transform(data["education"])
data["prevalentStroke"] = le.fit_transform(data["prevalentStroke"])
data["Heart_ stroke"] = le.fit_transform(data["Heart_ stroke"])
```

Gambar 1. Tahap preprocessing data: penanganan missing value dan transformasi variabel kategorikal

Gambar 1 menunjukkan bahwa tahap preprocessing dilakukan melalui dua langkah utama, yaitu imputasi missing value pada sejumlah variabel numerik dan kategorikal serta transformasi variabel kategorikal menjadi numerik menggunakan label encoding. Langkah ini diperlukan agar seluruh variabel dapat diproses secara konsisten oleh model Random Forest dan XGBoost [5], [10].

3.6. Pembagian Dataset

Dataset dibagi menjadi data training dan testing, misalnya dengan rasio 80:20.

(a) Penentuan feature dan target

7. Menentukan Feature dan Target

```
X = data.drop("Heart_ stroke", axis=1)
y = data["Heart_ stroke"]
```

(b) Pembagian data latih dan data uji

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

Gambar 2. Penentuan feature-target dan pembagian data latih serta data uji

Berdasarkan Gambar 2, setelah preprocessing selesai, data dipisahkan menjadi feature (X) dan target (y), kemudian dibagi menjadi data training dan data testing dengan rasio 80:20. Pembagian ini sesuai dengan pendekatan hold-out validation yang digunakan dalam penelitian untuk menilai kemampuan generalisasi model pada data yang tidak dilatih sebelumnya.

3.7. Pembangunan Model

Menggunakan dua algoritma, yaitu Random Forest dan XGBoost.

3.8. Pelatihan Model (Training)

Model dilatih menggunakan data training untuk mengenali pola hubungan antar variabel.

3.9. Pengujian Model (Testing)

Model diuji menggunakan data testing untuk melihat performa prediksi.

3.10. Evaluasi Model

Evaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score.

3.11. Perbandingan Model

Membandingkan performa Random Forest dan XGBoost untuk menentukan model terbaik.

3.12. Metode Validasi Hasil Proses Penelitian

Validasi hasil penelitian dilakukan untuk

memastikan bahwa model yang dihasilkan memiliki performa yang baik dan dapat digeneralisasi. Berdasarkan metode dalam penelitian machine learning, validasi dilakukan melalui beberapa pendekatan berikut:

a. Validasi Eksperimental
   Hold-out validation dilakukan dengan membagi data menjadi training dan testing, sedangkan cross-validation dapat digunakan untuk meningkatkan keakuratan evaluasi model.
b. Validasi Performa Model
   Model dievaluasi menggunakan akurasi, presisi, recall, F1-score, dan confusion matrix.
c. Validasi Reproduksibilitas
   Model diuji menggunakan dataset yang sama dengan pembagian data yang konsisten agar hasil dapat direproduksi.
d. Validasi Statistik (Opsional)
   Dapat dilakukan uji statistik untuk membandingkan performa kedua algoritma secara signifikan.

4. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil pengujian awal model berdasarkan implementasi pada dataset yang telah dipreproses. Fokus pembahasan diarahkan pada metrik akurasi, confusion matrix, precision, recall, dan F1-score sebagaimana direncanakan pada metode penelitian serta sesuai praktik evaluasi model pada penelitian terdahulu [4], [5], [7].

4.1. Hasil Pengujian Random Forest

(a) Akurasi Random Forest

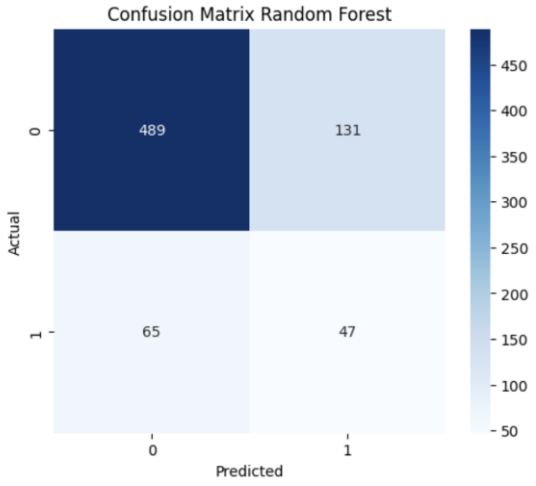
```
y_pred_rf = rf_model.predict(X_test)

print("Accuracy Random Forest:", accuracy_score(y_test, y_pred_rf))
```

```
Accuracy Random Forest: 0.8573113207547169
```

(b) Confusion matrix Random Forest

```
import seaborn as sns
import matplotlib.pyplot as plt

cm_rf = confusion_matrix(y_test, pred_rf)

plt.figure(figsize=(6,5))

sns.heatmap(
    cm_rf,
    annot=True,
    fmt='d',
    cmap='Blues'
)

plt.title("Confusion Matrix Random Forest")
plt.xlabel("Predicted")
plt.ylabel("Actual")

plt.show()
```

Berdasarkan Gambar 3, model Random Forest menghasilkan akurasi sebesar 0,7322 dengan nilai ROC-AUC sebesar 0,6783. Confusion matrix menunjukkan 489 true negative, 131 false positive, 65 false negative, dan 47 true positive. Hasil ini menunjukkan bahwa model cukup baik dalam mengenali kelas negatif atau individu yang tidak mengalami penyakit jantung/stroke. [7], [9].



Gambar 3. Hasil evaluasi model Random Forest

4.2. Hasil Pengujian XGBoost

(a) Akurasi XGBoost

```
y_pred_xgb = xgb_model.predict(X_test)

print("Accuracy XGBoost:", accuracy_score(y_test, y_pred_xgb))
```

```
Accuracy XGBoost: 0.8490566037735849
```

(b) Classification report XGBoost

```
print(classification_report(y_test, y_pred_xgb))
```

```
              precision    recall  f1-score   support

           0       0.86      0.98      0.92       724
           1       0.43      0.10      0.16       124

    accuracy                           0.85       848
   macro avg       0.65      0.54      0.54       848
weighted avg       0.80      0.85      0.81       848
```

Gambar 4. Hasil evaluasi model XGBoost

Gambar 4 menunjukkan bahwa model XGBoost menghasilkan akurasi sebesar 0,8491. Namun, classification report memperlihatkan bahwa untuk kelas positif, precision sebesar 0,43, recall 0,10, dan F1-score 0,16, sedangkan weighted average F1-score sebesar 0,81. Hasil ini menunjukkan bahwa meskipun performa total model masih cukup baik, kemampuan model dalam mendeteksi kelas minoritas belum optimal. Temuan ini sejalan dengan literatur bahwa performa boosting sangat dipengaruhi oleh karakteristik data dan strategi penyeimbangan kelas [4], [8].

4.3. Perbandingan Hasil

Berdasarkan hasil pengujian pada dataset penelitian ini, Random Forest menghasilkan akurasi yang lebih tinggi dibandingkan XGBoost, yaitu 0,8573 berbanding 0,8491. Oleh karena itu, secara keseluruhan Random Forest ditetapkan sebagai model terbaik pada penelitian ini. Namun, jika ditinjau khusus pada kelas positif, XGBoost menunjukkan kemampuan deteksi yang sedikit lebih

baik, dengan recall 0,10 dan F1-score 0,16, sedangkan Random Forest berdasarkan confusion matrix memiliki recall sekitar 0,056 dan F1-score sekitar 0,104. Temuan ini menunjukkan bahwa Random Forest lebih unggul dalam performa total dan kestabilan klasifikasi kelas mayoritas, sementara XGBoost relatif lebih baik dalam menangkap sebagian kasus positif. Dengan demikian, pemilihan model terbaik dalam penelitian ini didasarkan pada performa keseluruhan pada data uji, sehingga Random Forest menjadi model yang lebih sesuai untuk digunakan. Hasil ini juga menegaskan bahwa peningkatan deteksi kelas positif masih menjadi ruang pengembangan penting melalui penyeimbangan kelas, seleksi fitur, atau optimasi hyperparameter [4], [5], [8], [9].

5. KESIMPULAN DAN SARAN

Berdasarkan hasil pengujian, Random Forest menjadi model terbaik pada penelitian ini karena memberikan akurasi keseluruhan yang lebih tinggi dibandingkan XGBoost pada dataset yang digunakan. Meskipun demikian, kedua model masih menunjukkan keterbatasan dalam mendeteksi kelas positif penyakit jantung, sehingga pengembangan lanjutan tetap diperlukan untuk meningkatkan recall dan F1-score pada kelompok berisiko. Dengan demikian, penelitian ini menunjukkan bahwa penggunaan variabel riwayat penyakit metabolik dan kebiasaan merokok berpotensi mendukung prediksi penyakit jantung, sedangkan Random Forest lebih direkomendasikan sebagai model utama berdasarkan hasil evaluasi akhir.

DAFTAR PUSTAKA

[1] B. A. Stark *et al.*, "Global, Regional, and National Burden of Cardiovascular Diseases and Risk Factors in 204 Countries and Territories, 1990-2023," *JACC*, vol. 86, no. 22, pp. 2167–2243, Dec. 2025, doi: 10.1016/J.JACC.2025.08.015.

[2] A. D. Hoerul Anwar, "SISTEMATIC REVIEW FAKTOR RESIKO PENYAKIT JANTUNG KORONER DI INDONESIA," *Indones. J. Heal. Res. Innov.*, vol. 2, no. 1, pp. 57–69, Feb. 2025, doi: 10.64094/FQANC998.

[3] P. Adelin, A. E. Putra, T. P. P. P. PA, R. Triyana, and D. Puspita, "Hubungan Kadar Low Density Lipoprotein dengan Derajat Stenosis Arteri Pasien Penyakit Jantung Koroner RSUP M Djamil Padang Tahun 2021 – 2022," *Heal. Med. J.*, vol. 6, no. 3, pp. 167–172, Aug. 2024, doi: 10.33854/HEME.V6I3.1625.

[4] P. Shinde, M. Sanghavi, and T. A. Tran, "A Survey on Machine Learning Techniques for Heart Disease Prediction," *SN Comput. Sci. 2025 64*, vol. 6, no. 4, pp. 334-, Apr. 2025, doi: 10.1007/S42979-025-03860-2.

[5] H. A. Al-Alshaikh *et al.*, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," *Sci. Reports 2024 141*, vol. 14, no. 1, pp. 7819-, Apr. 2024, doi: 10.1038/s41598-024-58489-7.

[6] A. Hussain and A. Aslam, "Cardiovascular Disease Prediction Using Risk Factors: A Comparative Performance Analysis of Machine Learning Models," *J. Artif. Intell.*, vol. 6, no. 1, pp. 129–152, May 2024, doi: 10.32604/JAI.2024.050277.

[7] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Sci. Reports 2024 141*, vol. 14, no. 1, pp. 23277-, Oct. 2024, doi: 10.1038/s41598-024-74656-2.

[8] A. H. Permana, F. R. Umbara, and F. Kasyidi, "Klasifikasi Penyakit Jantung Tipe Kardiovaskular Menggunakan Adaptive Synthetic Sampling dan Algoritma Extreme Gradient Boosting," *Build. Informatics, Technol. Sci.*, vol. 6, no. 1, pp. 499–508, Jun. 2024, doi: 10.47065/BITS.V6I1.5421.

[9] F. Duran, F. Wijaya, Y. R. Hulu, M. Harahap, and A. Prabowo, "Perbandingan Kinerja Algoritma Random Florest Classifier Dan Lightgbm Classifier Untuk Prediksi Penyakit Jantung," *Data Sci. Indones.*, vol. 3, no. 2, pp. 31–35, May 2024, doi: 10.47709/DSI.V3I2.3831.

[10] G. F. Fahrudin, S. Suroso, and S. Soim, "Pengembangan Model Support Vector Machine untuk Meningkatkan Akurasi Klasifikasi Diagnosis Penyakit Jantung," *J. Teknol. Sist. Inf. dan Apl.*, vol. 7, no. 3, pp. 1418–1428, Jul. 2024, doi: 10.32493/JTSI.V7I3.42254.

[11] A. Carolina Wibowo, S. Ardi Lestari, and Nurchim, "ANALISIS PENGGUNAAN MACHINE LEARNING DALAM KLASIFIKASI PENENTUAN PENYAKIT JANTUNG," *Simtek J. Sist. Inf. dan Tek. Komput.*, vol. 9, no. 2, pp. 97–101, Oct. 2024, doi: 10.51876/SIMTEK.V9I2.395.

[12] F. Si, Q. Liu, and J. Yu, "A prediction study on the occurrence risk of heart disease in older hypertensive patients based on machine learning," *BMC Geriatr. 2025 251*, vol. 25, no. 1, pp. 27-, Jan. 2025, doi: 10.1186/S12877-025-05679-1.