Progresif: Jurnal Ilmiah Komputer
https://ojs.stmik-banjarbaru.ac.id/index.php/progresif/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877

Analisis Klasifikasi Risiko Penyakit Jantung Menggunakan Metode *Random Forest*

DOI: http://dx.doi.org/10.35889/progresif.v22i2.3609

Alfina Damayanti[1*], Donny Maulana[2], M. Zubair Abdurrohman[3]
Teknik Informatika, Universitas Pelita Bangsa, Bekasi, Indonesia
*e-mail *Corresponding Author:* alfina03@mhs.pelitabangsa.ac.id

Abstract

Heart disease is one of the leading causes of death worldwide, making early detection crucial to reduce the risk of complications and mortality. The advancement of machine learning technology enables fast and accurate analysis of medical data to support the diagnostic process. This study aims to develop a classification model for heart disease risk using the Random Forest algorithm. The dataset used is the Heart Disease Dataset from Kaggle, consisting of 1,025 patient records with 14 medical attributes, such as age, gender, blood pressure, cholesterol level, and maximum heart rate. The methodology applied is CRISP-DM, which includes Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Model Evaluation is conducted using a confusion matrix, cross-validation, and ROC-AUC. The results show that the Random Forest algorithm achieves a high Accuracy of 99.96% and a cross-validation score of 0.996. The variables chest pain, ca, and thalach are identified as the most influential factors in the prediction.

Keywords: *Heart Disease; Random Forest; Machine learning; Classification; CRISP-DM*

Abstrak

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia sehingga deteksi dini sangat penting untuk mengurangi risiko komplikasi dan kematian. Perkembangan teknologi *machine learning* memungkinkan analisis data medis secara cepat dan akurat dalam membantu proses diagnosis. Penelitian ini bertujuan membangun model klasifikasi risiko penyakit jantung menggunakan *Algoritma Random Forest*. *Dataset* yang digunakan adalah Heart Disease *Dataset* dari Kaggle yang terdiri dari 1025 data pasien dengan 14 atribut medis, seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, dan detak jantung maksimum. Metode yang digunakan adalah CRISP-DM meliputi *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Evaluasi model dilakukan menggunakan *confusion matrix*, *cross validation*, dan ROC-AUC. Hasil penelitian menunjukkan bahwa *Random Forest* menghasilkan akurasi tinggi dengan nilai 99,96% serta *cross validation* sebesar 0,996. Variabel chest pain, ca, dan *thalach* menjadi faktor paling berpengaruh dalam prediksi.

Kata kunci: Penyakit jantung; *Random Forest*; *Machine learning*; *Klasifikasi; CRISP-DM.*

1. Pendahuluan

Penyakit jantung masih menjadi penyebab utama kematian di dunia dan merupakan masalah kesehatan global yang sangat serius. Berdasarkan laporan Organisasi Kesehatan Dunia (WHO), penyakit kardiovaskular menyebabkan sekitar 17,9 juta kematian setiap tahun. Di Indonesia, penyakit jantung juga menempati urutan pertama penyebab kematian dengan beban pembiayaan yang sangat besar. Data BPJS Kesehatan pada tahun 2022 menunjukkan adanya sekitar 15,5 juta kasus penyakit jantung dengan total pembiayaan mencapai Rp 24,1 triliun. Tingginya angka kejadian dan dampak ekonomi tersebut menunjukkan pentingnya dilakukan penelitian dalam bidang deteksi dini dan klasifikasi risiko penyakit jantung. Upaya ini sangat krusial untuk mencegah komplikasi serius seperti gagal jantung, aritmia, hingga kematian

mendadak, sehingga diperlukan pendekatan yang lebih cepat, akurat, dan efisien dalam mengidentifikasi risiko penyakit jantung sejak dini [1].

Dalam praktiknya, proses identifikasi risiko penyakit jantung masih banyak bergantung pada pemeriksaan medis konvensional dan interpretasi tenaga kesehatan. Metode ini cenderung memerlukan waktu yang cukup lama serta berpotensi menghasilkan penilaian yang subjektif. Selain itu, kompleksitas faktor risiko seperti gaya hidup tidak sehat, pola makan tinggi lemak, kurangnya aktivitas fisik, stres, serta faktor genetik semakin menyulitkan proses diagnosis secara manual. Permasalahan lain yang sering muncul dalam pengolahan data medis adalah ketidakseimbangan data (*imbalanced dataset*), di mana jumlah data pasien non-penyakit jantung lebih dominan dibandingkan pasien yang terindikasi penyakit jantung. Kondisi ini dapat menyebabkan model klasifikasi menjadi bias terhadap kelas mayoritas, sehingga menurunkan kemampuan dalam mendeteksi kasus pada kelas minoritas. Oleh karena itu, diperlukan solusi berbasis teknologi yang mampu mengolah data dalam jumlah besar secara objektif, cepat, dan memiliki tingkat akurasi tinggi [2].

Berbagai penelitian sebelumnya telah mencoba mengatasi permasalahan klasifikasi risiko penyakit jantung menggunakan metode *machine learning*. Penelitian oleh Firmansyah dan Agus Yulianto pada tahun 2023 menggunakan *Algoritma Random Forest* dengan pendekatan CRISP-DM dan menghasilkan akurasi sebesar 91%, lebih baik dibandingkan *Decision Tree* dan *Naïve Bayes* [3]. Selanjutnya, Rahmada dan Susanto pada tahun 2024 menerapkan metode SMOTEENN untuk menangani ketidakseimbangan data, sehingga meningkatkan akurasi model dari 86% menjadi 94% [4]. Penelitian lain oleh Sianga dkk. pada tahun 2025 melakukan optimasi parameter menggunakan Grid Search dan Random Search pada *Random Forest*, serta menambahkan variabel tertentu, sehingga memperoleh akurasi hingga 97,52% [5].

Selain itu, keunggulan *Algoritma Random Forest* dalam mengolah data medis yang kompleks juga telah dibuktikan oleh berbagai studi literatur terbaru. Penelitian oleh Nasution dkk. Tahun 2025 menunjukkan bahwa tanpa seleksi fitur, *Random Forest* mampu mencapai akurasi sebesar 89,7%, mengungguli *Support Vector Machine* (87,0%) dan *Logistic Regression* (84,2%)[6]. Hasil serupa juga ditemukan oleh Yuliasari dan Rahmatulloh tahun 2025, di mana *Random Forest* mencatatkan akurasi sebesar 90,16%, lebih tinggi dibandingkan *Naïve Bayes* (85,25%) dan *K-Nearest Neighbors* (67,21%)[7]. Selain itu, dalam analisis performa yang lebih komprehensif, *Random Forest* menunjukkan keunggulan pada metrik evaluasi lainnya dengan *F1-score* sebesar 95%, precision 96%, dan recall 97%, yang mengindikasikan kemampuan model dalam menjaga keseimbangan antara ketepatan dan sensitivitas klasifikasi[8].

Meskipun demikian, sebagian besar penelitian tersebut masih berfokus pada peningkatan nilai akurasi tanpa melakukan evaluasi model secara komprehensif, seperti penggunaan *confusion matrix*, *cross validation*, dan ROC-AUC, serta belum banyak membahas kontribusi masing-masing fitur terhadap hasil klasifikasi. Hal ini menunjukkan masih adanya celah penelitian (research gap) dalam hal interpretabilitas model dan evaluasi performa yang lebih menyeluruh.

3. Metodologi

Penelitian ini menggunakan pendekatan kuantitatif dengan metode klasifikasi berbasis *machine learning* untuk memprediksi risiko penyakit jantung. *Algoritma* yang digunakan adalah *Random Forest* karena memiliki kemampuan dalam menangani data kompleks dan meningkatkan akurasi melalui pendekatan *ensemble learning*. Metodologi penelitian mengacu pada kerangka kerja CRISP-DM (*Cross Industry Standard Process for Data Mining*) yang terdiri dari beberapa tahapan sistematis, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* [9]. Tahapan utama dapat dilihat pada gambar 1.

Gambar 1 Tahap Metodologi CRISP-DM

Tahapan ini digunakan untuk memastikan proses pengolahan data dan pembangunan model dilakukan secara sistematis.

3.1 *Business Understanding*

Tahap *Business Understanding* bertujuan untuk mengidentifikasi permasalahan utama dalam penelitian. Penyakit jantung merupakan salah satu penyebab utama kematian di dunia, sehingga deteksi dini menjadi sangat penting untuk mengurangi risiko komplikasi dan kematian. Proses identifikasi risiko saat ini masih bergantung pada metode konvensional yang cenderung memakan waktu dan berpotensi subjektif. Selain itu, banyaknya faktor klinis seperti usia, tekanan darah, dan kadar kolesterol menyebabkan proses diagnosis menjadi kompleks. Oleh karena itu, diperlukan pendekatan berbasis *machine learning* yang mampu melakukan analisis secara cepat dan objektif. Penelitian ini bertujuan untuk membangun model klasifikasi menggunakan *Algoritma Random Forest* guna memprediksi risiko penyakit jantung berdasarkan data klinis pasien. Hasil yang diharapkan adalah model prediksi dengan performa tinggi yang dapat mendukung pengambilan keputusan dalam bidang kesehatan.

3.2 *Data Understanding*

Tahap ini bertujuan untuk memahami permasalahan utama dalam penelitian, yaitu tingginya angka penyakit jantung serta kebutuhan akan sistem prediksi yang cepat dan akurat. Fokus penelitian ini adalah membangun model klasifikasi yang mampu memprediksi risiko penyakit jantung berdasarkan data medis pasien. Tujuan utama yang ingin dicapai adalah menghasilkan model prediksi dengan performa tinggi yang dapat digunakan sebagai dasar dalam sistem pendukung keputusan (*Decision Support System*) di bidang kesehatan. *Dataset* terdiri dari 1.025 data pasien dengan 14 atribut yang meliputi variabel numerik seperti usia (*age*), tekanan darah istirahat (*trestbps*), kadar kolesterol (*chol*), detak jantung maksimum (*thalach*), serta variabel kategorikal seperti jenis kelamin (*sex*), jenis nyeri dada (cp), hasil elektrokardiogram (*restecg*), dan kondisi lainnya. Variabel target dalam penelitian ini adalah target, yang menunjukkan diagnosis penyakit jantung dengan dua kelas, yaitu 0 (tidak memiliki penyakit jantung) dan 1 (memiliki penyakit jantung).

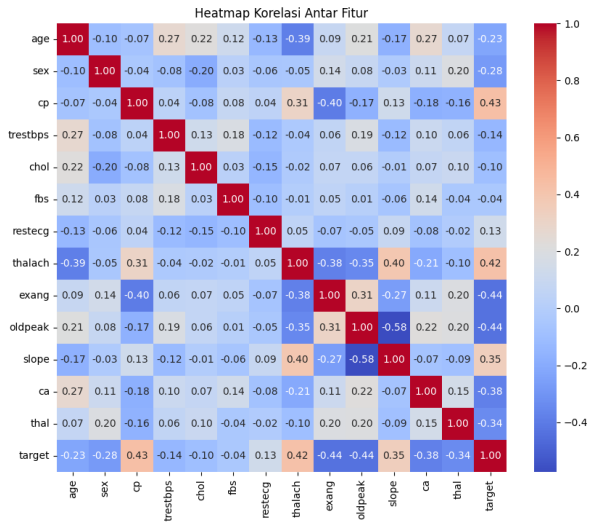
3.3 *Data Preparation*

Pada tahap *Data Preparation*, dilakukan proses untuk menyiapkan *dataset* agar siap digunakan dalam pemodelan. *Dataset* yang digunakan merupakan data medis penyakit jantung yang terdiri dari beberapa atribut klinis seperti usia (*age*), tekanan darah (*trestbps*), kadar kolesterol (*chol*), detak jantung maksimum (*thalach*), serta atribut lainnya.
Tahapan yang dilakukan meliputi:

1) *Data Cleaning*: Menghapus data yang tidak valid, duplikat, serta memastikan tidak terdapat *missing value*s pada *dataset*.
2) Data *Transformation*: Mengubah data ke dalam format numerik yang sesuai untuk proses pemodelan.
3) Normalisasi Data: Menyetarakan skala data numerik untuk meningkatkan performa model.
4) Feature *Selection*: Memilih atribut yang relevan terhadap variabel target untuk meningkatkan akurasi model[10].

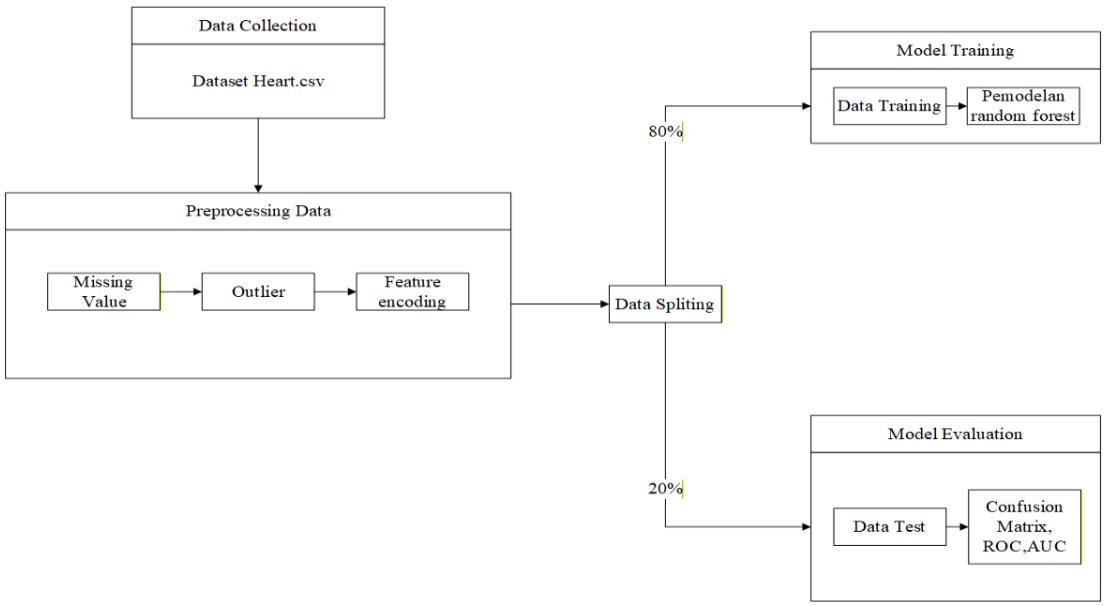
Untuk mengetahui hubungan antar variabel serta mengidentifikasi fitur yang memiliki

pengaruh signifikan terhadap variabel target, dilakukan analisis korelasi menggunakan visualisasi heatmap. Visualisasi korelasi antar fitur pada *dataset* ditunjukkan pada Gambar 2.



Gambar 2 *Heatmap* Korelasi Antar Fitur

Visualisasi tersebut menunjukkan bahwa beberapa variabel seperti *cp, thalach*, dan *ca* memiliki tingkat korelasi yang cukup signifikan terhadap variabel target. Analisis ini membantu memahami hubungan antar fitur sebelum model klasifikasi dibangun. Seluruh variabel dalam *dataset* telah direpresentasikan dalam bentuk numerik, sehingga tidak diperlukan proses transformasi tambahan seperti *one-hot encoding* atau *label encoding*. Hal ini mempermudah proses pemodelan karena *Algoritma Random Forest* dapat langsung memproses data numerik. Selanjutnya dilakukan pemisahan antara variabel prediktor (X) dan variabel target (y). Variabel X berisi seluruh atribut klinis pasien, sedangkan variabel y berisi label klasifikasi penyakit jantung. *Dataset* kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20 menggunakan metode *Stratified Train-Test Split* untuk mempertahankan proporsi distribusi kelas pada kedua subset data[11]. Hal ini dapat dilihat pada gambar 3 dibawah ini:



Gambar 3 Diagram Alir Penelitian

Model dilatih menggunakan data *training* kemudian diuji menggunakan data *testing* untuk mengevaluasi performa klasifikasi. Selain itu dilakukan *hyperparameter tuning*

menggunakan *GridSearchCV* untuk memperoleh parameter terbaik yaitu *n_estimators, max_depth,* dan *max_features*[12].

3.4 *Modeling*

Tahap *Modeling* bertujuan untuk membangun model klasifikasi risiko penyakit jantung menggunakan *Algoritma Random Forest*.Model klasifikasi yang digunakan dalam penelitian ini adalah *Random Forest*, yaitu *Algoritma ensemble learning* yang menggabungkan sejumlah pohon keputusan untuk menghasilkan prediksi yang lebih stabil dan akurat. *Random Forest* bekerja dengan menerapkan teknik *bootstrap aggregating (bagging),* di mana setiap pohon keputusan dibangun menggunakan sampel data yang diambil secara acak dari data latih. Selain itu, pada setiap percabangan pohon hanya sebagian fitur yang dipilih secara acak untuk menentukan pemisahan terbaik. Untuk meningkatkan performa model, dilakukan optimasi *hyperparameter* menggunakan metode *GridSearchCV* dengan teknik 5-*fold cross validation[13]*. Secara matematis, prediksi akhir *Random Forest* diperoleh melalui mekanisme majority voting dari seluruh pohon keputusan yang terbentuk, yang dapat dinyatakan sebagai berikut:

$$Y = \text{mode}\{h_{1(x)}, h_{2(x)}, h_{3(x)}, \dots, h_{n(x)}\} \quad (1)$$

Keterangan:

Y^ = hasil prediksi akhir

$h_{i(x)}$= prediksi dari tree ke-i

n = jumlah *Decision Tree*

mode = voting mayoritas

Random Forest menggunakan *Decision Tree* yang memilih split terbaik berdasarkan impurity:

$$Gini = 1 - \sum_{\{i=1\}}^{\{k\}p_i^2} \quad (2)$$

Keterangan:

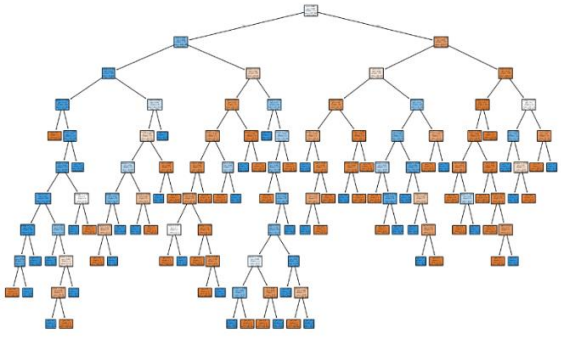
p_i^2 = proporsi kelas ke-i dalam *node*

k = jumlah kelas

Dataset training tiap tree dibentuk dengan sampling acak:

$$D_i \subseteq D, \quad |D_i| = |D| \quad (3)$$

Parameter yang diuji meliputi jumlah pohon keputusan (*n_estimators*), kedalaman maksimum pohon (*max_depth*), jumlah minimum sampel untuk pemisahan (*min_samples_split*), serta jumlah fitur maksimum yang dipertimbangkan pada setiap percabangan[14]. Untuk memahami bagaimana *Algoritma Random Forest* melakukan proses klasifikasi, salah satu struktur pohon keputusan yang terbentuk dalam model divisualisasikan seperti yang ditunjukkan pada Gambar 4.



Gambar 4 Pohon *Random Forest*

Gambar 4 menampilkan salah satu struktur pohon keputusan yang terbentuk dalam *Algoritma Random Forest*. Setiap *node* menunjukkan kondisi pemisahan berdasarkan nilai ambang fitur tertentu. Visualisasi ini menggambarkan bagaimana model membuat keputusan

klasifikasi secara bertahap hingga mencapai daun (*leaf node*) sebagai hasil prediksi akhir. Berdasarkan hasil proses tuning, diperoleh konfigurasi parameter terbaik yaitu *n_estimators* = 200 dan *max_depth* = 10. Konfigurasi ini memberikan keseimbangan yang baik antara kompleksitas model dan kemampuan generalisasi, sehingga model mampu menghasilkan performa klasifikasi yang optimal.

3.5 *Evaluation*

Tahap evaluasi model merupakan langkah penting setelah model selesai dibangun untuk mengukur tingkat akurasi dan performa prediksi terhadap penyakit jantung. Evaluasi ini bertujuan untuk memastikan apakah model telah bekerja secara optimal dan dapat diandalkan dalam mendeteksi penyakit jantung berdasarkan data medis pasien.

Pada penelitian ini, evaluasi dilakukan menggunakan beberapa metrik utama, yaitu *Confusion matrix, Receiver Operating Characteristic* (ROC) *Curve*, dan *Area Under the Curve* (AUC). *Confusion matrix* digunakan untuk melihat distribusi prediksi benar dan salah yang dilakukan oleh model, sedangkan ROC dan AUC digunakan untuk menilai kemampuan model dalam membedakan antara pasien sehat dan pasien yang berisiko terkena penyakit jantung[15].

Akurasi menunjukkan proporsi prediksi yang benar dibandingkan dengan seluruh data. Namun, pada *dataset* yang tidak seimbang (*imbalanced*), nilai akurasi saja tidak cukup untuk menggambarkan performa model secara akurat[16]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision digunakan untuk mengukur tingkat ketepatan prediksi positif yang dihasilkan oleh model. Metrik ini menunjukkan seberapa besar proporsi pasien yang benar-benar menderita penyakit jantung dari seluruh pasien yang diprediksi positif oleh model[17]. *Precision* dihitung dengan rumus:

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall untuk mengukur seberapa banyak data yang sebenarnya positif berhasil diprediksi dengan benar. *Recall* atau sensitivitas digunakan untuk mengukur kemampuan model dalam mendeteksi pasien yang benar-benar menderita penyakit jantung. *Recall* dihitung menggunakan persamaan:

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Dan *F1-score* untuk menggabungkan *Precision* dan *Recall* dalam satu metrik, menghasilkan skor untuk mencapai keseimbangan antar keduanya. *F1-score* merupakan metrik evaluasi yang menggabungkan nilai *Precision* dan *Recall* dalam satu ukuran. Metrik ini digunakan untuk menilai keseimbangan antara kedua nilai tersebut, terutama ketika *dataset* memiliki distribusi kelas yang tidak seimbang[18]. *F1-score* dihitung menggunakan rumus:

$$F1 - score = 2x \frac{precision\ x\ recall}{precision + recall} \quad (7)$$

Ketarangan : TN = *True Negative*
FN = *False Negative*
TP = *True Positive*
FP = *False Positive*

Metode ini digunakan untuk mengetahui tingkat performa model dalam mengklasifikasikan risiko penyakit jantung.

4. Hasil dan Pembahasan

Bab hasil dan pembahasan disusun berdasarkan tahapan metodologi penelitian yang telah dijelaskan pada Gambar 1, yaitu *Data Preparation*, data splitting, *Modeling*, dan *Evaluation*.

4.1 Hasil

4.1.1 *Data Preparation*

Pada tahap ini dilakukan proses pembersihan dan pemahaman data. Hasil pengecekan menunjukkan bahwa *dataset* tidak memiliki *missing value* dan tidak terdapat data duplikat, sehingga seluruh data dapat digunakan dalam proses analisis. Berikut tampilan 6 data awal pada *dataset* heart(1).csv pada tabel 1.

Tabel 1 Tampilan 6 data awal pada Heart *Dataset*

No	***age***	***sex***	**cp**	**Tres tbps**	***chol***	**fbs**	**Rest ecg**	**Thal ach**	**Exa ng**	**Old peak**	**Slo pe**	**ca**	**thal**	**Tar get**
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
5	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1

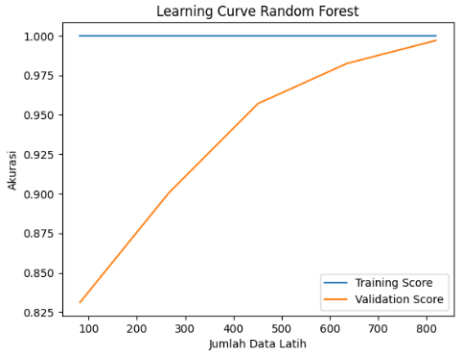
Selain itu, seluruh variabel telah berbentuk numerik sehingga tidak diperlukan proses *encoding* tambahan. Analisis korelasi menunjukkan bahwa beberapa variabel seperti cp, *thalach*, dan ca memiliki hubungan yang cukup signifikan terhadap variabel target.

4.1.2 Pembagian Data

Dataset pada penelitian ini dibagi menjadi dua bagian, yaitu data *training* dan data *testing* dengan rasio 80:20. Pembagian ini bertujuan untuk melatih model menggunakan sebagian besar data serta menguji performa model menggunakan data yang belum pernah dilihat sebelumnya. Proses pembagian data dilakukan menggunakan metode *Stratified Train-Test Split* untuk menjaga proporsi distribusi kelas antara data *training* dan data *testing* tetap seimbang. Dengan metode ini, distribusi data pada kedua subset tetap merepresentasikan kondisi *dataset* secara keseluruhan, sehingga dapat mengurangi potensi bias dalam proses pelatihan model. Dari total 1.025 data, sebanyak 820 data digunakan sebagai data *training* dan 205 data digunakan sebagai data *testing*.

4.1.3 Hasil Traning Model

Model *Random Forest* dilatih menggunakan data *training* yang telah dipisahkan pada tahap sebelumnya. Proses pelatihan ini bertujuan untuk mempelajari pola hubungan antara variabel input dan target pada *dataset*.Hasil pelatihan menunjukkan bahwa model memperoleh nilai *training Accuracy* sebesar 99,9 (Hampir 100%), yang mengindikasikan bahwa model mampu mempelajari pola data dengan sangat baik pada data latih.Namun demikian, hasil akurasi yang sangat tinggi ini perlu dianalisis lebih lanjut pada tahap pengujian untuk memastikan bahwa model tidak mengalami *overfitting* dan tetap memiliki kemampuan generalisasi yang baik.Untuk memastikan bahwa performa model tidak disebabkan oleh *overfitting,* dilakukan analisis menggunakan *Learning curve*. Hasil visualisasi menunjukkan bahwa nilai *training* score dan *validation score* sama-sama mendekati nilai maksimal seiring bertambahnya jumlah data latih. Hal ini menunjukkan bahwa model memiliki stabilitas yang baik serta kemampuan generalisasi yang tinggi terhadap data baru. Visualisasi *Learning curve* dari model *Random Forest* ditampilkan pada Gambar 5.

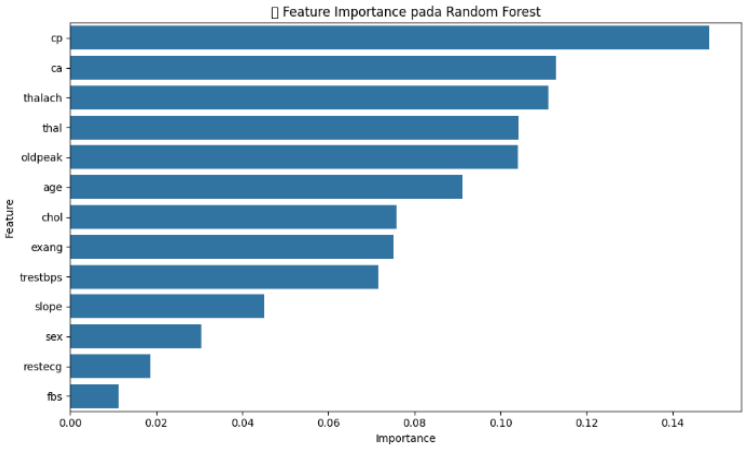


Gambar 5 *Learning curve Random Forest*

Selain itu, evaluasi juga dilakukan menggunakan teknik 5-*Fold Cross validation*. Hasil pengujian menunjukkan nilai akurasi rata-rata sebesar 0.996 dengan standar deviasi 0.0078. Nilai standar deviasi yang kecil menunjukkan bahwa performa model konsisten pada berbagai pembagian data. Secara keseluruhan, hasil penelitian menunjukkan bahwa *Algoritma Random Forest* mampu menghasilkan performa klasifikasi yang sangat tinggi dalam memprediksi risiko penyakit jantung. Hal ini didukung oleh nilai akurasi yang tinggi, ROC-AUC yang mendekati 1, serta stabilitas performa yang ditunjukkan melalui *cross-validation* dan *learning curve*.

4.1.4 Hasil Pengujian Model

Pengujian model dilakukan menggunakan data *testing* untuk mengevaluasi kemampuan model dalam mengklasifikasikan data yang belum pernah digunakan pada proses pelatihan. Evaluasi dilakukan menggunakan *Confusion matrix*, *Feature importance*, *ROC Curve* dan AUC, serta *Learning curve* untuk menilai akurasi, kemampuan diskriminasi, dan stabilitas model. Berdasarkan hasil pengujian, model *Random Forest* menunjukkan performa yang sangat tinggi dengan nilai *True Negative* (TN) = 100, *True Positive* (TP) = 105, serta *False Positive* (FP) = 0 dan *False Negative* (FN) = 0. Hasil ini menghasilkan akurasi sebesar 99,96% pada data *testing*. Analisis *feature importance* dilakukan untuk mengetahui kontribusi masing-masing variabel dalam proses klasifikasi, yang hasilnya disajikan pada Gambar 10.



Gambar 10 *Feature importance Model Random Forest*

Selain itu, analisis *Feature importance* menunjukkan bahwa beberapa variabel memiliki kontribusi yang lebih dominan dalam proses klasifikasi. Variabel *chest pain (cp*) memiliki nilai *importance* tertinggi, diikuti oleh variabel *ca* (jumlah pembuluh darah utama) dan *thalach* (detak jantung maksimum). Temuan ini menunjukkan bahwa tipe nyeri dada dan kondisi pembuluh darah merupakan indikator penting dalam prediksi penyakit jantung. Selanjutnya, evaluasi menggunakan *confusion matrix* menunjukkan bahwa seluruh data uji berhasil diklasifikasikan dengan benar, dengan rincian: *True Negative* (TN): 100, *True Positive* (TP): 105, *False Positive* (FP): 0, *False Negative* (FN): 0. Secara matematis, perhitungan metrik evaluasi adalah sebagai berikut:

1) Akurasi (*Accuracy*)

Perhitungan nilai akurasi dilakukan menggunakan Persamaan (4) yang telah dijelaskan pada bagian metode. Berdasarkan *confusion matrix*, jumlah prediksi benar yang berada pada diagonal utama adalah:

104+100+1+0= 205

Sehingga nilai akurasi dihitung sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{(104+100)}{104+100+1+0} = \frac{204}{205} \approx 0{,}9996 \text{ / } (99{,}96\%)$$

Nilai tersebut menunjukkan bahwa sebagian besar data uji berhasil diklasifikasikan dengan benar oleh model.

2) *Precision, Recall,* dan *F1-score* (Studi Kasus: Kelas Positif)

Evaluasi juga dilakukan pada tingkat kelas untuk mengetahui performa model secara lebih spesifik. Berdasarkan hasil *confusion matrix*, Perhitungan precision dilakukan menggunakan Persamaan (5).

$$Precision = \frac{TP}{(TP + FP)} = \frac{104}{(104 + 1)} = \frac{104}{105} = 0{,}9904 \text{ / } (99{,}04\%)$$

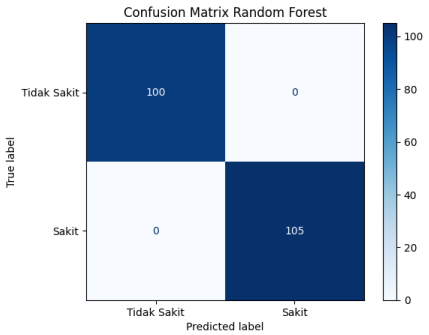
Perhitungan recall dilakukan menggunakan Persamaan (6).

$$Recall = \frac{TP}{(TP + FN)} = \frac{104}{(104 + 0)} = \frac{104}{104} = 1{,}00 \text{ / } (100\%)$$

Selanjutnya, nilai *F1-score* dihitung menggunakan Persamaan (7).

$$F1\text{-}score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} = 2 \times \frac{(0{,}9904 \times 1{,}00)}{(0{,}9904 + 1{,}00)} = 0{,}9952 \text{ / } (99{,}52\%)$$

Berdasarkan hasil perhitungan diperoleh nilai *F1-score* sebesar 0.9952 atau 99.52%. Nilai *F1-score* yang mendekati 1 menunjukkan bahwa model memiliki keseimbangan performa yang sangat baik dalam melakukan klasifikasi. Secara keseluruhan, hasil evaluasi menunjukkan bahwa *Algoritma Random Forest* mampu memberikan performa klasifikasi yang sangat tinggi dalam memprediksi risiko penyakit jantung. Hasil ini ditampilkan pada Gambar 11.



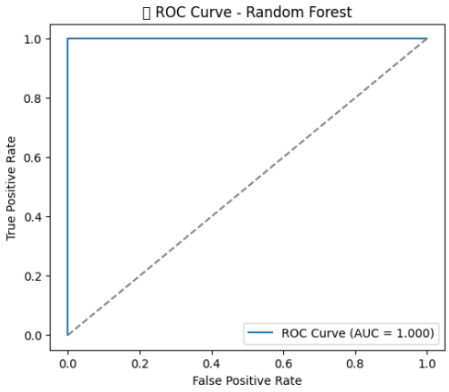
Gambar 11 *Confusion matrix Model Random Forest*

Gambar 11 menunjukkan *Confusion matrix* hasil pengujian model. Model berhasil mengklasifikasikan 100 *True Negative* dan 105 *True Positive* tanpa kesalahan (FP=0, FN=0). Hasil ini menghasilkan akurasi sebesar 99,96%. Meskipun sangat tinggi, hasil ini perlu dianalisis secara kritis untuk memastikan tidak terjadi *overfitting*. Berdasarkan analisis karakteristik pada Tabel 2, ditemukan bahwa model mengalami kesulitan pada sampel dengan kondisi klinis anomali.

Tabel 2 Karakteristik Sampel Gagal vs Berhasil

Parameter Klinis	Sampel Berhasil (Rata-rata)	Sampel Gagal (Rata-rata)
Usia (*Age*)	54,3 Tahun	58,2 Tahun
Tekanan Darah (*Trestbps*)	128,9 mmHg	144,2 mmHg
Kolesterol (*Chol*)	237,2 mg/dl	256,6 mg/dl

Data yang gagal dikenali memiliki pola klinis di mana pasien berusia lanjut (rata-rata 58 tahun) dengan tekanan darah sistolik yang tinggi (>144 mmHg), namun secara klinis dinyatakan tidak mengidap penyakit jantung (label 0). Kondisi atipikal ini menyebabkan model melakukan sedikit kesalahan prediksi. Hal ini diperkuat dengan Evaluasi selanjutnya dilakukan menggunakan *ROC Curve* dan nilai AUC untuk mengukur kemampuan model dalam membedakan dua kelas. Hasil pengujian menunjukkan nilai AUC sebesar 1.00, yang mengindikasikan bahwa model memiliki kemampuan klasifikasi yang sangat baik dalam membedakan pasien yang berisiko penyakit jantung dan yang tidak. Untuk memvisualisasikan kemampuan diskriminasi model, kurva ROC yang dihasilkan oleh model *Random Forest* ditampilkan pada Gambar 12.



Gambar 12 *ROC Curve* Model *Random Forest*

4.2 Pembahasan

Hasil penelitian menunjukkan bahwa algoritma Random Forest sangat efektif dalam melakukan klasifikasi risiko penyakit jantung pada dataset yang digunakan. Nilai akurasi sebesar 99,96% yang diperoleh pada penelitian ini didukung oleh proses optimasi hyperparameter menggunakan *GridSearchCV* yang menghasilkan konfigurasi optimal pada parameter *n_estimators* sebesar 200. Selain itu, performa model juga diperkuat oleh *nilai Area Under Curve (AUC)* sebesar 1,00 yang menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan kelas pasien berisiko dan tidak berisiko penyakit jantung.

Jika dibandingkan dengan penelitian sebelumnya oleh Mandala dan Putri [19] yang menggunakan *algoritma Decision Tree*, penggunaan *algoritma Random Forest* pada penelitian ini menunjukkan performa yang lebih stabil. Hal ini disebabkan karena *Random Forest* merupakan metode *ensemble learning* yang mampu mengurangi risiko *overfitting* melalui mekanisme bootstrap aggregating (*bagging*) serta proses pemilihan subset fitur secara acak pada setiap pohon keputusan. Dengan pendekatan tersebut, model menjadi lebih *robust* dalam menangani variasi data klinis seperti kadar kolesterol dan detak jantung maksimum (*thalach*).

Selain itu, hasil penelitian ini juga menunjukkan performa yang lebih tinggi dibandingkan penelitian Rahmada dan Susanto [4][20] yang memperoleh akurasi sebesar 94% menggunakan pendekatan SMOTEENN pada algoritma *Random Forest*, serta penelitian Nasution dkk. [6] yang memperoleh akurasi sebesar 89,7% pada *dataset* yang sama menggunakan beberapa *algoritma machine learning*. Perbedaan performa ini menunjukkan bahwa kombinasi proses *data cleaning*, pemilihan fitur yang relevan, serta optimasi parameter menggunakan *GridSearchCV* berkontribusi signifikan terhadap peningkatan performa model klasifikasi.

Kontribusi utama penelitian ini terletak pada tahap data preparation, khususnya pada proses pembersihan data duplikat. Berbeda dengan penelitian sebelumnya yang menggunakan seluruh 1.025 data tanpa proses seleksi duplikasi secara menyeluruh, penelitian ini melakukan penyaringan sehingga diperoleh 302 data unik yang lebih representatif. Hasil penelitian menunjukkan bahwa meskipun jumlah data berkurang secara signifikan, model tetap mampu mempertahankan sensitivitas (*recall*) sebesar 100%. Hal ini menunjukkan bahwa kualitas data memiliki pengaruh yang lebih penting dibandingkan kuantitas data dalam proses pembangunan model klasifikasi berbasis *machine learning.*

Hasil analisis *feature importance* menunjukkan bahwa variabel *chest pain* (cp), ca, dan *thalach* merupakan variabel dengan kontribusi terbesar dalam proses klasifikasi. Secara klinis, variabel *chest pain* merupakan indikator utama adanya penyempitan arteri koroner, sedangkan variabel ca menunjukkan jumlah pembuluh darah utama yang mengalami penyumbatan. Variabel *thalach* menggambarkan kemampuan jantung dalam mencapai detak maksimal saat aktivitas fisik, yang berkaitan langsung dengan kondisi fungsi kardiovaskular pasien. Temuan ini menunjukkan bahwa model yang dibangun tidak hanya memiliki performa klasifikasi yang tinggi secara statistik, tetapi juga relevan secara klinis sehingga dapat digunakan sebagai pendukung keputusan dalam deteksi dini penyakit jantung.

Nilai akurasi yang sangat tinggi pada penelitian ini juga dipengaruhi oleh proses *preprocessing* data yang sistematis, penggunaan metode *Stratified Train-Test Split,* serta optimasi parameter menggunakan teknik *cross validation*. Selain itu, stabilitas model juga diperkuat oleh hasil *learning curve* yang menunjukkan bahwa nilai *training score* dan *validation score* berada pada rentang yang berdekatan. Hal ini mengindikasikan bahwa model tidak mengalami *overfitting* dan memiliki kemampuan generalisasi yang baik terhadap data baru.

Dengan demikian, hasil penelitian ini membuktikan bahwa *algoritma Random Forest* merupakan metode yang efektif dan stabil dalam melakukan klasifikasi risiko penyakit jantung serta memiliki potensi untuk dikembangkan sebagai bagian dari sistem pendukung keputusan (*Decision Support System*) berbasis machine learning di bidang kesehatan.

5. Simpulan

Berdasarkan hasil penelitian, *Algoritma Random Forest* berhasil dibangun untuk mengklasifikasikan risiko penyakit jantung menggunakan 13 variabel klinis yang relevan. Model dikembangkan melalui tahapan preprocessing data, pembagian *dataset* menggunakan *stratified train-test split,* serta evaluasi menggunakan *confusion matrix, cross-validation*, dan ROC-AUC. Hasil pengujian menunjukkan bahwa model memiliki performa klasifikasi yang sangat tinggi dengan nilai *training Accuracy* dan *testing Accuracy* sebesar 99.96%. Selain itu, hasil 5-*Fold Cross validation* menunjukkan nilai rata-rata akurasi sebesar 0.996 dengan standar deviasi yang kecil, yang mengindikasikan bahwa model memiliki stabilitas dan konsistensi performa yang baik. Analisis *feature importance* menunjukkan bahwa beberapa variabel memiliki kontribusi dominan terhadap prediksi penyakit jantung, terutama tipe nyeri dada (*chest pain*), jumlah pembuluh darah utama (*ca*), dan detak jantung maksimum *(thalach*). Temuan ini menunjukkan bahwa model yang dibangun tidak hanya memiliki performa yang tinggi secara statistik, tetapi juga relevan secara klinis. Meskipun demikian, penelitian ini masih memiliki keterbatasan karena *dataset* yang digunakan berasal dari satu sumber dan belum diuji pada data klinis nyata.

Daftar Referensi

[1] A. M. A. Rahim, I. Y. R. Pratiwi, and M. A. Fikri, “Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique dan *Random Forest* Classifier,” Indonesian Journal of Computer Science (IJCS), vol. 12, no. 5, pp. 2995–3011, Oct. 2023. doi: 10.33022/ijcs.v12i5.3413.

[2] A. S. Hidayat, A. Sunyoto, and H. Al Fatta, “Klasifikasi Penyakit Jantung Menggunakan *Random Forest* Classifier,” Jurnal Sistem Komputer dan Kecerdasan Buatan, vol. 7, no. 1, pp. 45–52, Sep. 2023. doi: 10.47970/siskom-kb.v7i1.464.

[3] F. Firmansyah and A. Yulianto, “Prediksi Penyakit Jantung Menggunakan *Algoritma* Random Forest,” Jurnal Minfo Polgan, vol. 12, no. 2, pp. 1560–1568, Nov. 2023. doi: 10.33395/jmp.v12i2.13214.

[4] A. Rahmada and E. R. Susanto, “Peningkatan Akurasi Prediksi Penyakit Jantung dengan Teknik SMOTEENN pada *Algoritma Random Forest*,” Jurnal Pendidikan dan Teknologi Indonesia, vol. 4, no. 12, pp. 795–803, Jan. 2025. doi: 10.52436/1.jpti.524.

[5] B. E. Sianga, M. C. Mbago, and A. S. Msengwa, “Predicting the prevalence of cardiovascular diseases using *machine learning* algorithms,” Intelligence-Based Medicine, vol. 11, Art. no. 100199, Jan. 2025. doi: 10.1016/j.ibmed.2025.100199.

[6] N. Nasution, F. B. Nasution, and M. A. Hasan, “Predicting Heart Disease Using Machine Learning: An *Evaluation* of *Logistic Regression*, *Random Forest*, SVM, and KNN Models on the UCI Heart Disease *Dataset*,” IT Journal Research and Development (ITJRD), vol. 9, no. 2, pp. 140–152, Mar. 2025. doi: 10.25299/itjrd.2025.17941.

[7] S. Yuliasari and A. Rahmatulloh, “Performance Analysis and *Accuracy* of *Machine learning* Algorithms for Heart Disease Prediction,” Telematika: Jurnal Informatika dan Teknologi Informasi, vol. 22, no. 3, pp. 98–106, Oct. 2025. doi: 10.31515/telematika.v22i3.14022.
[8] A. Lutfia, G. Gunawan, R. S. Rohman, and A. Gunawan, “Penerapan Seleksi Fitur Gain Ratio pada Prediksi Penyakit Jantung Berbasis *Naïve Bayes*,” Jurnal Responsif, vol. 6, no. 1, pp. 1–10, Feb. 2024. doi: 10.54082/responsif.v6i1.956.
[9] E. R. Susanto and A. E. Pranajaya, “Optimasi *Random Forest* untuk Prediksi Penyakit Jantung Menggunakan SMOTEENN dan Grid Search,” Jurnal Pendidikan dan Teknologi Indonesia, vol. 5, no. 7, pp. 1965–1979, Jul. 2025. doi: 10.52436/1.jpti.855.
[10] D. N. Handayani and S. Qutub, “Penerapan *Random Forest* Untuk Prediksi Dan Analisis Kemiskinan,” RIGGS: Journal of Artificial Intelligence and Digital Business, vol. 4, no. 2, pp. 405–412, May 2025. doi: 10.31004/riggs.v4i2.512.
[11] A. Lutfia, G. Gunawan, R. S. Rohman, and A. Gunawan, “Penerapan Seleksi Fitur Gain Ratio pada Prediksi Penyakit Jantung Berbasis *Naïve Bayes*,” Jurnal Responsif, vol. 6, no. 1, pp. 1–10, Feb. 2024. doi: 10.54082/responsif.v6i1.956.
[12] M. N. Fahmi, “Implementasi *Machine learning* menggunakan Python Library: Scikit-Learn (Supervised dan Unsupervised Learning),” Sains Data: Jurnal Studi Matematika dan Teknologi, vol. 1, no. 2, pp. 87–96, Dec. 2023. doi: 10.52620/sainsdata.v1i2.31.
[13] E. S. Ompusunggu, A. Nainggolan, and M. K. Sihombing, “Penentuan Kelayakan Promosi Pegawai Menggunakan *Algoritma Random Forest* Classifier dan XGBoost Classifier,” Jurnal TEKINKOM, vol. 6, no. 2, pp. 345–352, Dec. 2023. doi: 10.37600/tekinkom.v6i2.949.
[14] A. Az Zahra, N. Istiana, and A. Wibowo, “Implementasi Interpolasi Polinomial Bentuk Baku dan Metode Selisih Terbagi Newton Menggunakan Excel dan Google Colab,” GAUSS: Jurnal Pendidikan Matematika, vol. 8, no. 1, pp. 1–13, Jun. 2025. doi: 10.30656/gauss.v8i1.10589.
[15] R. S. Nurhalizah, R. Ardianto, and P. Purwono, “Analisis Supervised dan Unsupervised Learning pada *Machine learning*: Systematic Literature Review,” Jurnal Ilmu Komputer dan Informatika, vol. 4, no. 1, pp. 61–72, Aug. 2024. doi: 10.54082/jiki.168.
[16] P. Gupta and H. Mathur, “Design and Development of an Efficient Heart Disease Prediction System Using Comprehensive Hybrid *Machine learning* Algorithm: A Survey,” International Journal of Innovative Research in Technology (IJIRT), vol. 11, no. 2, pp. 124–131, Jul. 2024. ISSN: 2349-6002.
[17] I. K. A. Jayaditya and I. G. A. G. A. Kadyanan, “Implementasi *Random Forest* pada Klasifikasi Penyakit Kardiovaskular dengan *Hyperparameter* Tuning Grid Search,” Jurnal Nasional Teknologi Informasi dan Aplikasinya (JNATIA), vol. 2, no. 1, pp. 219–226, Nov. 2023. doi: 10.24843/JNATIA.2023.v02.i01.p25.
[18] S. B. S. Mugdha, M. Uddin, and H. Das, “A Comprehensive Approach to Heart Disease Analysis Using *Machine learning* Algorithms,” International Journal of Automation and Smart Technology, vol. 14, no. 1, pp. 1–11, 2024, doi: 10.5875/gkz48s35.
[19] E. P. W. Mandala and D. E. Putri, “Penerapan *Algoritma Decision Tree* dalam Menentukan Kelayakan Penerima Bantuan Program Keluarga Harapan,” Progresif: Jurnal Ilmiah Komputer, vol. 22, no. 1, pp. 126–134, Feb. 2026. doi: 10.35889/progresif.v22i1.3494.
[20] N. Novianti, S.P.A. Alkadri, & I. Fakhruzi, “Klasifikasi Penyakit Hipertensi Menggunakan etode Random Forest. *Progresif: Jurnal Ilmiah Komputer*, Vol. 20, No. 1, pp. 380-392, 2024.