The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 5, October 2024
https://doi.org/10.33022/ijcs.v13i5.4214

# Optimasi Pemilihan Fitur untuk Prediksi Penyakit Jantung Menggunakan Algoritma Genetika dan Random Forest

**Takhamo Gori[1], Annisa Hestiningtyas[2]**
takhamo.gori@students.amikom.ac.id[1], annisahestiningtyas@poltara.ac.id[2]
[1]Universitas Amikom Yogyakarta
[2]Politeknik Nusantara Balikpapan

**Informasi Artikel**

Diterima : 29 Jun 2024
Direview : 3 Okt 2024
Disetujui : 30 Okt 2024

**Kata Kunci**

Penyakit Jantung, Pemilihan Fitur, Optimasi Parameter, Algoritma Genetika, Random Forest

**Abstrak**

Penyakit jantung merupakan salah satu penyebab utama kematian di seluruh dunia, menekankan urgensi prediksi dini dan manajemen risiko yang efektif. Dalam upaya meningkatkan akurasi prediksi penyakit jantung, penelitian ini mengusulkan pendekatan metode GridSearchCV (GS) dan Genetic Algorithm Feature Selection (GA-FS) pada model Random Forest (RF). Setelah proses seleksi fitur dengan GA-FS, dari sebelas atribut awal dimasukkan, delapan atribut terpilih, yakni Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, ExerciseAngina, dan ST_Slope, sementara atribut Age, MaxHR, dan Oldpeak dieliminasi. Hasil penelitian menunjukkan bahwa model RF yang dioptimalkan dengan GS dan GA-FS (RF-GS-GAFS) mencapai akurasi 91.85%, presisi 95.10%, recall 90.65%, dan F1-Score 92.82%, mengungguli model RF dengan optimasi GS (89.67%) dan RF tanpa optimalisasi (88.04%). Temuan ini memberikan kontribusi positif yang signifikan dalam meningkatkan kinerja model prediksi penyakit jantung melalui optimalisasi parameter dan pemilihan fitur menggunakan algoritma genetika.

***Keywords***

*Heart Disease, Feature Selection, Parameter Optimization, Genetic Algorithm, Random Forest*

***Abstract***

*Heart disease is one of the leading causes of death worldwide, emphasizing the urgency of early prediction and effective risk management. In an effort to enhance the accuracy of heart disease prediction, this study proposes an approach using GridSearchCV (GS) and Genetic Algorithm Feature Selection (GA-FS) methods with the Random Forest (RF) model. Following the feature selection process with GA-FS, out of eleven initial attributes, eight were selected: Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, ExerciseAngina, and ST_Slope, while Age, MaxHR, and Oldpeak were eliminated. The research results indicate that the RF model optimized with GS and GA-FS (RF-GS-GAFS) achieved an accuracy of 91.85%, precision of 95.10%, recall of 90.65%, and F1-Score of 92.82%, outperforming the GS-optimized RF model (89.67%) and RF without optimization (88.04%). These findings provide a significant positive contribution to improving the performance of heart disease prediction models through parameter optimization and feature selection using genetic algorithms.*

## A. Pendahuluan

Penyakit jantung merupakan masalah serius yang dihadapi oleh kesehatan global, penyebab utama kematian di seluruh dunia. Menurut WHO, pada tahun 2016, sekitar 17,9 juta jiwa meninggal dunia, di mana sekitar 31 persen dari total kematian di seluruh dunia disebabkan oleh penyakit jantung [1]. Deteksi dini dan prediksi risiko penyakit jantung menjadi kunci untuk mengambil langkah-langkah preventif yang efektif dalam menyelamatkan pasien [2]. Dalam upaya meningkatkan akurasi prediksi penyakit jantung, pengembangan model klasifikasi telah menjadi fokus utama dalam penelitian.

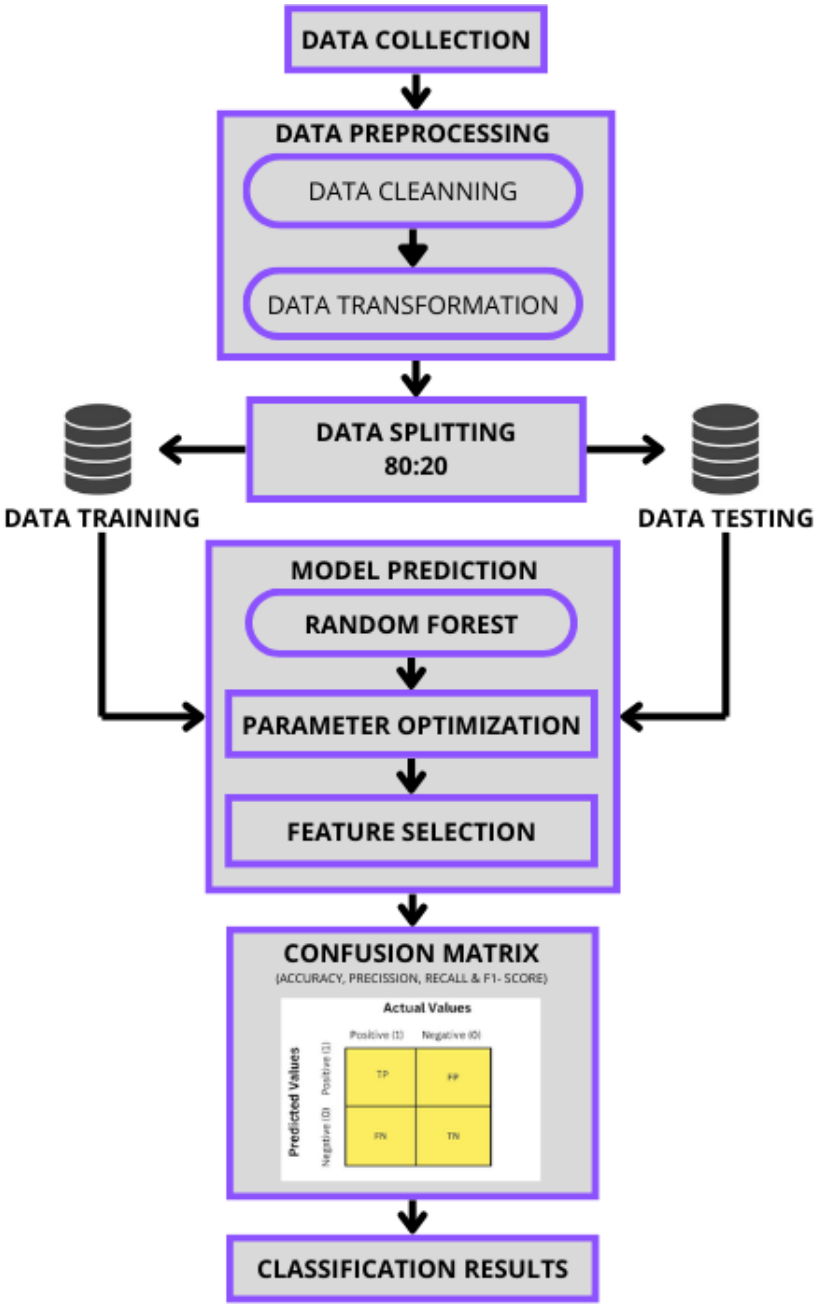
Meskipun banyak penelitian telah dilakukan tentang prediksi penyakit jantung [3], [4], [5], [6], salah satu tantangan utama dalam membangun model prediktif adalah memilih atribut yang paling informatif dan relevan [7], [8], [9]. Pemilihan fitur yang optimal dapat meningkatkan kinerja model, mengurangi overfitting, dan memberikan pemahaman yang lebih baik terhadap faktor-faktor yang mempengaruhi penyakit jantung [10].

Salah satu pendekatan dalam pemilihan fitur adalah penggunaan algoritma genetika. Algoritma genetika dapat secara efisien mengeksplorasi kombinasi atribut yang berpotensi memberikan hasil terbaik [11]. Dalam penelitian ini, kami mengusulkan pengoptimalan pemilihan fitur untuk prediksi penyakit jantung dengan menggabungkan algoritma genetika dan model Random Forest.

Penelitian ini bertujuan untuk mengoptimalkan pemilihan fitur dalam prediksi penyakit jantung dengan menggabungkan algoritma genetika (GA) dan model Random Forest (RF), serta menggunakan optimasi Gridsearch untuk menentukan parameter terbaik pada model Random Forest. Dengan pendekatan ini, diharapkan dapat mengidentifikasi subset fitur yang paling informatif untuk meningkatkan akurasi prediksi penyakit jantung serta memberikan dasar bagi pengembangan model yang lebih efektif dalam konteks praktek klinis.

## B. Tahapan Penelitian

Tahapan penelitian yang dilakukan dalam penelitian ini dimulai dari pengumpulan data, preprocessing data yang meliputi: data cleaning dan data transformation. Setelah itu, melakukan proses splitting data, membangun model prediksi menggunakan algoritma Random Forest (RF), optimasi parameter Random Forest menggunakan GridSearchCV (GS), Genetic Algoritm feature selection (GA-FS), mengevaluasi model prediksi dengan Confusion Matrix dan analisis hasil. Tahapan penelitian ditunjukkan pada gambar 1.




**Gambar 1.** Tahapan Penelitian

1. Data Collection

Data yang digunakan dalam penelitian ini adalah Heart Failure Prediction Dataset yang diperoleh melalui situs web Kaggle dan dapat diakses melalui UCI Machine Learning Repository [12]. Dataset ini merupakan hasil penggabungan dari lima dataset penyakit jantung, meliputi dataset Cleveland, Hungarian, Swiss, Long Beach VA, dan Stalog. Dataset terdiri dari sebelas atribut prediktor meliputi Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak dan ST_Slope dan satu atribut sebagai kelas target yaitu HeartDisease. Informasi atribut dataset ditunjukkan pada Tabel 1.

**Tabel 1**. Informasi Atribut Dataset

| *Atribute* | *Description* | *Value* |
|---|---|---|
| Age | Age of the patient in years | 28–77 |
| Sex | Sex of the patient | M: Male<br>F: Female |
| ChestPainType | Chest pain type | TA: Typical Angina<br>ATA: Atypical Angina<br>NAP: Non-Anginal Pain<br>ASY: Asymptomatic |
| RestingBP | Resting blood pressure in mm Hg | 80–200 |
| Cholesterol | Serum cholesterol in mm/dl | 0 – 603 |
| FastingBS | Fasting blood sugar | 1: if FastingBS > 120 mg/dl<br>0: otherwise |
| RestingECG | Resting electrocardiogram results | Normal: Normal<br>ST: having ST-T wave abnormality<br>LVH: showing probable or definite left ventricular hypertrophy |
| MaxHR | Maximum heart rate achieved | 60–202 |
| ExerciseAngina | Exercise-induced angina | Y: Yes<br>N: No |
| Oldpeak | Numeric value measured in depression | -2.6 – 6.2 |
| ST_Slope | The slope of the peak exercise ST segment | Up: upsloping<br>Flat: flat<br>Down: downsloping |
| HeartDisease | Output class | 1: heart disease<br>0: Normal |

2. Data Preprocessing

Pada tahap pra-pemrosesan data, kami melakukan dua langkah utama: pembersihan data dan transformasi data. Pembersihan data melibatkan identifikasi dan penanganan nilai yang hilang, penghapusan data yang tidak konsisten (outlier/noisy), deteksi serta penghapusan data duplikat, dan penanganan data yang tidak lengkap [13]. Sementara itu, dalam proses transformasi data, kami mengubah data kategorikal menjadi data numerik menggunakan metode Label Encoding. Label Encoding adalah suatu teknik pengolahan data yang mengonversi data kategori atau data ordinal menjadi bentuk numerik [14]. Dengan demikian, setiap tabel yang berisi data dengan tipe string atau teks akan mengalami transformasi ke bentuk numerik.

3. Data Splitting

Setelah proses preprocessing data, langkah berikutnya melibatkan pembagian data (*data splitting*). *Data splitting* merupakan proses membagi dataset menjadi dua bagian utama, yakni data training dan data testing. Data training berperan dalam melatih model, sementara data testing digunakan untuk menguji sejauh mana kinerja model yang telah dilatih [15]. Dalam penelitian ini, pembagian data dilakukan dengan rasio 80% data training dan 20% data testing.

4. Random Forest

Algoritma Random Forest (RF) merupakan salah satu teknik klasifikasi yang termasuk dalam kategori Supervised Learning. Teknik ini melibatkan penggunaan beberapa pohon keputusan yang bekerja bersama untuk membuat prediksi kelas. Prediksi akhir diambil berdasarkan mayoritas suara dari seluruh kelompok pohon yang membentuk model prediktif. Kelebihan utama dari RF adalah kemampuannya meningkatkan akurasi dan mengurangi risiko overfitting pada dataset, menjadikannya metode yang efektif bahkan saat diterapkan pada dataset yang besar dan mungkin memiliki sejumlah besar nilai catatan yang hilang. Fleksibilitas algoritma ini juga terlihat dari kemampuannya menyimpan dan menggunakan hasil prediksi dari masing-masing pohon keputusan untuk berbagai jenis data [7].

Hyperparameter tuning pada model RF disesuaikan menggunakan metode GridSearchCV (GS). Grid search merupakan teknik tuning hyperparameter yang sering digunakan dalam machine learning untuk menemukan kombinasi hyperparameter yang optimal pada model tertentu. Proses ini melibatkan pencarian menyeluruh dari semua kombinasi hyperparameter dalam rentang atau set nilai yang telah ditentukan [16]. Pada penelitian ini, kami menguji berbagai nilai parameter sebagaimana tercantum dalam Tabel 2. Set parameter terbaik yang kami peroleh dari hasil GS kemudian kami manfaatkan dalam melatih model RF, dengan tujuan mencapai akurasi klasifikasi yang maksimal.

**Tabel 2**. Hyperparameter Tuning RF

| Parameter | Nilai |
|---|---|
| n_estimators | 10, 50, 100, 200 |
| max_features | auto, sqrt, log2 |
| max_depth | None, 10, 20, 30, 40, 50 |
| min_samples_split | 2, 5, 10 |
| min_samples_leaf | 1, 2, 4 |

5. Feature Selection

Pemilihan fitur adalah tahap preprocessing yang sangat penting untuk menemukan subset fitur yang optimal dan relevan dari fitur aslinya dengan kriteria tertentu, yang bertujuan untuk meningkatkan kinerja model *machine learning* secara signifikan [17]. Algoritma genetika (GA) merupakan salah satu metode yang efektif dalam pemilihan fitur, mencari subset fitur optimal untuk meningkatkan kinerja klasifikasi. GA adalah metode evolusioner yang efisien dan terinspirasi dari proses evolusi biologis kromosom [18].

GA dimulai dengan membuat sekelompok solusi acak yang disebut kromosom. Setiap kromosom mewakili cara potensial untuk memecahkan masalah. GA bekerja dengan menciptakan kromosom baru yang lebih baik melalui evaluasi dan perpaduan kromosom yang paling baik. Pada setiap generasi, pasangan kromosom dipilih sebagai orang tua menggunakan metode turnamen atau roda roulette berdasarkan kinerja mereka. Operasi crossover dan mutasi genetik digunakan untuk menciptakan keturunan baru untuk generasi berikutnya. Crossover menukar bagian-bagian gen antara dua kromosom, sementara mutasi mengubah nilai gen secara acak. Proses ini berulang secara iteratif melalui pemilihan, crossover, dan mutasi hingga kriteria terminasi terpenuhi.

Dalam konteks pemilihan fitur dengan GA, setiap kelompok fitur direpresentasikan oleh individu (kromosom) menggunakan angka biner, di mana 1 menunjukkan fitur yang dipilih dan 0 menunjukkan fitur yang tidak dipilih. Langkah selanjutnya melibatkan pembuatan fungsi kebugaran yang menggunakan akurasi klasifikasi sebagai ukuran evaluasi. Pendekatan ini termasuk dalam kategori wrapper karena individu dipilih berdasarkan kinerja dalam fungsi kebugaran. Operator genetika seperti crossover (pertukaran gen) dan mutasi digunakan untuk menciptakan kromosom baru. Pada tahap crossover, dua kromosom induk dipilih untuk menciptakan kromosom baru dengan kombinasi genetika dari kedua orang tua. Mutasi pada GA untuk pemilihan fitur melibatkan perubahan nilai gen secara acak, kadang-kadang mengubah 1 menjadi 0 atau sebaliknya. Tujuan dari langkah ini adalah meningkatkan variasi dalam populasi, mendukung evolusi fitur yang lebih baik seiring berjalannya waktu.

Dalam penelitian ini, hyperparameter tuning diterapkan pada GA-FS seperti ditunjukkan pada Tabel 3 [8].

**Tabel 3**. Hyperparameter Tuning GA-FS

| Parameter | Nilai |
|---|---|
| Generations | 30 |
| Population | 50 |
| Crossover | 0.5 |
| Mutation | 0.07 |

6. Confusion Matrix

Model prediksi yang dibangun dievaluasi menggunakan confusion matrix. Confusion Matrix merupakan sebuah matrik dua dimensi yang menggambarkan perbandingan antara hasil prediksi dengan kelas data sebenarnya, kemudian menghitung jumlah prediksi yang benar dan yang salah pada setiap kategori kelas [19]. Confusion Matrix terdiri dari empat elemen yang membantu menilai seberapa baik kinerja suatu model klasifikasi yaitu True Positive (TP): Hasil prediksi benar, menyatakan bahwa pasien memiliki penyakit jantung. True Negative (TN): Hasil prediksi benar, menyatakan bahwa pasien tidak memiliki penyakit jantung. False Negative (FN): Hasil prediksi salah, menyatakan bahwa pasien sebenarnya memiliki penyakit jantung, dan False Positive (FP): Hasil prediksi salah, menyatakan bahwa pasien sebenarnya tidak memiliki penyakit jantung [8].

Dalam penelitian ini, kinerja model prediksi Random Forest (RF) dievaluasi berdasarkan nilai Accuracy, Sensitivity (Recall), Precision, dan F1-Score [20]. Accuracy menunjukkan seberapa baik model dalam memprediksi dengan tepat, Recall mengukur kemampuan model dalam mengidentifikasi kasus positif secara akurat, Precision mengukur ketepatan prediksi kasus positif, dan F1-Score merupakan rata-rata harmonik dari Precision dan Recall.

## C. Hasil dan Pembahasan

Pengolahan data dalam penelitian ini menggunakan bahasa pemrograman Python dan platform Google Colab mulai dari tahap preprocessing data, hyperparameter tuning, pemilihan fitur, model prediksi, hingga evaluasi model,

dengan memanfaatkan tools (Libraries dan Frameworks) yang disediakan oleh python dan Google Colab.

Pada tahap preprocessing data, dalam proses data cleaning, tidak ditemukan data yang hilang atau duplikat. Selanjutnya, tahap transformasi data menggunakan proses Label Encoding, ditunjukkan pada Gambar 2 (Sebelum proses label encoding) dan Gambar 3 (Setelah proses label encoding).

| | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 40 | M | ATA | 140 | 289 | 0 | Normal | 172 | N | 0.0 | Up | 0 |
| **1** | 49 | F | NAP | 160 | 180 | 0 | Normal | 156 | N | 1.0 | Flat | 1 |
| **2** | 37 | M | ATA | 130 | 283 | 0 | ST | 98 | N | 0.0 | Up | 0 |
| **3** | 48 | F | ASY | 138 | 214 | 0 | Normal | 108 | Y | 1.5 | Flat | 1 |
| **4** | 54 | M | NAP | 150 | 195 | 0 | Normal | 122 | N | 0.0 | Up | 0 |

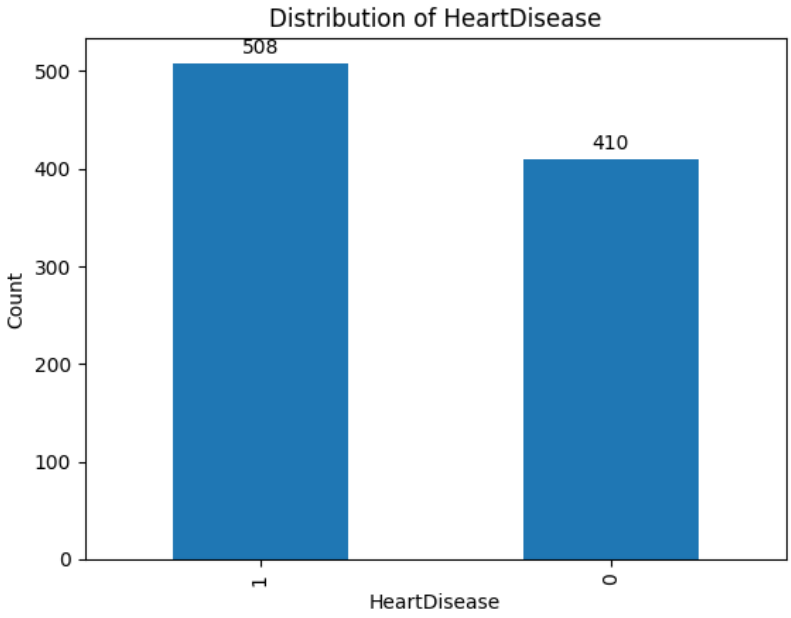
**Gambar 2.** Sebelum Proses Label Encoding

Sebelum proses label encoding, terdapat atribut pada dataset dengan jenis data kategorikal yang memiliki nilai berupa string, mencakup atribut seperti Sex, ChestPainType, RestingECG, ExerciseAngina, dan ST_Slope.

| | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 40 | 1 | 1 | 140 | 289 | 0 | 1 | 172 | 0 | 0.0 | 2 | 0 |
| **1** | 49 | 0 | 2 | 160 | 180 | 0 | 1 | 156 | 0 | 1.0 | 1 | 1 |
| **2** | 37 | 1 | 1 | 130 | 283 | 0 | 2 | 98 | 0 | 0.0 | 2 | 0 |
| **3** | 48 | 0 | 0 | 138 | 214 | 0 | 1 | 108 | 1 | 1.5 | 1 | 1 |
| **4** | 54 | 1 | 2 | 150 | 195 | 0 | 1 | 122 | 0 | 0.0 | 2 | 0 |

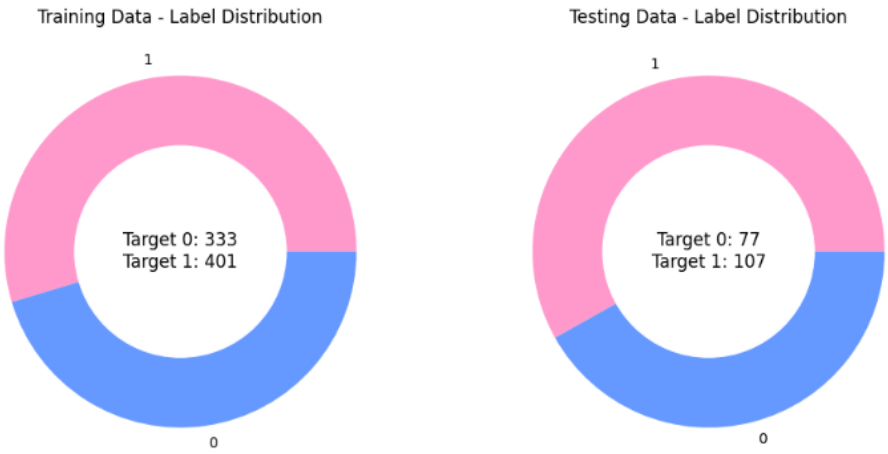
**Gambar 3.** Setelah Proses Label Encoding

Setelah proses label encoding seperti yang ditunjukkan pada Gambar 3, data pada atribut Sex, ChestPainType, RestingECG, ExerciseAngina, dan ST_Slope yang semula berupa tipe kategorikal (string) telah diubah menjadi tipe numerik.



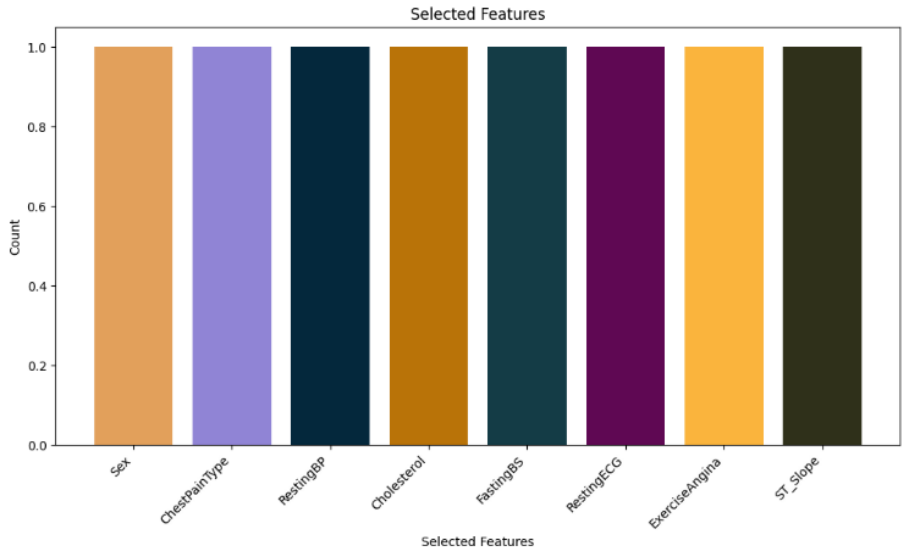

**Gambar 4.** Distribusi Data

Pada Gambar 4, terlihat bahwa jumlah distribusi data dalam Dataset mencapai 918 data, yang terdiri dari 508 data berlabel 1, mengindikasikan bahwa mereka mengalami penyakit jantung, dan 410 data berlabel 0, yang menunjukkan mereka dalam keadaan normal.



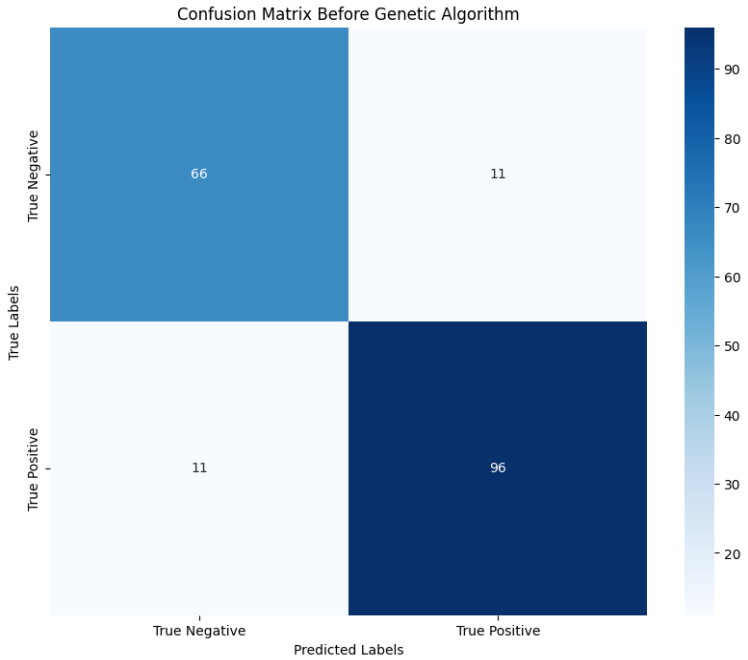

**Gambar 5.** Distribusi Data Training dan Data Testing

Pada tahap pembagian data, total entri data pada data training mencapai 734 data, dengan 333 data pada kelas 0 dan 401 data untuk kelas 1. Selain itu, pada data testing, total entri data mencapai 184 data, dengan 77 entri data pada kelas 0 dan 107 entri untuk data kelas 1.



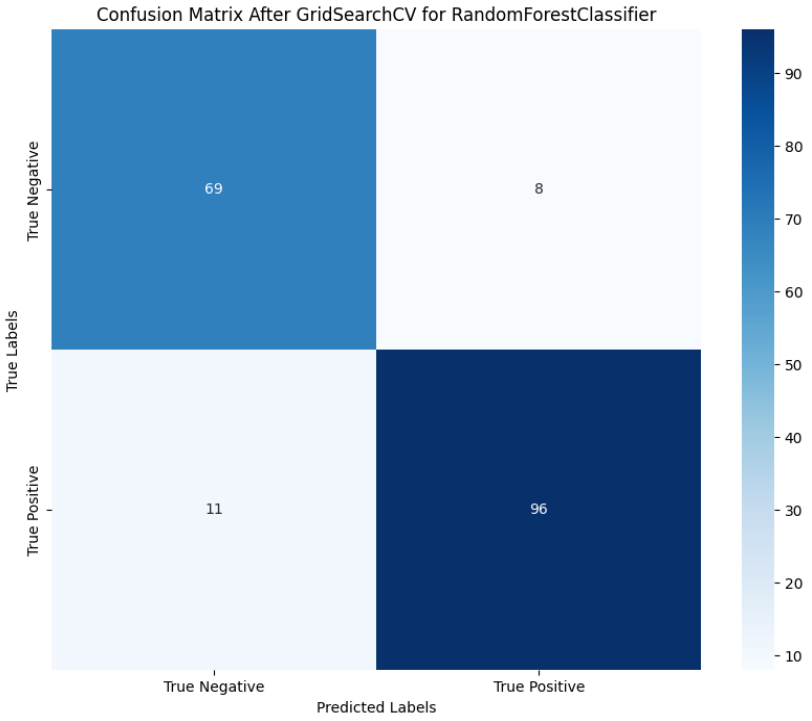

**Gambar 6.** Atribut Terpilih Setelah Proses GA-FS

Setelah melalui tahap seleksi fitur menggunakan algoritma genetika (GA-FS) dan optimalisasi parameter dengan GridSearchCV (GS) pada algoritma Random Forest (RF), hasil penelitian menunjukkan bahwa dari sebelas atribut yang awalnya dimasukkan, terdapat delapan atribut terpilih karena memberikan kontribusi yang relevan dan optimal dalam meningkatkan kinerja model. Atribut-atribut tersebut meliputi: Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, ExerciseAngina, dan ST_Slope. Selain itu, atribut Age, MaxHR, dan Oldpeak dieliminasi karena tidak memberikan kontribusi yang signifikan terhadap kinerja model prediksi.



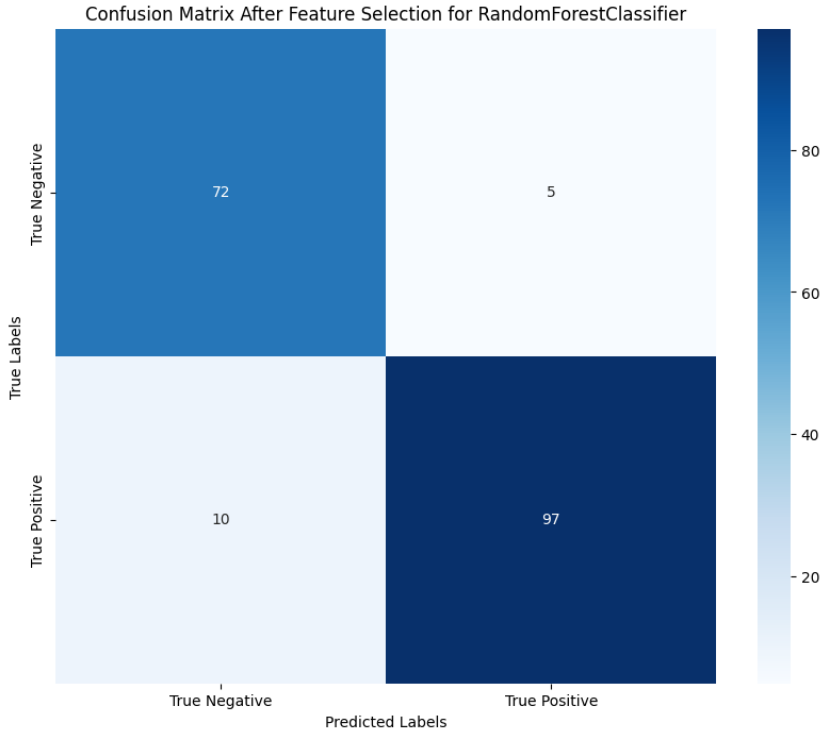

**Gambar 7.** Confusion Matrix Prediksi RF Sebelum GS dan GA-FS

Gambar 7 menyajikan nilai confusion matrix hasil prediksi Random Forest (RF) sebelum proses hyperparameter tuning menggunakan GridSearchCV (GS) dan *Genetic Algoritm Feature Selection* (GA-FS). Hasil tersebut menunjukkan bahwa RF berhasil memprediksi dengan benar 96 data kasus 1 (positif) yang mengindikasikan bahwa mereka mengalami penyakit jantung dan 66 data kasus 0 (negatif) yang mengindikasikan mereka tidak mengalami penyakit jantung. Sedangkan 11 data diprediksi positif yang sebenarnya negatif dan 11 data diprediksi negatif yang sebenarnya positif.




**Gambar 8.** Confusion Matrix Prediksi RF Setelah Optimasi GS

Pada Gambar 8, nilai confusion matrix hasil prediksi Random Forest (RF) setelah proses optimasi parameter menggunakan GridSearchCV (GS) menunjukkan bahwa RF berhasil memprediksi dengan benar 96 data kasus 1 (positif) yang mengindikasikan bahwa mereka mengalami penyakit jantung dan 69 data kasus 0 (negatif) yang mengindikasikan mereka tidak mengalami penyakit jantung. Sedangkan 8 data diprediksi positif yang sebenarnya negatif dan 11 data diprediksi negatif yang sebenarnya positif.




**Gambar 9.** Confusion Matrix Prediksi RF Setelah Optimasi GS dan GA-FS

Gambar 9 menyajikan nilai confusion matrix hasil prediksi Random Forest setelah proses optimasi parameter menggunakan GridSearchCV dan Genetic Algorithm Feature Selection (RF-GS-GAFS). Hasil tersebut menunjukkan bahwa RF berhasil memprediksi dengan benar 97 data kasus 1 (positif) yang mengindikasikan bahwa mereka mengalami penyakit jantung dan 72 data kasus 0 (negatif) yang mengindikasikan mereka tidak mengalami penyakit jantung. Sedangkan 5 data diprediksi positif yang sebenarnya negatif dan 10 data diprediksi negatif yang sebenarnya positif.

**Tabel 4.** Perbandingan Performa RF, RF-GS, RF-GS-GAFS

| ***Performance*** | ***Accuracy*** | ***Precission*** | ***Recall*** | ***F1-Score*** |
|---|---|---|---|---|
| RF | 88.04% | 89.72% | 89.72% | 89.72% |
| RF-GS | 89.67% | 92.31% | 89.72% | 91.00% |
| RF-GS-GAFS | 91.85% | 95.10% | 90.65% | 92.82% |

Dalam Tabel 4, kami membandingkan nilai evaluasi model prediksi random forest (RF) berdasarkan tingkat Accuracy, Precission, Recall, dan F1-Score sebelum proses GridSearchCV (GS), setelah proses GridSearchCV (RF-GS), dan setelah proses pemilihan fitur menggunakan Algoritma Genetika (RF-GS-GAFS). Hasil penelitian menunjukkan peningkatan yang cukup signifikan pada nilai evaluasi model setelah proses GS. Akurasi model meningkat dari 88.04% menjadi 89.67%, sementara nilai Precission dan F1-Score juga mengalami peningkatan masing-masing dari 89.72% menjadi 92.31% dan dari 89.72% menjadi 91.00%. Meskipun

begitu, nilai Recall tidak mengalami perubahan, yaitu 89.72%. Selanjutnya, setelah proses RF-GS-GAFS, terjadi peningkatan yang sangat signifikan pada Accuracy, Precission, Recall, dan F1-Score, masing-masing mencapai 91.85%, 95.10%, 90.65%, dan 92.82%. Dengan demikian, dapat disimpulkan bahwa optimasi parameter dengan GS dan pemilihan fitur menggunakan GA sangat efektif dalam meningkatkan kinerja model prediksi RF.

## D. Kesimpulan

Berdasarkan hasil dan pembahasan penelitian, dapat disimpulkan bahwa penerapan metode GridSearchCV (GS) dan Genetic Algorithm Feature Selection (GA-FS) berhasil meningkatkan kinerja model Random Forest (RF) dalam memprediksi penyakit jantung. Dari sebelas atribut awal, delapan atribut terpilih, seperti Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, ExerciseAngina, dan ST_Slope, memberikan kontribusi yang relevan dan optimal dalam meningkatkan kinerja model. Sebaliknya, atribut lainnya, yakni Age, MaxHR, dan Oldpeak, dieliminasi karena tidak memberikan kontribusi yang signifikan. Metode RF-GS-GAFS mencapai akurasi 91.85%, mengungguli model RF-GS (89.67%) dan RF tanpa optimalisasi (88.04%). Temuan ini menunjukkan bahwa kombinasi optimasi parameter dan pemilihan fitur menggunakan algoritma genetika sangat efektif dalam meningkatkan kinerja model prediksi penyakit jantung.

Pada penelitian berikutnya, disarankan menggunakan metode optimasi seperti RandomSearch atau PSO. Pada tahap preprocessing data, perlu dilakukan normalisasi data dan proses One-Hot Encoding. Untuk menangani ketidakseimbangan data, direkomendasikan menggunakan teknik undersampling atau oversampling. Selain itu, penting untuk membandingkan berbagai metode seleksi fitur seperti correlaiton-based dan information gain. Disarankan juga membandingkan berbagai model prediksi seperti SVM, Decision Tree, dan Multilayer Perceptron untuk menemukan model terbaik dalam prediksi penyakit jantung.

## E. Ucapan Terima Kasih

Kami mengucapkan banyak terima kasih kepada pihak-pihak yang telah memberikan dukungan terhadap penelitian ini.

## F. Referensi

[1] R. Buettner and M. Schunter, "Efficient machine learning based detection of heart disease," 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom), Bogota, Colombia, 2019, pp. 1-6, doi: 10.1109/HealthCom46333.2019.9009429.

[2] P. Gupta and D. Seth, “Comparative analysis and feature importance of machine learning and deep learning for heart disease prediction,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 29, no. 1, p. 451, Jan. 2022, doi: https://doi.org/10.11591/ijeecs.v29.i1.pp451-459.

[3] A. Prasetyo, Erick, D. A. Mulia, & K. K. Octavina, “Comparative Study of Heart Failure Prediction Algorithm: Logistic Regression and SVM”, International Journal of Intelligent Systems and Applications in Engineering, 11(2), 2023.

[4] M. Zeng, "The Prediction of Heart Failure based on Four Machine Learning Algorithms," Highlights in Science, Engineering and Technology, vol. 39, pp. 1377–1382, Apr. 2023, doi: https://doi.org/10.54097/hset.v39i.6771.
[5] M. T. García-Ordás, M. Bayón-Gutiérrez, C. Benavides, J. Aveleira-Mata, and J. A. Benítez-Andrades, "Heart disease risk prediction using deep learning techniques with feature augmentation," Multimedia Tools and Applications, Mar. 2023, doi: https://doi.org/10.1007/s11042-023-14817-z.
[6] Ş. Ay, E. Ekinci, and Z. Garip, "A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases," The Journal of Supercomputing, Mar. 2023, doi: https://doi.org/10.1007/s11227-023-05132-3.
[7] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, vol. 16, no. 2, p. 88, Feb. 2023, doi: https://doi.org/10.3390/a16020088.
[8] M. G. El-Shafiey, A. Hagag, E.-S. A. El-Dahshan, and M. A. Ismail, "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest," Multimedia Tools and Applications, vol. 81, no. 13, pp. 18155–18179, Mar. 2022, doi: https://doi.org/10.1007/s11042-022-12425-x.
[9] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," Healthcare Analytics, p. 100060, May 2022, doi: https://doi.org/10.1016/j.health.2022.100060.
[10] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "Improving the Heart Disease Detection and Patients' Survival Using Supervised Infinite Feature Selection and Improved Weighted Random Forest," IEEE Access, vol. 10, pp. 67363–67372, Jan. 2022, doi: https://doi.org/10.1109/access.2022.3185129.
[11] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," Journal of Big Data, vol. 9, no. 1, Feb. 2022, doi: https://doi.org/10.1186/s40537-022-00573-8.
[12] Fedesoriano (2021) Heart failure prediction dataset. https://www.kaggle.com/fedesoriano/heart-failure-prediction, Sept 2021.
[13] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," Measurement: Sensors, p. 100605, Dec. 2022, doi: https://doi.org/10.1016/j.measen.2022.100605.
[14] M. A. Talukder, K. F. Hasan, M. M. Islam, M. A. Uddin, A. Akhter, M. A. Yousuf, F. Alharbi, anda M. A. Moni, "A dependable hybrid machine learning model for network intrusion detection," Journal of Information Security and Applications, vol. 72, p. 103405, Feb. 2023, doi: https://doi.org/10.1016/j.jisa.2022.103405.
[15] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," Technometrics, pp. 1–23, Apr. 2021, doi: https://doi.org/10.1080/00401706.2021.1921037.
[16] M. Y. Shams, A. M. Elshewey, E.-S. M. El-kenawy, A. Ibrahim, F. M. Talaat, and Zahraa Tarek, "Water quality prediction using machine learning models based on grid search method," Multimedia Tools and Applications, Sep. 2023, doi: https://doi.org/10.1007/s11042-023-16737-4.

[17] W. Ali and A. Ahmed, “Hybrid Intelligent Phishing Website Prediction Using Deep Neural Networks with Genetic Algorithm-based Feature Selection and Weighting,” IET Information Security, Jul. 2019, doi: https://doi.org/10.1049/iet-ifs.2019.0006.
[18] W. Ali and F. Saeed, “Hybrid Filter and Genetic Algorithm-Based Feature Selection for Improving Cancer Classification in High-Dimensional Microarray Data,” Processes, vol. 11, no. 2, p. 562, Feb. 2023, doi: https://doi.org/10.3390/pr11020562.
[19] T. Gori, A. Sunyoto, and H. A. Fatta, “Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa,” Jurnal Teknologi Informasi dan Ilmu Komputer/Jurnal teknologi informasi dan ilmu komputer, vol. 11, no. 1, pp. 215–224, Feb. 2024, doi: https://doi.org/10.25126/jtiik.20241118074.
[20] D. Valero-Carreras, J. Alcaraz, and M. Landete, “Comparing two SVM models through different metrics based on the confusion matrix,” Computers & Operations Research, vol. 152, p. 106131, Apr. 2023, doi: https://doi.org/10.1016/j.cor.2022.106131.