Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 5, No. 7, Juli 2025, Hal. 1965-1979
DOI: https://doi.org/10.52436/1.jpti.855
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Optimasi Random Forest untuk Prediksi Penyakit Jantung Menggunakan SMOTEENN dan Grid Search

Akbar Eka Pranajaya[1], Erliyan Redy Susanto[*2]

[1,2]Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia, Indonesia
Email: [1]akbar_eka_pranajaya@teknokrat.ac.id, [2]erliyan.redy@teknokrat.ac.id

Abstrak

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia, dengan sekitar 17,9 juta kematian setiap tahun. Diagnosis dini dan akurat sangat penting untuk pengobatan yang efektif, namun ketidakseimbangan kelas dalam dataset medis sering menyebabkan bias pada model prediktif, khususnya dalam mengidentifikasi pasien dengan penyakit jantung (kelas minoritas). Studi ini bertujuan untuk mengoptimalkan kinerja algoritma Random Forest dalam memprediksi penyakit jantung dengan mengatasi ketidakseimbangan data menggunakan teknik SMOTEENN *(Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors*) serta penyetelan hiperparameter melalui GridSearchCV. Dataset dibagi menjadi data pelatihan (80%) dan pengujian (20%), dengan evaluasi kinerja menggunakan metrik akurasi, presisi, recall, spesifisitas, F1-score, dan AUC ROC. Hasil penelitian menunjukkan bahwa model yang dioptimalkan mencapai akurasi sebesar 94%, presisi 87%, recall 100%, spesifisitas 91%, F1-score 93%, dan AUC sebesar 0,99. Teknik SMOTEENN terbukti efektif dalam meningkatkan representasi kelas minoritas tanpa menimbulkan noise yang signifikan, sementara GridSearchCV berhasil menemukan kombinasi hiperparameter terbaik untuk meningkatkan performa model. Model Random Forest yang dihasilkan menunjukkan potensi tinggi sebagai alat bantu diagnosis dini penyakit jantung, yang dapat berkontribusi dalam menurunkan angka kematian dan meningkatkan efisiensi biaya perawatan.

Kata kunci: *Penyakit jantung, Random Forest, SMOTEENN, GridSearchCV, klasifikasi medis.*

Random Forest Optimization for Heart Disease Prediction Using SMOTEENN and Grid Search

Abstract

Heart disease is one of the leading causes of death in the world, with about 17.9 million deaths each year. Early and accurate diagnosis is essential for effective treatment, but class imbalances in medical datasets often lead to bias in predictive models, particularly in identifying patients with heart disease (minority class). This study aims to optimize the performance of the Random Forest algorithm in predicting heart disease by overcoming data imbalances using the SMOTEENN (Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors) technique as well as hyperparameter tuning through GridSearchCV. The dataset was divided into training (80%) and testing (20%) data, with performance evaluations using accuracy, precision, recall, specificity, F1-score, and AUC ROC metrics. The results showed that the optimized model achieved an accuracy of 94%, precision of 87%, recall of 100%, specificity of 91%, F1-score of 93%, and an AUC of 0.99. The SMOTEENN technique proved effective in improving the representation of minority classes without causing significant noise, while GridSearchCV managed to find the best combination of hyperparameters to improve model performance. The resulting Random Forest model shows high potential as an aid in the early diagnosis of heart disease, which can contribute to lowering mortality and improving the cost efficiency of treatment.

Keywords: *Heart disease, Random Forest, SMOTEENN, GridSearchCV, medical classification.*

p-ISSN: 2775-4227
e-ISSN: 2775-4219

1. PENDAHULUAN

Penyakit jantung masih menjadi salah satu penyebab utama kematian di seluruh dunia, menyumbang sekitar 17,9 juta kematian setiap tahun menurut Organisasi Kesehatan Dunia[1]. Diagnosis dini dan akurat terhadap penyakit jantung sangat penting untuk pengobatan yang efektif dan pencegahan komplikasi[2]. Dalam beberapa tahun terakhir, *machine learning* (ML) telah muncul sebagai alat yang powerful untuk analisis prediktif di bidang kesehatan[3], menawarkan potensi untuk meningkatkan akurasi diagnosis dan hasil perawatan pasien. Di antara berbagai algoritma ML, *Random Forest* mendapatkan perhatian signifikan karena ketangguhannya, kemampuan menangani data berdimensi tinggi, dan resistensi terhadap *overfitting*[4]. Namun, kinerja *Random Forest* dan model ML lainnya sering terhambat oleh ketidakseimbangan dataset[5]. Masalah umum dalam dataset medis di mana jumlah individu sehat jauh melebihi jumlah individu dengan penyakit[6]. Ketidakseimbangan ini dapat menyebabkan model yang bias dan performa buruk dalam memprediksi kelas minoritas, yang seringkali paling krusial dalam diagnosis medis[7].

Untuk mengatasi tantangan ini, teknik canggih seperti *Synthetic Minority Over-sampling Technique Edited Nearest Neighbors* (SMOTEENN) telah diusulkan[8].SMOTEENN menggabungkan over-sampling pada kelas minoritas dengan under-sampling pada kelas mayoritas, secara efektif menyeimbangkan dataset dan meningkatkan kinerja model[9]. Selain itu, penyetelan hiperparameter menggunakan metode seperti GridSearchCV telah terbukti mengoptimalkan kinerja model dengan mengeksplorasi ruang hiperparameter secara sistematis[10]. Meskipun potensi teknik-teknik ini cukup besar, penerapan gabungannya dalam konteks prediksi penyakit jantung menggunakan *Random Forest* masih kurang dieksplorasi, sehingga membuka peluang studi yang signifikan[11].

Studi sebelumnya [12]telah menunjukkan efektivitas *Random Forest* dalam berbagai tugas diagnosis medis. Misalnya[13], menerapkan *Random Forest* untuk memprediksi penyakit jantung dengan akurasi 87%, menegaskan potensinya dalam aplikasi kesehatan. Demikian pula, penggunaan teknik resampling seperti SMOTE dan variannya telah banyak dipelajari[14]. melaporkan bahwa SMOTEENN secara signifikan meningkatkan kinerja model ML pada dataset tidak seimbang, mencapai peningkatan 15% dalam recall untuk kelas minoritas[15]. Namun, sebagian besar studi yang ada fokus pada teknik resampling atau penyetelan hiperparameter secara terpisah, dengan eksplorasi terbatas terhadap dampak kombinasi keduanya pada kinerja model[16]. Selain itu, hanya sedikit p yang secara khusus membahas optimasi *Random Forest* untuk prediksi penyakit jantung menggunakan teknik-teknik canggih ini, sehingga meninggalkan celah didalam studi yang cukup mencolok[17].

Studi sebelumnya juga telah mengevaluasi kinerja akurasi dari tiga algoritma machine learning yang berbeda untuk memprediksi penyakit jantung. Dari ketiga algoritma yang diuji, di antaranya, yaitu Support Vector Machine (SVM), Logistic Regression, dan Artificial Neural Network (ANN), mencapai tingkat akurasi yang berbeda beda[12]. Logistic Regression umumnya menampilkan akurasi yang cukup memuaskan, sementara SVM dan Neural Networks cenderung menawarkan performa yang lebih unggul meskipun memerlukan waktu komputasi yang lebih lama. Dalam studi ini, algoritma Random Forest dipilih karena kemampuannya dalam menangani dataset yang rumit, menghasilkan akurasi yang tinggi, serta menyediakan interpretasi yang lebih mudah dipahami dibandingkan dengan metode lainnya.

Tantangan utama dalam studi ini adalah meningkatkan kinerja prediktif algoritma *Random Forest* untuk klasifikasi penyakit jantung, terutama dalam konteks dataset tidak seimbang[18]. Meskipun *Random Forest* secara *inheren robust*, kinerjanya bisa menjadi suboptimal ketika diterapkan pada dataset dengan ketidakseimbangan kelas yang signifikan, seperti yang sering terjadi dalam diagnosis medis[19]. Selain itu, kurangnya penyetelan hiperparameter secara sistematis semakin membatasi potensinya[20]. Studi ini bertujuan untuk mengatasi keterbatasan ini dengan mengintegrasikan SMOTEENN untuk penyeimbangan data dan GridSearchCV untuk optimasi hiperparameter, sehingga meningkatkan akurasi, presisi, dan recall model[21].

Tujuan dari studi ini adalah untuk mengoptimalkan kinerja algoritma *Random Forest* untuk prediksi penyakit jantung dengan memanfaatkan SMOTEENN untuk penyeimbangan data dan GridSearchCV untuk penyetelan hiperparameter[18]. Secara spesifik, studi ini bertujuan untuk mengevaluasi dampak SMOTEENN terhadap kinerja *Random Forest* dalam menangani dataset penyakit jantung yang tidak seimbang, mengidentifikasi hiperparameter optimal untuk *Random Forest* menggunakan GridSearchCV guna memaksimalkan akurasi prediktif, Membandingkan kinerja model *Random Forest* yang dioptimalkan dengan model *baseline* dan teknik *state-of-the-art* lainnya[9].

Studi ini menggunakan pendekatan sistematis untuk mengoptimalkan *Random Forest* dalam prediksi penyakit jantung. Dataset pertama-tama diproses dan diseimbangkan menggunakan SMOTEENN[22]. Selanjutnya, GridSearchCV diterapkan untuk mengidentifikasi hiperparameter optimal bagi model *Random Forest*[23]. Kinerja model yang dioptimalkan dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score, serta dibandingkan dengan model *baseline* dan teknik *state-of-the-art* lainnya[24]. Hasil studi ini

p-ISSN: 2775-4227
e-ISSN: 2775-4219

menunjukan akurasi sebesar 0.94%, menunjukkan keandalan yang tinggi dalam memprediksi hasil dengan benar. Presisi untuk kelas positif tercatat sebesar 0.87%, yang berarti sebagian besar prediksi positif adalah benar. Selain itu, recall mencapai 1.0%, menandakan bahwa model tidak melewatkan satupun kasus penyakit jantung dan Specificity mencapai 0.91%, menunjukan bahwa model dalam mengukur dan mengidentifikasi kasus negatif dengan benar. yang sangat krusial dalam bidang kesehatan. F1-score sebesar 0.93% juga menunjukkan keseimbangan yang baik antara presisi dan recall.

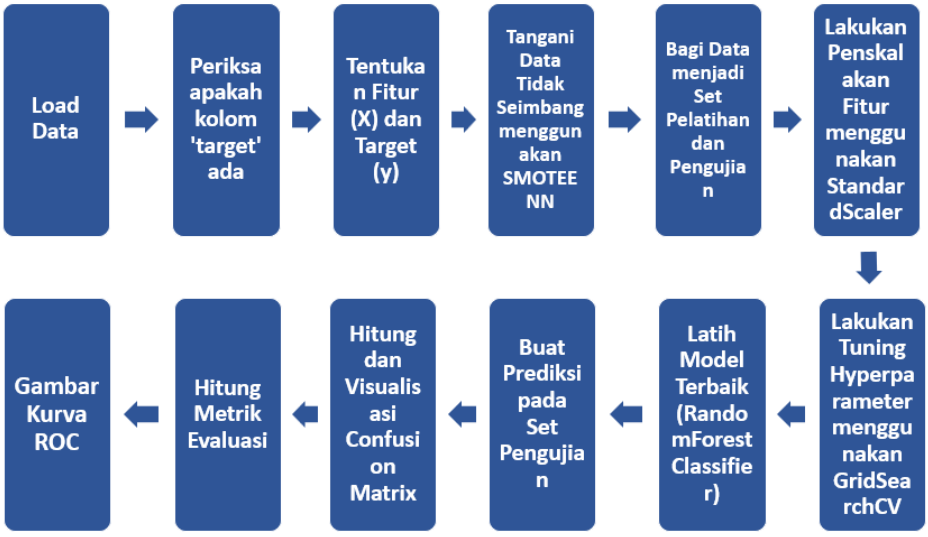
Studi ini berkontribusi pada bidang diagnosis medis dengan menunjukkan efektivitas kombinasi SMOTEENN dan GridSearchCV dalam mengoptimalkan *Random Forest* untuk prediksi penyakit jantung[25]. Temuan ini memberikan wawasan berharga tentang penanganan dataset tidak seimbang dan optimasi hiperparameter, yang dapat diperluas ke aplikasi kesehatan lainnya. Dari perspektif praktis, model yang dioptimalkan memiliki potensi untuk meningkatkan diagnosis dini dan pengobatan penyakit jantung, yang pada akhirnya dapat mengurangi angka kematian dan biaya kesehatan[26]. Secara teoretis, studi ini memperluas pemahaman tentang bagaimana teknik *resampling* dan penyetelan hiperparameter dapat secara sinergis meningkatkan kinerja model ML dalam skenario data tidak seimbang[27].

2. METODE PENELITIAN

Studi ini menggunakan satu metode prediksi pembelajaran *machine learning* dalam prediksi penyakit jantung yaitu metode *Random Forest* dengan Teknik SMOTEENN dan Hyperparameter Tuning menggunakan GridSearchCV.

2.1. Tahap Studi

Berikut tahapan-tahapan dalam penelitian ini dapat dilihat pada Gambar 1.



Gambar 1 Tahap penelitian

2.2. Sumber Data

Dalam melakukan penelitian atau study, keakuratan dan keandalan data menjadi fondasi utama yang menentukan kualitas hasil yang diperoleh. Sumber data akan membahas asal-usul data yang digunakan, metode pengumpulannya, serta validitas informasi yang menjadi dasar pembahasan. Dengan memahami sumber data, pembaca dapat menilai sejauh mana temuan atau kesimpulan yang dihasilkan dapat diandalkan dan relevan. Pada bagian ini, penulis akan menjelaskan secara rinci dari mana data diperoleh, bagaimana proses pengumpulannya dilakukan, serta langkah-langkah yang diambil untuk memastikan integritas data tersebut. Berikut adalah sumber data yang didapat:

- **Pengumpulan Data**

  Langkah pertama dalam studi ini adalah pengumpulan dan persiapan data. Dataset yang digunakan dalam studi ini adalah dataset penyakit jantung yang berisi berbagai fitur klinis yang relevan dengan kondisi kesehatan jantung. Fitur-fitur tersebut mencakup informasi seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, dan hasil tes medis lainnya. Kolom 'target' dalam dataset ini merupakan variabel

p-ISSN: 2775-4227
e-ISSN: 2775-4219

yang menunjukkan adanya penyakit jantung, di mana nilai 1 menunjukkan pasien memiliki penyakit jantung dan nilai 0 menunjukkan pasien tidak memiliki penyakit jantung. Dengan melakukan langkah-langkah persiapan data ini, diharapkan dataset yang digunakan dalam studi ini telah siap untuk proses analisis lebih lanjut, sehingga model yang dibangun dapat memberikan hasil yang akurat dan dapat diandalkan.

- **Pemeriksaan kolom Target**

  Langkah pertama dalam proses preprocessing data adalah memeriksa keberadaan dan integritas kolom 'target' dalam dataset. Kolom 'target' merupakan komponen krusial karena berisi label kelas yang menunjukkan status kesehatan pasien, yaitu apakah seorang pasien memiliki penyakit jantung (biasanya direpresentasikan dengan nilai 1) atau tidak (direpresentasikan dengan nilai 0). Pemeriksaan ini meliputi beberapa aspek penting. Pertama, kami memverifikasi keberadaan kolom 'target' dengan mengecek nama kolom yang tersedia dan memastikan tidak ada kesalahan penulisan atau ketidaksesuaian nama kolom. Setelah memastikan keberadaannya, kami menganalisis distribusi kelas dalam kolom tersebut dengan menghitung jumlah sampel untuk setiap kelas (positif dan negatif) guna mengidentifikasi apakah terdapat ketidakseimbangan kelas, yang dapat mempengaruhi kinerja model jika tidak ditangani dengan teknik yang tepat. Selain itu, kami juga memeriksa apakah terdapat nilai yang hilang (missing value) dalam kolom 'target'. Jika ditemukan, langkah-langkah penanganan seperti penghapusan baris atau imputasi nilai akan dipertimbangkan untuk memastikan kualitas data. Terakhir, kami memastikan bahwa kolom 'target' memiliki tipe data yang sesuai, biasanya dalam bentuk integer atau kategori, untuk memfasilitasi proses pelatihan model. Pemeriksaan ini sangat penting untuk memastikan dataset siap untuk proses selanjutnya, termasuk penanganan ketidakseimbangan kelas dan pelatihan model. Dengan memastikan integritas kolom 'target', kami dapat membangun model yang lebih akurat dan dapat diandalkan untuk prediksi penyakit jantung.

- **Tentukan Fitur (X) dan Target (y)**

  Setelah memastikan keberadaan kolom target dalam dataset, langkah selanjutnya adalah menentukan variabel fitur (X) dan variabel target (y). Variabel fitur terdiri dari semua kolom yang relevan dalam dataset yang digunakan untuk memprediksi status penyakit jantung. Kolom-kolom ini mencakup berbagai parameter klinis seperti usia, jenis kelamin, tingkat kolesterol, tekanan darah, dan indikator lainnya yang secara medis dianggap berpengaruh terhadap kondisi jantung. Variabel target (y) adalah kolom 'target' yang menunjukkan adanya atau tidak adanya penyakit jantung. Kolom ini biasanya berisi nilai biner, di mana 1 menunjukkan pasien memiliki penyakit jantung dan 0 menunjukkan pasien tidak memiliki penyakit jantung. Penentuan variabel fitur dan target ini sangat penting karena akan menjadi dasar bagi model untuk mempelajari pola dan hubungan antara fitur-fitur tersebut dengan kondisi kesehatan jantung. Proses ini juga melibatkan pemeriksaan lebih lanjut terhadap data untuk memastikan tidak ada nilai yang hilang (missing values) atau ketidakonsistenan dalam kolom target. Jika ditemukan masalah, langkah preprocessing seperti imputasi atau penghapusan data yang tidak valid akan dilakukan. Dengan demikian, dataset yang digunakan untuk pelatihan dan pengujian model akan lebih bersih dan siap untuk proses analisis lebih lanjut.

- **Penanganan Data Tidak Seimbang**

  Dataset yang tidak seimbang, di mana proporsi antara kelas mayoritas dan minoritas signifikan, dapat menyebabkan bias dalam model klasifikasi. Model cenderung lebih baik dalam memprediksi kelas mayoritas, sementara performanya pada kelas minoritas kurang optimal. Untuk mengatasi masalah ini, studi ini menggunakan teknik SMOTEENN (*Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors*). SMOTEENN adalah pendekatan hybrid yang menggabungkan dua metode: SMOTE untuk oversampling dan ENN untuk undersampling. SMOTE bekerja dengan menghasilkan sampel sintetis untuk kelas minoritas dengan cara interpolasi antara sampel yang sudah ada. Hal ini membantu meningkatkan representasi kelas minoritas dalam dataset. Namun, oversampling saja dapat menyebabkan noise atau tumpang tindih antara kelas. Oleh karena itu, ENN digunakan untuk membersihkan data dengan menghapus sampel yang mungkin salah diklasifikasikan oleh tetangga terdekatnya, terutama dari kelas mayoritas. Dengan menggabungkan kedua teknik ini, SMOTEENN tidak hanya menyeimbangkan distribusi kelas tetapi juga meningkatkan kualitas data dengan mengurangi noise dan ambiguitas. Hasilnya adalah dataset yang lebih seimbang dan bersih, yang diharapkan dapat

p-ISSN: 2775-4227
e-ISSN: 2775-4219

meningkatkan kinerja model klasifikasi, khususnya dalam mendeteksi kasus-kasus penyakit jantung yang termasuk dalam kelas minoritas.

Proses ini dilakukan sebelum pembagian dataset menjadi set pelatihan dan pengujian untuk memastikan bahwa model dilatih pada data yang representatif dan tidak bias terhadap kelas tertentu. Dengan demikian, model dapat belajar pola dari kedua kelas secara lebih efektif, yang pada akhirnya akan meningkatkan akurasi dan generalisasi model pada data yang belum pernah dilihat sebelumnya.

- **Pembagian Data**

Setelah data diproses dan diseimbangkan menggunakan teknik SMOTEENN, langkah selanjutnya adalah membagi dataset menjadi dua subset utama, yaitu set pelatihan (training set) dan set pengujian (testing set). Pembagian ini dilakukan dengan tujuan untuk memastikan bahwa model dapat dievaluasi pada data yang belum pernah dilihat sebelumnya, sehingga memberikan gambaran yang lebih akurat tentang kemampuan generalisasi model. Dalam studi ini, dataset dibagi dengan proporsi 80% untuk set pelatihan dan 20% untuk set pengujian. Proporsi ini dipilih karena dianggap sebagai rasio yang umum digunakan dalam banyak studi dan memberikan keseimbangan yang baik antara jumlah data yang cukup untuk melatih model dan data yang cukup untuk menguji performa model. Set pelatihan digunakan untuk melatih model, sementara set pengujian digunakan untuk mengevaluasi seberapa baik model dapat memprediksi data baru. Pembagian data dilakukan secara acak (random splitting) untuk memastikan bahwa distribusi kelas dalam set pelatihan dan pengujian tetap seimbang. Hal ini penting untuk menghindari bias dalam pelatihan dan evaluasi model. Selain itu, stratifikasi (stratified splitting) juga dapat diterapkan untuk memastikan bahwa proporsi kelas target dalam set pelatihan dan pengujian sama dengan proporsi dalam dataset asli. Ini terutama berguna ketika dataset memiliki ketidakseimbangan kelas yang signifikan. Dengan membagi dataset secara tepat, kita dapat memastikan bahwa model tidak hanya bekerja baik pada data pelatihan tetapi juga mampu menggeneralisasi dengan baik pada data yang belum pernah dilihat, yang merupakan indikator penting dari kinerja model dalam aplikasi dunia nyata.

- **Penskalaan Fitur**

Fitur dalam dataset sering kali memiliki skala yang berbeda-beda, yang dapat menyebabkan ketidakseimbangan dalam kontribusi masing-masing fitur terhadap model. Sebagai contoh, beberapa fitur mungkin memiliki nilai yang sangat besar, sementara yang lain memiliki nilai yang relatif kecil. Perbedaan skala ini dapat memengaruhi performa model, terutama pada algoritma yang bergantung pada jarak atau gradient, seperti *Random Forest*. Untuk mengatasi masalah ini, dilakukan proses penskalakan fitur menggunakan *StandardScaler. StandardScaler* adalah teknik penskalakan yang mentransformasikan fitur sehingga memiliki mean (rata-rata) nol dan deviasi standar satu. Proses ini dilakukan dengan mengurangi mean dari setiap fitur dan kemudian membaginya dengan deviasi standar. Secara matematis, transformasi ini dapat dinyatakan sebagai:

$$\mathcal{Z} = \frac{(x - \mu)}{\sigma} \qquad (1)$$

Dimana $\mathcal{X}$ adalah nilai fitur asli, μ adalah mean dari fitur tersebut, dan σ adalah deviasi standat. Hasil dari penskalakan ini adalah fitur yang memiliki distribusi dengan mean nol dan variansi satu, yang membantu dalam meningkatkan stabilitas dan kecepatan konvergensi selama proses pelatihan model. Dengan melakukan penskalakan fitur, kita memastikan bahwa semua fitur berkontribusi secara seimbang terhadap model, sehingga menghindari dominasi fitur dengan skala yang lebih besar. Hal ini sangat penting untuk meningkatkan akurasi dan kinerja model secara keseluruhan, terutama ketika menggunakan algoritma yang sensitif terhadap skala data seperti *Random Forest*. Selain itu, penskalakan juga membantu dalam interpretasi hasil model, karena semua fitur sekarang berada pada skala yang sama.

- **Hyperparameter Tuning dengan GridSearchCV**

Untuk meningkatkan kinerja model *Random Forest Classifier*, dilakukan proses tuning hyperparameter menggunakan teknik GridSearchCV. Hyperparameter adalah parameter yang nilainya ditentukan sebelum proses pelatihan model, dan pemilihan nilai yang tepat dapat secara signifikan memengaruhi performa model. GridSearchCV adalah metode yang sistematis untuk mencari kombinasi hyperparameter terbaik dengan melakukan pencarian grid (*grid search*) pada ruang parameter yang telah ditentukan. GridSearchCV bekerja dengan mencoba semua kombinasi hyperparameter yang telah ditentukan dalam *grid*. Untuk setiap kombinasi, model dilatih menggunakan validasi silang (*cross-validation*) untuk memastikan bahwa performa model tidak hanya bergantung pada satu pembagian data.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Proses ini menghasilkan kombinasi hyperparameter yang memberikan skor validasi terbaik, yang kemudian digunakan untuk melatih model akhir. Dengan menggunakan GridSearchCV, studi ini bertujuan untuk menemukan konfigurasi hyperparameter yang optimal, sehingga model *Random Forest* dapat mencapai akurasi yang tinggi dan generalisasi yang baik pada dataset penyakit jantung. Hasil dari proses tuning ini diharapkan dapat meningkatkan kemampuan model dalam memprediksi penyakit jantung dengan lebih akurat dan andal.

- **Pelatihan Model**

  Model yang digunakan dalam studi ini adalah Random Forest Classifier, sebuah algoritma ensemble learning yang membangun banyak pohon keputusan selama pelatihan dan menggabungkan prediksinya melalui voting untuk klasifikasi. Algoritma ini dipilih karena kemampuannya menangani data kompleks, mengurangi risiko overfitting, dan memberikan akurasi tinggi, bahkan pada dataset berdimensi besar. Proses pelatihan melibatkan penskalakan data dan penanganan ketidakseimbangan kelas, di mana setiap pohon dilatih pada subset data acak (bootstrap sampling). Parameter seperti jumlah pohon (n_estimators), kedalaman maksimum (max_depth), dan kriteria pemisahan (Gini impurity atau entropy) diatur untuk mengoptimalkan performa model. Setelah pelatihan, model dievaluasi menggunakan set pelatihan dan pengujian untuk memastikan generalisasi yang baik dan menghindari overfitting. Dengan menggunakan *Random Forest Classifier*, studi ini bertujuan untuk memanfaatkan kekuatan ensemble learning dalam meningkatkan akurasi prediksi dan mengurangi varians model, sehingga menghasilkan prediksi yang lebih stabil dan dapat diandalkan untuk kasus penyakit jantung.

- **Prediksi pada Set Pengujian**

  Setelah model *Random Forest Classifier* dilatih dan dioptimalkan melalui proses hyperparameter tuning menggunakan GridSearchCV, langkah selanjutnya adalah melakukan prediksi pada set pengujian. Set pengujian ini terdiri dari data yang tidak digunakan selama proses pelatihan, sehingga dapat memberikan gambaran yang lebih akurat tentang kemampuan generalisasi model. Prediksi dilakukan dengan menerapkan model yang telah dilatih ke dalam fitur-fitur pada set pengujian. Hasil prediksi ini kemudian dibandingkan dengan label aktual yang terdapat dalam set pengujian untuk mengevaluasi keakuratan model. Perbandingan antara prediksi dan label aktual memungkinkan untuk menghitung berbagai metrik evaluasi, seperti akurasi, presisi, recall, dan F1-score, yang memberikan wawasan mendalam tentang performa model dalam mengklasifikasikan kasus penyakit jantung. Selain itu, analisis lebih lanjut dapat dilakukan dengan memeriksa kasus-kasus yang salah diklasifikasikan (false *positives dan* false *negatives*) untuk memahami di mana model mungkin mengalami kesulitan. Informasi ini dapat berguna untuk perbaikan lebih lanjut pada model atau untuk memberikan rekomendasi dalam konteks klinis. Dengan demikian, prediksi pada set pengujian tidak hanya mengukur kinerja model, tetapi juga memberikan dasar untuk interpretasi yang lebih mendalam terhadap hasil yang diperoleh.

- **Perhitungan dan Visualisasi Confusion Matrix**

  langkah selanjutnya adalah Visualisasi Confusion Matrix. Visualisasi *confusion matri*x memungkinkan peneliti untuk melihat secara langsung di mana model mungkin melakukan kesalahan klasifikasi. Misalnya, jika model memiliki banyak false negative, ini mungkin menunjukkan bahwa model kurang sensitif dalam mendeteksi kasus positif. Sebaliknya, jika terdapat banyak false positive, model mungkin terlalu agresif dalam memprediksi kasus positif. Confusion matrix adalah alat evaluasi yang membandingkan prediksi model dengan label aktual data, terdiri dari empat komponen utama: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). TP mengindikasikan jumlah sampel positif yang diprediksi dengan benar, sementara TN menunjukkan jumlah sampel negatif yang diprediksi dengan benar. FP terjadi ketika sampel negatif salah diprediksi sebagai positif (kesalahan tipe I), yang dapat menyebabkan tindakan medis yang tidak perlu. Sementara itu, FN terjadi ketika sampel positif salah diprediksi sebagai negatif (kesalahan tipe II), yang berpotensi berbahaya karena pasien tidak mendapatkan perawatan yang diperlukan. Visualisasi confusion matrix membantu peneliti mengidentifikasi area kesalahan klasifikasi, seperti kecenderungan model menghasilkan banyak false negative (kurang sensitif) atau false positive (terlalu agresif). Dengan menganalisis metrik-metrik ini, peneliti dapat menilai keefektifan model dalam memprediksi penyakit jantung, mengidentifikasi kelemahan, dan merencanakan perbaikan lebih lanjut, seperti penyesuaian hyperparameter atau teknik preprocessing data yang lebih canggih.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

- **Perhitungan Metrik Evaluasi**

  Setelah confusion matrix dibuat, langkah selanjutnya adalah menghitung beberapa metrik evaluasi yang penting untuk mengukur kinerja model klasifikasi. Metrik-metrik ini memberikan pemahaman yang lebih mendalam tentang seberapa baik model bekerja, terutama dalam konteks dataset yang tidak seimbang seperti dataset penyakit jantung. Akurasi adalah metrik yang paling umum digunakan, mengukur seberapa sering model melakukan prediksi yang benar. **Akurasi** dihitung sebagai proporsi prediksi yang benar (True Positive + True Negative) terhadap total jumlah sampel. Meskipun akurasi memberikan gambaran umum tentang performa model, pada dataset yang tidak seimbang, metrik ini mungkin tidak cukup karena dapat memberikan gambaran yang menyesatkan jika satu kelas dominan. **Presisi** mengukur seberapa akurat model dalam memprediksi kelas positif, dihitung sebagai proporsi prediksi positif yang benar (True Positive) terhadap total prediksi positif (True Positive + False Positive). Presisi sangat penting dalam konteks di mana False Positive memiliki konsekuensi serius, seperti dalam diagnosis penyakit jantung, di mana False Positive dapat menyebabkan pasien yang sehat menjalani perawatan yang tidak perlu. **Recall**, di sisi lain, mengukur seberapa baik model dapat mengidentifikasi semua sampel positif yang ada, dihitung sebagai proporsi sampel positif yang benar diprediksi (True Positive) terhadap total sampel positif aktual (True Positive + False Negative). Recall sangat kritis dalam konteks medis karena False Negative dapat berakibat fatal, seperti gagal mendeteksi pasien yang menderita penyakit jantung. Untuk menyeimbangkan antara presisi dan recall, kami menggunakan **F1-Score**, yang merupakan rata-rata harmonik dari kedua metrik tersebut. F1-Score sangat berguna pada dataset yang tidak seimbang karena memberikan gambaran yang lebih baik tentang performa model ketika ada trade-off antara False Positive dan False Negative. Selain itu, **Specificity** mengukur seberapa baik model dapat mengidentifikasi sampel negatif, dihitung sebagai proporsi sampel negatif yang benar diprediksi (True Negative) terhadap total sampel negatif aktual (True Negative + False Positive). Specificity penting dalam konteks di mana False Positive perlu diminimalkan, seperti dalam skrining medis. Terakhir, **ROC-AUC** adalah metrik yang mengukur kemampuan model untuk membedakan antara kelas positif dan negatif. ROC-AUC dihitung dengan memplot True Positive Rate (Recall) terhadap False Positive Rate (1 - Specificity) dan menghitung area di bawah kurva (AUC). Nilai AUC yang mendekati 1 menunjukkan model yang sangat baik dalam membedakan antara kelas positif dan negatif, sedangkan nilai AUC yang mendekati 0.5 menunjukkan model yang tidak lebih baik dari tebakan acak. Dengan menghitung metrik-metrik ini, kami dapat mengevaluasi performa model secara komprehensif dan mengidentifikasi area yang perlu ditingkatkan untuk mencapai hasil yang lebih akurat dan dapat diandalkan dalam memprediksi penyakit jantung.

- **Kurva ROC**

  (*Receiver Operating Characteristic*) merupakan alat visual yang digunakan untuk mengevaluasi performa model klasifikasi, khususnya dalam membedakan antara kelas positif dan negatif. Kurva ini memplot **True Positive Rate (TPR)** atau ***sensitivity*** terhadap **False Positive Rate (FPR)** atau **1** - *specificity* pada berbagai *threshold* klasifikasi. True Positive Rate mengukur proporsi instance positif yang diklasifikasikan dengan benar, sedangkan False Positive Rate mengukur proporsi instance negatif yang salah diklasifikasikan sebagai positif. Area di bawah kurva ROC (AUC-ROC) adalah metrik kuantitatif yang digunakan untuk mengukur seberapa baik model dapat membedakan antara kedua kelas. Nilai AUC-ROC berkisar antara 0 dan 1, di mana nilai 1 menunjukkan model yang sempurna (mampu membedakan semua instance positif dan negatif dengan benar), sedangkan nilai 0.5 menunjukkan model yang tidak lebih baik daripada tebakan acak. Dalam studi ini, kurva ROC digunakan untuk memvisualisasikan trade-off antara TPR dan FPR pada berbagai *threshold*, sehingga memberikan gambaran yang komprehensif tentang kemampuan model dalam memprediksi penyakit jantung. Selain itu, analisis kurva ROC membantu mengidentifikasi *threshold* optimal yang dapat digunakan untuk memaksimalkan TPR sambil meminimalkan FPR, tergantung pada tujuan aplikasi. Misalnya, dalam konteks medis seperti prediksi penyakit jantung, mungkin lebih penting untuk memaksimalkan TPR (mendeteksi sebanyak mungkin kasus positif) meskipun dengan sedikit peningkatan FPR, karena biaya dari false negative (missed diagnosis) bisa sangat tinggi. Dengan memvisualisasikan kurva ROC, studi ini tidak hanya mengevaluasi performa model secara keseluruhan tetapi juga memberikan wawasan tentang bagaimana model berperilaku pada berbagai tingkat kepastian klasifikasi. Hal ini sangat berguna untuk memahami kelebihan dan kekurangan model dalam konteks prediksi penyakit jantung.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

2.3. Variabel

Dalam studi ini ada empat belas variabel untuk mendiagnosa penyakit jantung. dataset yang digunakan dalam studi ini terdiri dari beberapa variabel yang merepresentasikan faktor-faktor medis yang berhubungan dengan penyakit jantung. menampilkan deskripsi variabel dalam dataset yang digunakan, termasuk karakteristik demografis, hasil pemeriksaan medis, serta label target yang menunjukkan ada atau tidaknya indikasi penyakit jantung, berikut Tabel 1.

Tabel 1. Variabel menampilkan fitur-fitur utama yang digunakan dalam studi ini.

Variabel	Keterangan
Age	Usia pasien (dalam tahun)
Sex	Jenis kelamin (1 = laki-laki, 0 = perempuan)
Chest_pain_type	Tipe nyeri dada (0-3, menunjukkan jenis nyeri dada)
Resting_bp	Tekanan darah saat istirahat (mmHg)
Cholestoral	Kadar kolesterol dalam darah (mg/dl)
Fasting_blood_sugar	Gula darah puasa (> 120 mg/dl, 1 = benar, 0 = salah)
Restecg	Hasil elektrokardiografi saat istirahat (0-2, menunjukkan kelainan)
Max_hr	Denyut jantung maksimum yang dicapai
Exang	Angina akibat olahraga (1 = ya, 0 = tidak)
Oldpeak	Depresi ST akibat olahraga dibandingkan dengan saat istirahat
Slope	Kemiringan segmen ST saat puncak olahraga (0-2)
Num_major_vessels	Jumlah pembuluh darah utama yang terdeteksi (0-3)
Thal	Status thalassemia (1-3, hasil tes thallium)
Target	Indikasi penyakit jantung (1 = ada, 0 = tidak ada)

2.4. *Random Forest*

Random Forest adalah model *ensemble* yang menggabungkan banyak ***Decision Tree*** untuk meningkatkan akurasi dan mengurangi *overfitting*[28]. Algoritma ini merupakan pengembangan dari metode *Decision Tree*, di mana alih-alih menggunakan satu pohon keputusan (*Decision Tree*), *Random Forest* membangun banyak pohon (hutan) untuk membuat prediksi yang lebih akurat dan robust[29].

A. *Decision Tree* dalam *Random Forest*

Setiap pohon keputusan dalam Random Forest mengikuti aturan pemisahan berdasarkan **entropy** atau **Gini Index**.

a) ***Entropy (Information Gain)***

Entropy mengukur impurity (ketidakteraturan) dalam dataset, dengan rumus:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \qquad (2)$$

Keterangan:
- S = himpunan data
- c = jumlah kelas dalam dataset
- p_i = probabilitas suatu kelas iii

Information Gain (IG), digunakan untuk memilih fitur terbaik dalam pemisahan node:

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \qquad (3)$$

Keterangan:
- A = fitur yang diuji
- S_u = subset data setelah dibagi berdasarkan fitur A

Semakin tinggi IG , semakin baik fitur tersebut dalam membagi data.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

b) Gini Index
Alternatif lain untuk memilih fitur adalah ***Gini Impurity***, yang didefinisikan sebagai:

$$Gini(S) = 1 - \sum_{i-1}^{c} p_i^2 \quad (4)$$

Semakin rendah nilai **Gini**, semakin "murni" data dalam node.

B. *Voting* dalam *Random Forest*

Prediksi akhir dalam *Random Forest* diperoleh dari **mayoritas suara** (***majority voting***) dari semua pohon:

$$\hat{y} = mode\{T_1(X), T_2(X), \dots, T_N(X)\} \quad (5)$$

Keterangan:
- $T_i(X) =$ prediksi dari pohon ke-iii
- $\hat{y} =$ hasil akhir prediksi

2.5. SMOTEENN

SMOTEENN (*Synthetic Minority Oversampling Technique and Edited Nearest Neighbor*) adalah teknik pengolahan data yang digunakan untuk mengatasi ketidakseimbangan data (*imbalanced data*), yang sering terjadi ketika jumlah sampel dalam kelas minoritas jauh lebih sedikit dibandingkan kelas mayoritas[30]. SMOTEENN menggabungkan oversampling kelas minoritas dengan undersampling kelas mayoritas untuk menciptakan distribusi kelas yang lebih seimbang[31].

$$x_{new} = x_i + \lambda(x_k - x_i) \quad (6)$$

Keterangan:

x_i : Sampel minoritas asli
x_k : Tetangga terdekat xix_ixi
λ : Bilangan acak antara 0 dan 1

2.6. Hyperparameter Tuning menggunakan GridSearchCV

adalah proses optimasi untuk menemukan kombinasi hyperparameter terbaik yang menghasilkan kinerja model terbaik. Berikut adalah penjelasan dan rumus terkait proses ini:

A. Hyperparameter dalam Random Forest

Beberapa hyperparameter yang umum dioptimasi dalam Random Forest meliputi:
- ***n_estimators***: Jumlah pohon keputusan dalam forest.
- ***max_depth***: Kedalaman maksimum setiap pohon.
- ***min_samples_split***: Jumlah sampel minimum yang diperlukan untuk membagi sebuah node.
- ***min_samples_leaf***: Jumlah sampel minimum yang harus ada di leaf node.
- ***max_features***: Jumlah fitur yang dipertimbangkan untuk membagi sebuah node.

B. GridSearchCV

GridSearchCV melakukan pencarian exhaustif (menyeluruh) terhadap semua kombinasi hyperparameter yang diberikan. Proses ini melibatkan:
- **Grid Parameter**: Menentukan rentang nilai untuk setiap hyperparameter yang akan diuji.
- **Cross-Validation**: Membagi dataset menjadi beberapa subset (fold) untuk melatih dan menguji model

C. Rumus dan Proses Perhitungan

a. Cross-Validation Score

GridSearchCV menggunakan validasi silang (cross-validation) untuk mengevaluasi setiap kombinasi hyperparameter. Skor validasi silang dihitung sebagai rata-rata skor performa model pada setiap fold. Misalnya, untuk akurasi:

p-ISSN: 2775-4227
e-ISSN: 2775-4219

$$\text{CV score} = \frac{1}{k}\sum_{i=1}^{k}\text{Accuracy}_i \quad (7)$$

Keterangan:

- k: jumlah fold dalam cross-validation.
- Accuracy_i: Akurasi model pada fold ke-i

2.7. Confusion Matrix

Confusion Matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan hasil prediksi model dengan nilai aktual (data sebenarnya). *Confusion matrix* memberikan informasi tentang prediksi benar dan salah yang dilakukan model untuk setiap kelas.

$$[TP\ FP\ FN\ TN\ ] \quad (8)$$

Keterangan:
TP : True Positives.
FP : False Positives.
FN : False Negatives.
TN : True Negatives.

Beberapa perhitungan Confusion matrix terdiri dari sebagai berikut.

a) Akurasi

Mengukur seberapa sering model membuat prediksi yang benar.

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

b) Presisi

Mengukur proporsi prediksi positif yang benar-benar positif.

$$\frac{TP}{TP+FP} \quad (10)$$

c) Recall

Mengukur kemampuan model untuk mendeteksi semua kasus positif yang sebenarnya.

$$\frac{TP}{TP+FN} \quad (11)$$

d) Specifycity

Mengukur kemampuan model untuk mengenali semua kasus negatif yang sebenarnya.

$$\frac{TN}{TN+FP} \quad (12)$$

e) F1-Score

Menggabungkan Precision dan Recall dalam satu metrik dengan mengambil rata-rata harmonisnya.

$$F1 = 2 \times \frac{Presisi \times Recall}{Presisi+Recall} \quad (13)$$

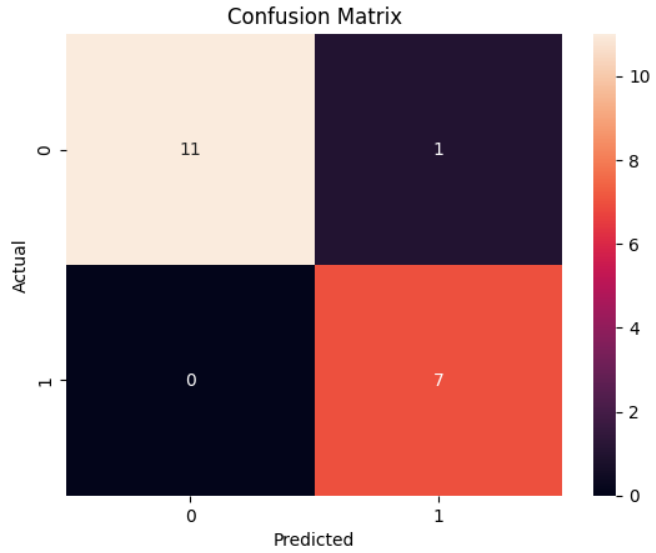
3. HASIL DAN PEMBAHASAN

Dalam studi ini kami menggunakan metode *Random Forest* dengan teknik SMOTEENN dan Hyperparameter Tuning menggunakan GridSearchCV

3.1. Metode Random Forest

Berikut ini optimasi model *Random Forest* untuk prediksi pada dataset penyakit jantung dengan teknik SMOTEENN dan Hyperparameter Tuning menggunakan GridSearchCV. *Confusion Matrix* algoritma *Random Forest* dengan teknik SMOTEENN dan Hyperparameter Tuning menggunakan GridSearchCV disajikan pada gambar 2.

p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 2 *Confusion Matrix Random Forest* dengan teknik SMOTEENN dan Hyperparameter Tuning menggunakan GridSearchCV

Berdasarkan confusion matrix yang ditampilkan, algoritma *Random Forest* yang telah diterapkan dengan teknik SMOTEENN dan hyperparameter tuning menggunakan GridSearchCV menunjukkan kinerja yang sangat baik dalam mendeteksi penyakit jantung. Model ini berhasil mengklasifikasikan 11 sampel negatif dengan benar (True Negative) dan 7 sampel positif dengan benar (True Positive). Hanya terdapat satu kasus False Positive, di mana satu sampel negatif diklasifikasikan sebagai positif, sementara tidak ada False Negative, yang berarti model tidak melewatkan satu pun kasus penyakit jantung.

Dari hasil evaluasi metrik, model ini mencapai akurasi sebesar 0.94%, menunjukkan keandalan yang tinggi dalam memprediksi hasil dengan benar. Presisi untuk kelas positif tercatat sebesar 0.87%, yang berarti sebagian besar prediksi positif adalah benar. Selain itu, recall mencapai 1.0%, menandakan bahwa model tidak melewatkan satupun kasus penyakit jantung dan Specificity mencapai 0.91%, menunjukan bahwa model dalam mengukur dan mengidentifikasi kasus negatif dengan benar. yang sangat krusial dalam bidang kesehatan. F1-score sebesar 0.93% juga menunjukkan keseimbangan yang baik antara presisi dan recall.

Dengan tidak adanya False Negative, model ini sangat cocok untuk aplikasi medis, di mana kesalahan dalam mendeteksi kasus penyakit jantung harus diminimalkan. Teknik SMOTEENN tampaknya berhasil dalam menangani ketidakseimbangan data, sementara GridSearchCV berperan dalam menemukan kombinasi hyperparameter optimal yang meningkatkan performa model. Meskipun demikian, jika ingin lebih meningkatkan presisi dan mengurangi False Positive lebih lanjut, dapat dipertimbangkan teknik threshold tuning atau metode ensemble tambahan. Secara keseluruhan, model ini menunjukkan hasil yang menjanjikan dalam klasifikasi penyakit jantung.

Dalam studi sebelumnya[12] menggunakan 3 metode pembelajaran mesin dalam prediksi penyakit jantung yang teridiri dari *Support Vector Machine* (SVM), *Logistic Regression* dan *Artificial Neural Network* (ANN) dengan hasil pembagian data dan uji 80% : 20% didapatkan hasil seperti pada Tabel 2.

Tabel 2 Hasil Pembagian Data Uji 80% : 20%[12]

Confusion Matrix	Prescision	Recall	F1 Score	Accuracy
SVM	0.67	0.88	0.76	0.70
Logistic Regression	0.88	0.88	0.88	0.86
ANN	0.87	0.84	0.86	0.85

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Berdasarkan pada Tabel 2, metode *Logistic Regression* memiliki nilai presisi tertinggi sebesar 0.88. Baik SVM maupun *Logistic Regression* sama-sama mencapai nilai recall tertinggi sebesar 0.88. Selain itu, *Logistic Regression* juga mencatat nilai F1-score tertinggi sebesar 0.88 dan akurasi tertinggi sebesar 0.86. Hasil pembagian data dengan uji 80 : 20 algoritma *Random Forest* dengan teknik SMOTEENN dan Hyperparameter Tuning menggunakan GridSearchCV didapatkan hasil seperti Tabel 3.

Tabel 3 Hasil Pembagian Data Uji 80% : 20%

Confusion Matrix	Accuracy	Prescision	Recall	*Specificity*	F1 Score
Random Forest	0.94	0.87	1.0	0.91	0.93

Berdasarkan Tabel 3, menunjukkan hasil evaluasi model machine learning menggunakan algoritma *Random Forest* dengan pembagian data uji 80%:20%. Pengujian dilakukan dengan menerapkan teknik SMOTEENN dan Hyperparameter Tuning menggunakan GridSearchCV model ini mencapai nilai akurasi sebesar 0.94, yang berarti model tersebut benar dalam memprediksi 94% dari total data uji. Akurasi ini mengindikasikan bahwa model memiliki kinerja yang sangat baik dalam mengklasifikasikan data. Selain itu, model memiliki presisi sebesar 0.87, yang menunjukkan bahwa 87% dari prediksi positif yang dibuat oleh model adalah benar. Nilai recall yang mencapai 1.0 mengindikasikan bahwa model berhasil mengidentifikasi semua instance positif dalam data, menandakan sensitivitas yang sempurna. Spesifisitas model sebesar 0.91 menunjukkan bahwa model dapat mengidentifikasi 91% dari instance negatif dengan benar. F1 Score, yang merupakan rata-rata harmonik dari presisi dan recall, mencapai 0.93, menandakan keseimbangan yang baik antara kedua metrik tersebut. Secara keseluruhan, model *Random Forest* ini menunjukkan performa yang sangat efektif dalam memprediksi kelas target pada data uji yang diberikan.

3.2. Diskusi

Dalam studi ini, kami membahas implementasi model klasifikasi menggunakan algoritma *Random Forest* untuk memprediksi penyakit jantung berdasarkan dataset "heart.csv". Langkah pertama yang dilakukan adalah memeriksa keberadaan kolom target, yang dalam hal ini adalah kolom 'target', untuk memastikan bahwa dataset yang digunakan sesuai dengan tujuan analisis. Jika kolom tersebut tidak ditemukan, program akan menghentikan eksekusi dan memberikan pesan kesalahan. Setelah memastikan keberadaan kolom target, dataset dibagi menjadi fitur (X) dan target (y). Mengingat dataset yang digunakan mungkin tidak seimbang, kami menerapkan teknik SMOTEENN (*Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors*) untuk menangani ketidakseimbangan data. Teknik ini menggabungkan oversampling pada kelas minoritas dan undersampling pada kelas mayoritas, sehingga menghasilkan distribusi data yang lebih seimbang.

Setelah data diresample, dataset dibagi menjadi data latih dan data uji dengan proporsi 80:20. Selanjutnya, dilakukan penskalaan fitur menggunakan StandardScaler untuk menormalisasi data agar memiliki mean nol dan deviasi standar satu. Hal ini penting untuk memastikan bahwa semua fitur memiliki skala yang sama, sehingga tidak ada fitur yang mendominasi proses pelatihan model. Untuk meningkatkan performa model, kami melakukan tuning hyperparameter menggunakan GridSearchCV dengan berbagai kombinasi parameter seperti jumlah estimator, kedalaman maksimum pohon, dan jumlah sampel minimum yang diperlukan untuk memecah node. Proses ini membantu menemukan kombinasi parameter terbaik yang menghasilkan akurasi tertinggi.

Setelah menemukan model terbaik, kami melakukan prediksi pada data uji dan mengevaluasi performa model menggunakan beberapa metrik evaluasi, termasuk akurasi, presisi, recall, spesifisitas, dan F1-score. Selain itu, kami juga memvisualisasikan confusion matrix untuk melihat seberapa baik model dapat memprediksi kelas positif dan negatif. Hasil evaluasi menunjukkan bahwa model memiliki performa yang baik dengan nilai akurasi, presisi, recall, Specificity, dan F1-score yang tinggi. Terakhir, kami memplot kurva ROC (Receiver Operating Characteristic) untuk menilai kemampuan model dalam membedakan antara kelas positif dan negatif. Area di bawah kurva ROC (AUC) yang mendekati 1 menunjukkan bahwa model memiliki kemampuan klasifikasi yang sangat baik.

Secara keseluruhan, model *Random Forest* yang diimplementasikan dalam artikel ini menunjukkan performa yang baik dalam memprediksi penyakit jantung. Namun, penting untuk dicatat bahwa hasil ini sangat

p-ISSN: 2775-4227
e-ISSN: 2775-4219

bergantung pada kualitas dan representasi data yang digunakan. Oleh karena itu, langkah-langkah seperti penanganan data tidak seimbang, penskalaan fitur, dan tuning hyperparameter menjadi krusial dalam mencapai performa model yang optimal.

4. KESIMPULAN

Studi ini berfokus pada optimasi kinerja algoritma *Random Forest* untuk prediksi penyakit jantung dengan menggunakan teknik SMOTEENN untuk menangani ketidakseimbangan data dan GridSearchCV untuk penyetelan hiperparameter. Hasil studi menunjukkan bahwa kombinasi kedua teknik ini secara signifikan meningkatkan kinerja model dalam mengklasifikasikan penyakit jantung. Model yang dioptimalkan mencapai akurasi sebesar 94% menunjukkan kemampuan model yang sangat baik dalam membedakan antara pasien yang memiliki dan tidak memiliki penyakit jantung. Penggunaan SMOTEENN terbukti efektif dalam menyeimbangkan dataset yang awalnya tidak seimbang, meningkatkan representasi kelas minoritas tanpa menimbulkan noise yang signifikan. Proses optimasi hiperparameter menggunakan GridSearchCV juga berhasil mengidentifikasi kombinasi optimal untuk parameter seperti *n_estimators, max_depth, dan min_samples_leaf*, sehingga model menjadi lebih stabil dan akurat. Perbandingan dengan model baseline menunjukkan peningkatan yang signifikan dalam akurasi, recall, dan F1-score, yang menegaskan pentingnya penanganan ketidakseimbangan data dan penyetelan hiperparameter dalam meningkatkan kinerja model.

Secara keseluruhan, studi ini memberikan kontribusi penting dalam bidang diagnosis medis dengan menunjukkan efektivitas kombinasi SMOTEENN dan GridSearchCV dalam mengoptimalkan *Random Forest* untuk prediksi penyakit jantung. Model yang dihasilkan memiliki potensi untuk digunakan sebagai alat bantu diagnosis dini, yang dapat membantu mengurangi angka kematian dan biaya perawatan. Namun, studi ini memiliki beberapa keterbatasan, seperti ukuran dataset yang relatif kecil dan fokus hanya pada algoritma *Random Forest*. Untuk studi selanjutnya, disarankan untuk mengeksplorasi teknik resampling dan optimasi hiperparameter pada dataset yang lebih besar dan beragam, serta membandingkan performa dengan algoritma lain seperti XGBoost atau Support Vector Machine (SVM).

DAFTAR PUSTAKA

[1] C. W. Tsao *et al.*, “Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association,” *Circulation*, vol. 147, no. 8, Feb. 2023, doi: 10.1161/CIR.0000000000001123.

[2] Donatila Mano S, Agnes Marcella, Yohanes Firmansyah, and Alexander Halim Santoso, “Peningkatan Pemahaman dan Kewaspadaan Masyarakat akan Penyakit Arteri Perifer,” *Jurnal Kabar Masyarakat*, vol. 1, no. 2, pp. 31–40, Jun. 2023, doi: 10.54066/jkb.v1i2.337.

[3] S. A. T. Al Azhima, D. Darmawan, N. F. Arief Hakim, I. Kustiawan, M. Al Qibtiya, and N. S. Syafei, “Hybrid Machine Learning Model untuk memprediksi Penyakit Jantung dengan Metode Logistic Regression dan Random Forest,” *Jurnal Teknologi Terpadu*, vol. 8, no. 1, pp. 40–46, Jul. 2022, doi: 10.54914/jtt.v8i1.539.

[4] A. M. A. Rahim, Inggrid Yanuar Risca Pratiwi, and Muhammad Ainul Fikri, “Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier,” *Indonesian Journal of Computer Science*, vol. 12, no. 5, Nov. 2023, doi: 10.33022/ijcs.v12i5.3413.

[5] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, “Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung,” *Informatik : Jurnal Ilmu Komputer*, vol. 18, no. 3, p. 239, Dec. 2022, doi: 10.52958/iftk.v18i3.4694.

[6] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, “Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 677–690, Jul. 2022, doi: 10.30812/matrik.v21i3.1726.

[7] S. P. Tamba and E. -, “PREDIKSI PENYAKIT GAGAL JANTUNG DENGAN MENGGUNAKAN RANDOM FOREST,” *Jurnal Sistem Informasi dan Ilmu Komputer Prima(JUSIKOM PRIMA)*, vol. 5, no. 2, pp. 176–181, Mar. 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2445.

[8] A. A. G. W. S. Erlangga, I. G. A. Gunadi, and I. M. G. Sunarya, “Kombinasi Oversampling dan Undersampling dalam Menangani Class Imbalanced dan Overlapping pada Klasifikasi Data Bank Marketing,” *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, vol. 7, no. 1, pp. 32–42, Apr. 2024, doi: 10.31598/jurnalresistor.v7i1.1515.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

[9] G. Husain *et al.*, "SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models," *Algorithms*, vol. 18, no. 1, p. 37, Jan. 2025, doi: 10.3390/a18010037.

[10] D. P. Mishra, H. K. Gupta, G. Saajith, and R. Bag, "Optimizing Heart Disease Prediction Model with GridsearchCV for Hyperparameter Tuning," in *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/IC-CGU58078.2024.10530772.

[11] 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," *Multimed Tools Appl*, vol. 81, no. 13, pp. 18155–18179, May 2022, doi: 10.1007/s11042-022-12425-x.

[12] F. Handayani, "Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network Dalam Prediksi Penyakit Jantung," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 7, no. 3, p. 329, Dec. 2021, doi: 10.26418/jp.v7i3.48053.

[13] N. Alotaibi and M. Alzahrani, "Comparative Analysis of Machine Learning Algorithms and Data Mining Techniques for Predicting the Existence of Heart Disease," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022, doi: 10.14569/IJACSA.2022.0130794.

[14] E. Mbunge *et al.*, "Implementation of ensemble machine learning classifiers to predict diarrhoea with SMOTEENN, SMOTE, and SMOTETomek class imbalance approaches," in *2023 Conference on Information Communications Technology and Society (ICTAS)*, IEEE, Mar. 2023, pp. 1–6. doi: 10.1109/ICTAS56421.2023.10082744.

[15] Y. Han and I. Joe, "Enhancing Machine Learning Models Through PCA, SMOTE-ENN, and Stochastic Weighted Averaging," *Applied Sciences*, vol. 14, no. 21, p. 9772, Oct. 2024, doi: 10.3390/app14219772.

[16] Y. A. Sir and A. H. H. Soepranoto, "Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas," *Jurnal Komputer dan Informatika*, vol. 10, no. 1, pp. 31–38, Mar. 2022, doi: 10.35508/jicon.v10i1.6554.

[17] R. Valarmathi and T. Sheela, "Heart disease prediction using hyper parameter optimization (HPO) tuning," *Biomed Signal Process Control*, vol. 70, p. 103033, Sep. 2021, doi: 10.1016/j.bspc.2021.103033.

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

[19] A. Masruriyah, H. Novita, C. Sukmawati, A. Ramadhan, S. Arif, and B. Dermawan, "Pengukuran Kinerja Model Klasifikasi dengan Data Oversampling pada Algoritma Supervised Learning untuk Penyakit Jantung," *Computer Science (CO-SCIENCE)*, vol. 4, no. 1, pp. 62–70, Jan. 2024, doi: 10.31294/coscience.v4i1.2389.

[20] M. Daviran, A. Maghsoudi, R. Ghezelbash, and B. Pradhan, "A new strategy for spatial predictive mapping of mineral prospectivity: Automated hyperparameter tuning of random forest approach," *Comput Geosci*, vol. 148, p. 104688, Mar. 2021, doi: 10.1016/j.cageo.2021.104688.

[21] K.-V. Tompra, G. Papageorgiou, and C. Tjortjis, "Strategic Machine Learning Optimization for Cardiovascular Disease Prediction and High-Risk Patient Identification," *Algorithms*, vol. 17, no. 5, p. 178, Apr. 2024, doi: 10.3390/a17050178.

[22] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *Information*, vol. 13, no. 10, p. 475, Oct. 2022, doi: 10.3390/info13100475.

[23] T. Gori and A. Hestiningtyas, "Optimasi Pemilihan Fitur untuk Prediksi Penyakit Jantung Menggunakan Algoritma Genetika dan Random Forest," *The Indonesian Journal of Computer Science*, vol. 13, no. 5, Oct. 2024, doi: 10.33022/ijcs.v13i5.4214.

[24] N. H. Alfajr and S. Defiyanti, "PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN METODE RANDOM FOREST DAN PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3S1.5055.

[25] S. A. Reddy, S. K. G.A.E., B. M, and L. Mosangi, "Hybrid Machine Learning Approaches for Robust Heart Disease Prediction: A Comprehensive Analysis," in *2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, IEEE, Nov. 2024, pp. 1–10. doi: 10.1109/ICECIE63774.2024.10815655.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

[26] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, no. 4, p. 1210, Apr. 2023, doi: 10.3390/pr11041210.

[27] M. Ahmed, M. H. Sulaiman, M. M. Hassan, and T. Bhuiyan, "Predicting the Classification of Heart Failure Patients Using Optimized Machine Learning Algorithms," *IEEE Access*, vol. 13, pp. 30555–30569, 2025, doi: 10.1109/ACCESS.2025.3541069.

[28] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, and P. Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm," *Inform Med Unlocked*, vol. 41, p. 101316, 2023, doi: 10.1016/j.imu.2023.101316.

[29] 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: 10.3390/a16020088.

[30] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," *Inf Sci (N Y)*, vol. 565, pp. 438–455, Jul. 2021, doi: 10.1016/j.ins.2021.03.041.

[31] M. H. Jamal, N. Naz, M. A. K. Khattak, F. Saeed, S. N. Altamimi, and S. N. Qasem, "A Comparison of Re-Sampling Techniques for Detection of Multi-Step Attacks on Deep Learning Models," *IEEE Access*, vol. 11, pp. 127446–127457, 2023, doi: 10.1109/ACCESS.2023.3332512.