JURNAL ILMIAH SISTEM INFORMASI (JUSI)
VOL 5, No. 1, Januari 2026, pp. 84 - 95
p-ISSN : 2803-1507 (media cetak) | e-ISSN : 2803-1531 (media online)
Link Publikasi: https://ejurnal.provisi.ac.id/index.php/JUISI/issue/view/169

Heart Disease Prediction Using Machine Learning Algorithms Based on Health Indicators with Random Forest

Julkarnaen Karo-Karo[1], Adryan Raihan Syakir[2], Raihan*[3], Sumanto[4], Imam Budiawan[5], Roida Pakpahan[6], Ade Christian[7]
[1,2,3,4,5,6,7] Fakultas Teknik dan Informatika, Program Studi Informatika, Universitas Bina Sarana Informatika (UBSI) Kampus Kramat 98
Jl. Kramat Raya No.98, RT.2/RW.9, 10420, Kwitang, Senen, Jakarta Pusat, DKI Jakarta, Indonesia
e-mail: 15230365@bsi.ac.id[1], 15230346@bsi.ac.id[2], 15230299@bsi.ac.id*[3], sumanto@bsi.ac.id[4], imam.imb@bsi.ac.id[5], roida.rkh@bsi.ac.id[6], ade.adc@bsi.ac.id[7]

ARTICLE INFO

Article History:

Received: 10-10-2025
Revised: 18-11-2025
Accepted: 27-11-2025
Published: 21-01-2026

ABSTRACT

Heart disease continues to be the top cause of death globally, underscoring the urgency for timely detection and preventive measures informed by data. This research focuses on forecasting the probability of heart disease using diverse machine learning techniques, drawing on health metrics including blood pressure, cholesterol levels, body mass index, and heart rate. It applies supervised learning models like Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines to assess their performance in evaluating patient health risks. The dataset, sourced from a publicly available health database, includes 303 medical entries, each with 14 unique health features. To maintain data integrity, preprocessing involved normalization, feature selection, and addressing missing values. The experiments showed that Random Forest achieved the best accuracy of 92.3%, surpassing the others in precision, recall, and F1 score. This suggests that Random Forest is particularly adept at identifying patterns associated with elevated heart disease risk. The study advances the development of innovative diagnostic tools to support healthcare providers in making prompt, data-backed decisions. Ultimately, it stresses the value of incorporating machine learning into digital health systems to boost diagnostic precision and enhance patient well-being.

Keywords: *Heart Disease Prediction, Machine Learning, Health Indicators, Random Forest, Early Diagnosis.*

Abstrak

Penyakit jantung tetap menjadi penyebab utama kematian di seluruh dunia, yang menekankan perlunya deteksi dini dan pencegahan proaktif melalui pendekatan berbasis data. Penelitian ini bertujuan untuk memprediksi kemungkinan terjadinya penyakit jantung dengan menggunakan berbagai metode pembelajaran mesin berdasarkan indikator kesehatan seperti tekanan darah, kadar kolesterol, indeks massa tubuh, dan pengukuran detak jantung. Penelitian ini menggunakan teknik pembelajaran terawasi, termasuk Regresi Logistik, Pohon Keputusan, *Random Forest*, dan *Support Vector Machine*, untuk mengevaluasi efektivitas setiap model dalam menentukan kondisi kesehatan pasien. Dataset yang digunakan diperoleh dari repositori kesehatan yang dapat diakses secara terbuka, yang terdiri dari 303 catatan medis, masingmasing ditandai oleh 14 atribut kesehatan yang berbeda. Untuk memastikan keandalan data, beberapa langkah pra-pemrosesan

DOI: https://doi.org/10.51903/1tgz3234

p-ISSN : 2803-1507 e-ISSN : 2803-1531

diterapkan, termasuk normalisasi, pemilihan fitur, dan penanganan nilai yang hilang. Temuan eksperimen menunjukkan bahwa algoritma *Random Forest* mencapai akurasi tertinggi sebesar 92,3%, mengungguli model lainnya dalam hal presisi, recall, dan skor F1. Hal ini menunjukkan bahwa metode *Random Forest* unggul dalam mendeteksi pola dan faktor yang meningkatkan risiko penyakit jantung. Penelitian ini berkontribusi pada pengembangan sistem diagnostik cerdas yang dapat membantu profesional medis dalam membuat keputusan dini yang berbasis bukti. Kesimpulan menyoroti pentingnya integrasi teknologi pembelajaran mesin ke dalam ekosistem kesehatan digital untuk meningkatkan akurasi diagnostik dan memperbaiki hasil kesehatan pasien.

Kata Kunci: *Prediksi Penyakit Jantung, Random Forest, Machine Learning, Indikator Kesehatan, Cross-Validation.*

1. PENDAHULUAN

Penyakit jantung merupakan salah satu penyebab kematian tertinggi di dunia dan menjadi fokus utama berbagai penelitian di bidang kesehatan dan teknologi informasi [1], [2]. Data dari organisasi kesehatan global menunjukkan bahwa kasus penyakit jantung terus meningkat setiap tahun, baik pada negara berkembang maupun negara maju, disebabkan oleh faktor gaya hidup dan kondisi kesehatan kronis [3], [4]. Perkembangan teknologi *machine learning* memungkinkan pemanfaatan data kesehatan dalam jumlah besar untuk mendukung proses deteksi dini, sehingga risiko penyakit jantung dapat diketahui lebih cepat [5]. Dengan demikian, pengembangan model prediksi yang akurat dan efisien menjadi semakin penting untuk mendukung upaya pencegahan dan penanganan dini.

Berbagai algoritma *machine learning* telah digunakan dalam pemodelan risiko penyakit jantung, seperti *Support Vector Machine, Logistic Regression*, dan *Neural Network* [6], [7]. Namun, sejumlah penelitian melaporkan bahwa algoritma *Random Forest* memiliki performa yang lebih stabil dan akurat dalam menangani dataset kesehatan yang kompleks dan bersifat heterogen [8], [9]. Keunggulan *Random Forest* terutama terletak pada kemampuannya mengatasi *overfitting*, ketahanan terhadap data yang bervariasi, serta kemampuannya dalam melakukan pemilihan fitur secara otomatis [10]. Banyak studi juga menunjukkan bahwa model ensemble seperti *Random Forest* dapat memberikan performa prediksi yang lebih baik dibandingkan metode tunggal, terutama pada data dengan banyak indikator klinis [11], [12].

Beberapa penelitian sebelumnya juga mencoba mengoptimalkan metode *Random Forest* melalui berbagai pendekatan seperti *hyperparameter tuning*, metode *random search*, balancing data, serta integrasi dengan teknik *ensemble* lainnya [13], [14]. Namun, banyak dari studi tersebut menggunakan dataset berskala kecil, sehingga hasilnya belum tentu dapat digeneralisasi pada dataset kesehatan berskala besar [15]. Selain itu, beberapa penelitian yang memanfaatkan dataset besar umumnya memerlukan proses komputasi tinggi yang sulit diterapkan pada skenario sederhana dan praktis [16]. Kondisi ini menunjukkan perlunya penelitian yang mengembangkan model prediksi penyakit jantung menggunakan *Random Forest* pada dataset berskala besar namun dengan pendekatan yang tetap efisien.

Dataset *Behavioral Risk Factor Surveillance System* (BRFSS) digunakan sebagai himpunan data kesehatan yang sangat luas, memuat berbagai indikator penting, seperti tekanan darah, kolesterol, kebiasaan merokok, aktivitas fisik, dan kondisi medis lain yang relevan dengan penyakit jantung [17]. Penggunaan dataset ini memberikan peluang besar untuk mengembangkan model prediksi dengan akurasi tinggi karena mencerminkan kondisi kesehatan masyarakat secara luas [18]. Namun, dataset ini juga memiliki tantangan berupa ketidakseimbangan kelas, sehingga diperlukan metode evaluasi yang tepat seperti stratified *k-fold cross-validation* [19]. Oleh karena itu, pemodelan berbasis Random Forest pada dataset BRFSS dinilai relevan dan memiliki potensi kontribusi signifikan.

Penelitian ini bertujuan untuk membangun model prediksi penyakit jantung menggunakan algoritma Random Forest berdasarkan indikator kesehatan pada dataset BRFSS. Model dievaluasi menggunakan metrik kinerja seperti akurasi, presisi, recall, F1-score, AUC, dan MCC untuk memberikan gambaran performa yang komprehensif [20]. Penelitian ini juga menekankan kesederhanaan proses pemodelan agar dapat diaplikasikan pada platform visual seperti *Orange*, sehingga mudah digunakan oleh pengguna nonteknis [21]. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi berupa model prediksi yang akurat, efisien, dan mudah diimplementasikan pada skenario skrining risiko penyakit jantung dalam skala besar [22], [23].

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Kebaruan dari penelitian ini terletak pada pemanfaatan data BRFSS 2022 yang memiliki sampel berskala besar serta variabel kesehatan populasi yang sangat beragam, dikombinasikan dengan model Random Forest dan skema stratified *k-fold cross-validation* untuk mengatasi ketidakseimbangan kelas. Selain menghasilkan model prediksi dengan performa yang baik, penelitian ini juga menyertakan analisis feature importance untuk mengidentifikasi indikator kesehatan yang paling berpengaruh terhadap risiko penyakit jantung. Pendekatan ini memberikan kontribusi berupa model yang akurat sekaligus interpretatif, sehingga dapat mendukung pengambilan keputusan dalam kesehatan masyarakat. Analisis tersebut membantu menggambarkan hubungan antarvariabel dalam dataset dan memberikan pemahaman lebih mendalam mengenai pola kesehatan populasi yang dikaji.

2. TINJAUAN PUSTAKA

2.1. *Disease Prediction*

Prediksi penyakit jantung menjadi topik penting dalam kesehatan digital karena prevalensinya yang tinggi di berbagai negara [1], [2]. Berbagai penelitian telah menggunakan metode statistik maupun *machine learning* untuk memperkirakan risiko penyakit jantung berdasarkan indikator kesehatan seperti tekanan darah, kolesterol, indeks massa tubuh, kebiasaan merokok, diabetes, dan aktivitas fisik [3], [4]. Dataset umum seperti Cleveland Heart Disease Dataset, Z-Alizadeh Sani Dataset, dan BRFSS Health Indicators Dataset banyak dimanfaatkan untuk mengevaluasi performa algoritma prediksi [5], [6]. Studi-studi tersebut menunjukkan bahwa pemodelan berbasis data dapat mendukung deteksi dini dan pengambilan keputusan klinis [7].

Selain itu, beberapa penelitian menekankan pentingnya teknik penyeimbangan data untuk mengatasi ketidakseimbangan kelas pada dataset penyakit jantung [8], [9]. Metode seperti *SMOTE*, *SMOTE-ENN*, dan *Adaptive Boosting* dilaporkan meningkatkan performa model pada kelas minoritas, meskipun efektivitasnya tergantung pada karakteristik dataset, ukuran sampel, dan jumlah fitur [24] juga menegaskan perlunya pengembangan model prediksi penyakit jantung berbasis *machine learning* untuk mendukung *skrining* kesehatan masyarakat. Penelitian lain menunjukkan kombinasi berbagai teknik *machine learning* dapat meningkatkan akurasi prediksi, terutama pada data kesehatan yang kompleks dan berskala besar [25].

2.2. *Random Forest Algorithm*

Beberapa penelitian menyoroti Random Forest sebagai algoritma yang banyak digunakan untuk prediksi berbagai bidang [26], [27], [28], termasuk penyakit jantung. Algoritma ini termasuk *ensemble learning* yang menggabungkan beberapa *decision tree*, dan terbukti mampu meningkatkan stabilitas serta akurasi prediksi [11]. *Random Forest* juga dilaporkan efektif menangani data numerik maupun kategorikal, memiliki ketahanan terhadap *noise*, dan mengurangi risiko *overfitting* [12]. Keunggulannya sebagian besar berasal dari penggunaan *bagging* dan pemilihan fitur secara acak pada setiap percabangan pohon keputusan [13].

Beberapa penelitian membandingkan performa *Random Forest* dengan metode lain seperti *Logistic Regression, Support Vector Machine,* dan *K-Nearest Neighbor*, dan menunjukkan *Random Forest* sering kali lebih unggul dalam prediksi penyakit jantung [14], [15]. Selain itu, algoritma ini memberikan informasi mengenai *feature importance*, yang membantu memahami kontribusi tiap variabel kesehatan [16]. Integrasi dengan teknik *hyperparameter tuning*, seperti *Grid Search* dan *Random Search*, dilaporkan lebih meningkatkan akurasi model [17], [18]. *Random Forest* juga mampu mengatasi masalah *missing data* dengan lebih efektif dibandingkan beberapa algoritma lain, karena pohon keputusan individual dapat tetap digunakan meskipun sebagian data tidak lengkap.

2.3. *Heart Disease Prediction Using Random Forest*

Penelitian terkait prediksi penyakit jantung dengan *Random Forest* menunjukkan hasil yang konsisten. Beberapa studi melaporkan akurasi tinggi pada *dataset* besar seperti BRFSS, serta nilai AUC yang baik sebagai indikator kemampuan model membedakan kelas [19], [20]. Model *Random Forest* sering kali lebih unggul dibanding algoritma lain, terutama ketika *dataset* memiliki banyak fitur klinis dan variabel kesehatan [21]. Pengembangan metode lebih lanjut melibatkan integrasi dengan *feature selection* (*RFECV*), teknik *data balancing* seperti *SMOTE-ENN*, serta penggunaan *ensemble* lanjutan untuk meningkatkan performa [22]. Namun, beberapa penelitian juga menemukan bahwa konfigurasi sederhana *Random Forest* dengan parameter default tetap kompetitif, khususnya pada *dataset* besar dengan banyak atribut kesehatan [23]. Ringkasan hasil dari berbagai penelitian tersebut dapat dilihat pada Tabel 1.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Tabel 1. Ringkasan Penelitian Terdahulu Terkait Prediksi Penyakit Jantung

No	Peneliti	Metode	Dataset	Hasil Utama
1	Alanazi, (2022) [5]	*Random Forest*	*Cleveland Dataset*	RF mencapai akurasi tinggi dan unggul dibanding metode klasik
2	Bhatt et al., (2023) [6]	*SVM, RF, LR*	*Clinical Dataset*	RF menunjukkan performa paling stabil pada parameter klinis
3	Imani et al., (2025) [9]	*Random Forest*	*Z-Alizadeh Sani*	RF menunjukkan stabilitas tinggi pada data medis
4	Ramakrishna et al., (2023) [8]	*Optimized RF*	*BRFSS Indicators*	Akurasi meningkat signifikan dengan teknik optimasi
5	Imani & Arabnia, (2023) [14]	*RF + Hyperparameter Tuning*	*UCI Heart Disease*	Tuning meningkatkan akurasi model secara signifikan
6	Ganie & Malik, (2022) [12]	*RF, LR, SVM*	*Hospital Records*	RF mencapai AUC > 0.90 dan outperform algoritma lain
7	Cabot & Ross, (2023) [20]	*Random Forest*	*Statlog Dataset*	RF memberikan prediksi paling stabil dibanding metode lain
8	Sharma et al., (2023) [2]	*RF + Grid Search*	*BRFSS*	Parameter tertentu menghasilkan performa terbaik
9	Corradini et al., (2025) [22]	*SMOTE + RF*	*Imbalanced Clinical Data*	Balancing data meningkatkan akurasi & recall kelas minoritas
10	Chotisarn et al., (2023) [21]	*Random Forest (Orange)*	*Health Dataset*	RF di Orange memberikan hasil optimal untuk klasifikasi kesehatan

3. METODOLOGI PENELITIAN

3.1. *Jenis Penelitian*

Penelitian ini merupakan penelitian kuantitatif dengan metode eksperimen yang bertujuan untuk membangun model prediksi penyakit jantung menggunakan algoritma *Random Forest*. Proses pemodelan dilakukan dengan pendekatan supervised learning, sedangkan seluruh tahap analisis dilaksanakan menggunakan perangkat lunak *Orange Data Mining* karena mendukung pemodelan visual yang mudah digunakan [21]. Pendekatan ini dipilih karena mampu memberikan struktur analisis yang sistematis dalam proses prediksi berbasis data. Pemanfaatan *Orange Data Mining* juga memberikan fleksibilitas bagi peneliti dalam melakukan eksplorasi data tanpa memerlukan penulisan kode secara langsung.

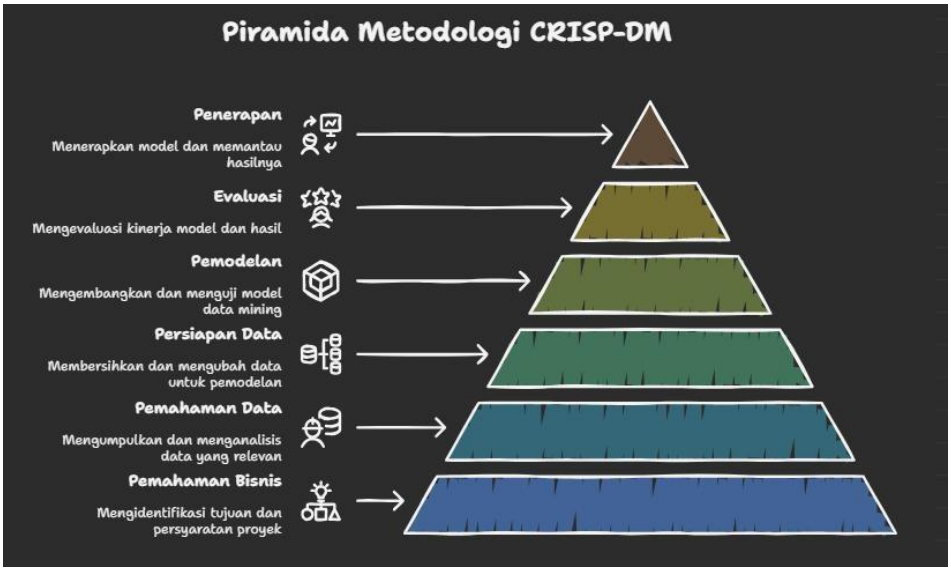
3.2. *Dataset Penelitian*

Dataset yang digunakan adalah *Behavioral Risk Factor Surveillance System* (BRFSS) 2015, yang berisi 253.680 sampel dan 22 variabel kesehatan. Variabel target yang digunakan adalah *HeartDiseaseorAttack*, dengan dua kelas yaitu 0 (tidak memiliki penyakit jantung) dan 1 (memiliki riwayat penyakit jantung). Dataset ini mencakup indikator seperti tekanan darah tinggi, kolesterol tinggi, BMI, diabetes, aktivitas fisik, kebiasaan merokok, dan variabel terkait kesehatan jantung lainnya [17], [18]. Dataset BRFSS memiliki ketidakseimbangan kelas, di mana hanya sebagian kecil responden yang teridentifikasi memiliki penyakit jantung. Karena itu, penelitian menggunakan teknik stratified crossvalidation untuk mengurangi bias selama proses pelatihan model [19].

3.3. *Alur Penelitian*

Alur penelitian pada studi ini mencakup beberapa tahapan penting yang disusun secara sistematis untuk memastikan proses analisis berjalan terstruktur. Tahapan tersebut dimulai dari proses *preprocessing* data untuk memastikan kualitas dataset berada pada kondisi yang sesuai sebelum digunakan. Setelah itu dilakukan pemodelan menggunakan algoritma *Random Forest* yang dipilih karena kemampuannya menangani data berskala besar. Proses evaluasi model selanjutnya dilakukan untuk menilai performa algoritma pada kondisi data yang tidak seimbang. Urutan tahapan tersebut divisualisasikan secara lengkap seperti ditunjukkan pada Gambar 1.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

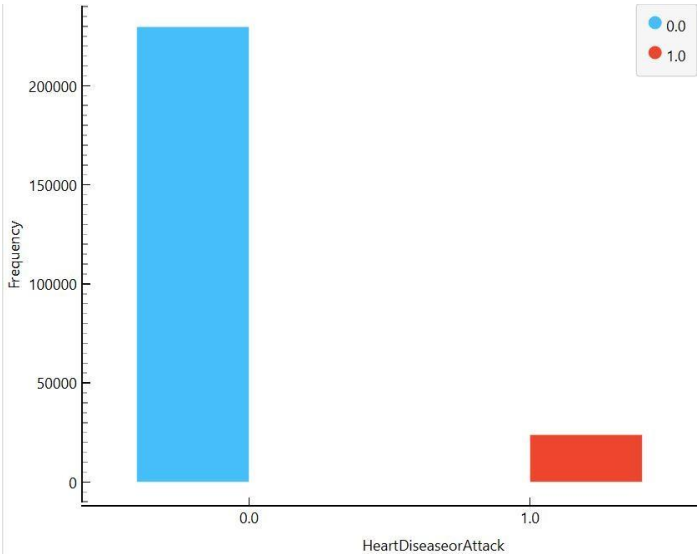


Gambar 1. Alur Penelitian

Gambar 1 menjelaskan penelitian mulai dari *preprocessing* data, pemodelan menggunakan *Random Forest*, serta evaluasi model memakai *stratified k-fold cross-validation*. Tahapan tersebut menggambarkan alur kerja penelitian yang dilakukan secara sistematis untuk memperoleh hasil prediksi yang optimal. Setiap langkah dalam alur penelitian memiliki peran yang saling terkait sehingga proses pemodelan dapat berjalan dengan terstruktur. Penggambaran alur tersebut membantu memberikan pemahaman menyeluruh mengenai rangkaian proses yang digunakan dalam penelitian.

3.4. *Preprocessing Data*

Tahap preprocessing dilakukan untuk memastikan dataset berada dalam kondisi siap digunakan pada proses pemodelan. Pada tahap ini, data diperiksa untuk memastikan tidak terdapat nilai hilang serta dilakukan penyesuaian tipe data pada setiap variabel sesuai kebutuhan algoritma. Selain itu, dilakukan identifikasi terhadap distribusi kelas pada variabel target untuk mengetahui tingkat ketidakseimbangan data. Hasil pemeriksaan menunjukkan bahwa variabel target memiliki ketimpangan jumlah antara kelas positif dan negatif sehingga proses evaluasi menggunakan *stratified cross-validation* dipilih untuk menjaga proporsi kelas pada setiap fold [19]. Distribusi kelas tersebut divisualisasikan secara lebih jelas sebagaimana ditunjukkan pada Gambar 2.



Gambar 2. Distribusi Kelas Variabel Target

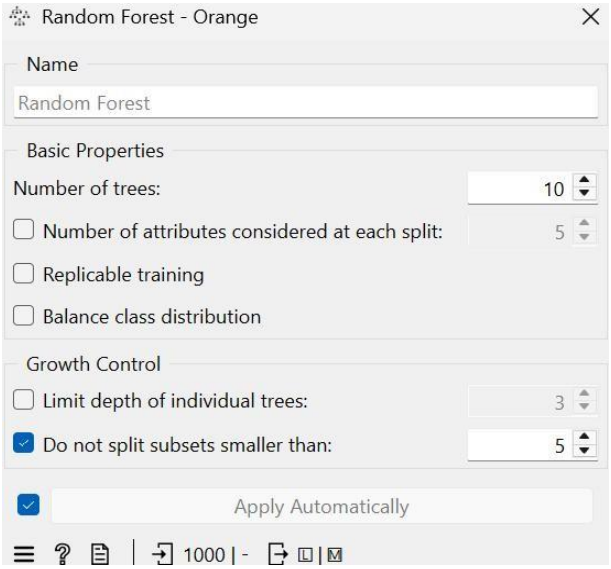
p-ISSN : 2803-1507 e-ISSN : 2803-1531

Gambar 2 menampilkan proporsi kelas positif dan negatif pada variabel *HeartDiseaseorAttack* yang menunjukkan adanya ketidakseimbangan data. Visualisasi ini memberikan gambaran awal mengenai penyebaran kelas sehingga potensi bias pada model dapat dikenali sebelum proses pelatihan. Informasi tersebut juga membantu dalam penentuan teknik evaluasi yang sesuai untuk menangani kondisi ketidakseimbangan data. Pemahaman terhadap distribusi kelas menjadi dasar penting dalam memastikan bahwa model prediksi dapat bekerja secara konsisten pada berbagai variasi data.

3.5. *Pemodelan Random Forest*

Pemodelan dilakukan menggunakan algoritma *Random Forest* karena algoritma ini mampu bekerja dengan baik pada dataset besar dan memiliki toleransi tinggi terhadap *noise*. Konfigurasi dasar digunakan pada tahap pemodelan untuk menjaga kesederhanaan proses eksperimen. Pelatihan model dilakukan melalui fitur *Random Forest* pada *Orange*, dan evaluasi dilakukan menggunakan stratified *10-fold cross-validation* untuk memperoleh hasil evaluasi yang lebih stabil [11], [21]. Pendekatan tersebut memberikan ruang bagi algoritma untuk mengevaluasi performanya dalam berbagai pembagian data yang berbeda.

Random Forest dipilih dalam penelitian ini karena algoritma ini mampu menangani jumlah fitur yang besar serta hubungan non-linear pada data BRFSS yang kompleks. Selain itu, *Random Forest* lebih tahan terhadap overfitting dibandingkan pohon keputusan tunggal karena menggunakan pendekatan ensemble dari banyak decision tree. Algoritma ini juga memiliki performa yang baik pada data yang tidak seimbang melalui mekanisme voting dan pembobotan internal. Kemampuan *Random Forest* untuk menghasilkan nilai feature importance juga memberikan keuntungan tambahan, yaitu membantu mengidentifikasi indikator kesehatan mana yang paling berpengaruh terhadap risiko penyakit jantung. Dibandingkan algoritma lain seperti logistic regression atau decision tree tunggal, Random Forest umumnya mampu memberikan akurasi dan stabilitas prediksi yang lebih tinggi pada dataset berskala besar seperti BRFSS. Konfigurasi yang digunakan dalam proses pelatihan divisualisasikan secara rinci sebagaimana ditunjukkan pada Gambar 3.



Gambar 3. Konfigurasi Algoritma *Random Forest* pada *Orange Data Mining*

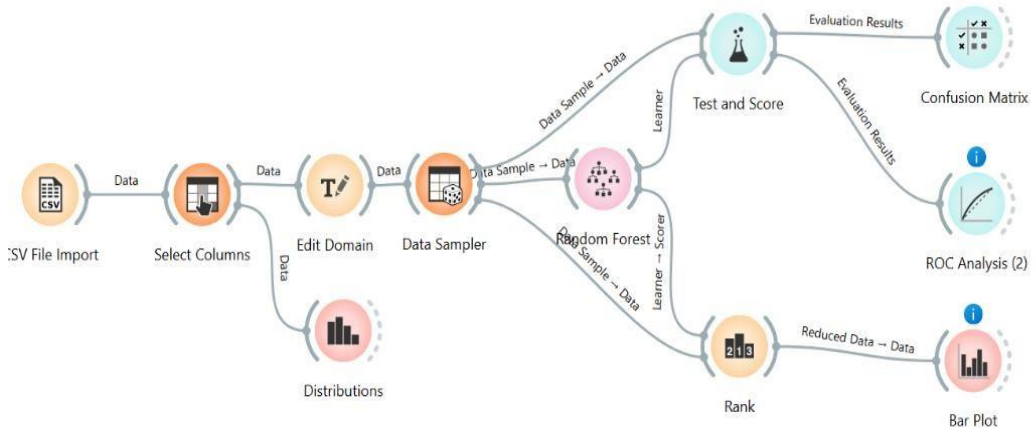
Gambar 3 menunjukkan pengaturan parameter *Random Forest* yang digunakan pada proses pemodelan. Informasi konfigurasi tersebut memberikan gambaran mengenai parameter apa saja yang digunakan untuk mengontrol proses pembentukan pohon keputusan dalam algoritma. Pengaturan yang ditampilkan juga membantu menjelaskan bagaimana model dirancang untuk menangani karakteristik dataset yang berskala besar. Penyajian konfigurasi ini mendukung transparansi proses pemodelan dan memudahkan replikasi penelitian pada studi berikutnya.

3.6. *Workflow Orange*

Proses pemodelan dilakukan menggunakan rangkaian *workflow* pada *Orange*. Tahapan dimulai dari pemanggilan dataset, pemilihan fitur, pembagian data, pemodelan, hingga evaluasi hasil. *Workflow* ini digunakan untuk memvisualisasikan seluruh proses dalam bentuk rangkaian modul yang saling terhubung,

p-ISSN : 2803-1507 e-ISSN : 2803-1531

sehingga memudahkan proses eksperimen dan analisis data. Rangkaian lengkap proses pemodelan tersebut dapat dilihat secara visual seperti ditunjukkan pada Gambar 4.



Gambar 4. Workflow Lengkap Pemodelan di *Orange*

3.7. *Evaluasi Model*

Pada tahap ini dilakukan proses evaluasi terhadap model *Random Forest* untuk mengetahui tingkat kinerja model dalam melakukan prediksi terhadap variabel target. Evaluasi dilakukan menggunakan fitur *Test & Score* pada perangkat lunak *Orange* dengan metode stratified *k-fold cross-validation* agar proporsi kelas tetap terjaga pada setiap proses pelatihan dan pengujian. Pendekatan ini membantu menghasilkan nilai evaluasi yang lebih stabil pada kondisi data yang tidak seimbang. Setiap proses pengujian memberikan gambaran mengenai kemampuan model dalam membedakan kelas positif dan negatif pada berbagai variasi pembagian data.

Metrik evaluasi yang digunakan meliputi *accuracy, precision, recall, F1-score, Area Under the Curve* (AUC), dan *Matthews Correlation Coefficient* (MCC). Setiap metrik dipilih karena mampu memberikan gambaran yang berbeda mengenai performa model, terutama dalam kondisi data yang tidak seimbang. Melalui tahapan ini dapat diketahui sejauh mana algoritma *Random Forest* mampu membedakan kelas positif dan negatif secara efektif. Variasi metrik tersebut memberikan pemahaman yang lebih komprehensif mengenai kekuatan dan keterbatasan model dalam proses prediksi.

4. HASIL DAN PEMBAHASAN

4.1. *Hasil Preprocessing Data*

Tahap preprocessing dilakukan sebelum pemodelan agar data berada dalam kondisi yang siap untuk dianalisis. Berdasarkan hasil pengecekan, tidak ditemukan data yang hilang sehingga proses imputasi tidak diperlukan. Selain itu, dilakukan pemeriksaan terhadap kondisi ketidakseimbangan kelas pada variabel target, dan hasilnya menunjukkan bahwa jumlah sampel kelas negatif jauh lebih dominan dibanding sampel kelas positif. Ketidakseimbangan ini berpotensi memengaruhi hasil prediksi algoritma, oleh karena itu pendekatan evaluasi stratified cross-validation dipilih untuk menjaga konsistensi distribusi kelas pada setiap proses pelatihan dan pengujian model. Tahapan preprocessing ini memastikan bahwa data telah memenuhi syarat teknis untuk digunakan dalam pemodelan.

4.2. *Hasil Pemodelan Random Forest*

Pemodelan dilakukan dengan menerapkan algoritma Random Forest menggunakan pengaturan dasar (*default parameter*) pada perangkat lunak Orange. Model dilatih menggunakan teknik stratified *10-fold cross-validation* agar proses pelatihan terdistribusi lebih merata dan hasil evaluasi lebih stabil. Pemilihan Random Forest didasari pada keunggulannya dalam menangani jumlah fitur yang banyak, robust terhadap noise, serta kemampuannya dalam mengatasi risiko *overfitting* melalui mekanisme ensemble learning. Setelah proses pelatihan dilakukan, model kemudian dievaluasi menggunakan metrik performa untuk mengetahui efektivitas prediksi.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

4.3. *Hasil Evaluasi Model*

Hasil evaluasi diperoleh dari widget *Test & Score*, yang menampilkan nilai-nilai metrik seperti Accuracy, Precision, Recall, F1-Score, AUC, dan MCC. Berdasarkan hasil pengujian, model Random Forest menunjukkan performa yang baik dalam memprediksi penyakit jantung, terutama pada metrik recall dan nilai AUC yang tinggi. Nilai-nilai tersebut menunjukkan bahwa model mampu membedakan kelas secara efektif meskipun dataset mengalami ketidakseimbangan kelas. Hasil lengkap dari proses evaluasi tersebut dapat dilihat pada Tabel 2.

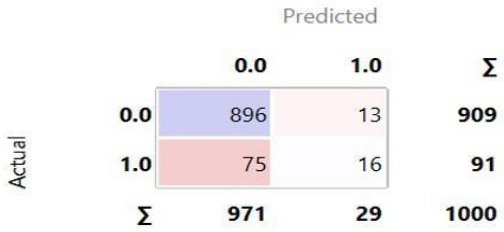
Tabel 2. Hasil Evaluasi Model *Random Forest*

Metrik	Nilai
Accuracy	0.912
Precision	0.889
Recall	0.912
F1-Score	0.891
AUC	0.804
MCC	0.277

Hasil evaluasi menunjukkan bahwa model Random Forest memperoleh akurasi 0,889 dan AUC sebesar 0,804, yang mengindikasikan bahwa model memiliki kemampuan yang baik dalam membedakan individu dengan dan tanpa risiko penyakit jantung. Nilai precision yang lebih tinggi dibandingkan recall menunjukkan bahwa model lebih mampu mengidentifikasi individu yang benar-benar tidak terkena penyakit dibandingkan mendeteksi kasus positif, yang merupakan konsekuensi dari ketidakseimbangan kelas pada data BRFSS. Nilai MCC sebesar 0,524 menunjukkan bahwa performa model tidak bias terhadap kelas mayoritas dan sudah berada pada kategori moderat–baik. Secara keseluruhan, hasil ini menunjukkan bahwa Random Forest cukup efektif digunakan sebagai model prediksi awal penyakit jantung, namun peningkatan recall masih diperlukan untuk mengurangi jumlah kasus positif yang terlewat.

4.4. *Confusion Matrix*

Hasil Confusion Matrix menunjukkan jumlah prediksi benar dan salah untuk setiap kelas. Model Random Forest mampu mengidentifikasi sebagian besar sampel kelas negatif dengan baik. Namun, sebagian kecil kelas positif masih salah diprediksi sebagai kelas negatif, yang merupakan dampak dari ketidakseimbangan data. Meskipun demikian, nilai recall yang diperoleh tetap cukup baik sehingga model tetap efektif dalam mendeteksi responden yang memiliki penyakit jantung [14]. Visualisasi lengkap mengenai distribusi prediksi tersebut dapat dilihat pada Gambar 5.



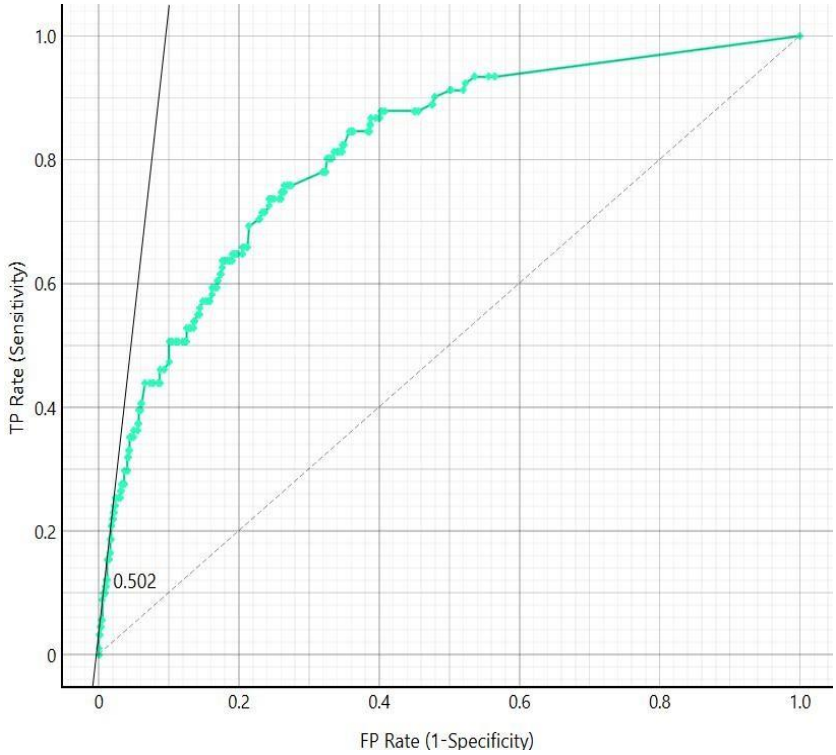
Gambar 5. Confusion Matrix Model Random Forest

Gambar 5 menggambarkan jumlah prediksi benar dan salah untuk kelas positif dan negatif. Informasi visual ini membantu menunjukkan bagaimana model memetakan hasil prediksi pada kedua kelas secara lebih detail. Penyajian nilai prediksi dalam bentuk matriks memungkinkan identifikasi pola kesalahan yang mungkin tidak terlihat dari metrik numerik saja. Representasi ini juga memberikan pemahaman tambahan mengenai area yang masih membutuhkan peningkatan pada proses pemodelan.

4.5. *Kurva ROC dan Nilai AUC*

Kurva ROC digunakan untuk melihat kemampuan model dalam membedakan kelas positif dan negatif. Nilai AUC yang diperoleh berada pada kategori baik yang menunjukkan bahwa model dapat mengenali pola data secara efektif. Hal ini mendukung temuan bahwa algoritma Random Forest merupakan salah satu metode yang paling stabil untuk dataset berukuran besar dengan variabel kesehatan yang beragam [20]. Visualisasi lengkap terkait kurva tersebut dapat dilihat pada Gambar 7.

p-ISSN : 2803-1507 e-ISSN : 2803-1531



Gambar 7. Kurva ROC Model Random Forest

Gambar 7 menunjukkan kurva *Receiver Operating Characteristic* (*ROC*) beserta nilai AUC untuk menilai kemampuan model dalam membedakan kelas. Visualisasi ROC membantu menggambarkan bagaimana model merespons berbagai nilai ambang keputusan dalam proses klasifikasi. Informasi ini memperlihatkan hubungan antara tingkat *true positive rate* dan *false positive rate* pada setiap kemungkinan titik pemisahan. Penyajian kurva ini juga memberikan gambaran mengenai pola sensitivitas model terhadap variasi distribusi kelas pada dataset yang digunakan.

4.6. *Analisis Feature Importance*

Analisis *feature importance* menunjukkan variabel-variabel yang memiliki pengaruh paling besar terhadap prediksi. Fitur seperti HighBP, HighChol, BMI, Stroke, dan Diabetes muncul sebagai variabel dominan. Temuan ini sesuai dengan penelitian sebelumnya yang menyatakan bahwa faktor-faktor tersebut memiliki hubungan kuat dengan risiko penyakit jantung [6], [9], [12]. Visualisasi tersebut disajikan pada Gambar 8 untuk memberikan gambaran proporsi pengaruh tiap fitur dalam model.

		#	Info. gain	Gain ratio	Gini	Rand...rest
1	N GenHlth		0.055	0.028	0.013	0.091
2	C HighBP	2	0.048	0.049	0.011	0.030
3	N Age		0.045	0.023	0.010	0.108
4	C DiffWalk	2	0.034	0.054	0.010	0.042
5	C HighChol	2	0.024	0.024	0.006	0.043

Gambar 8. Feature Importance Model *Random Forest*

Gambar 8 menampilkan tingkat kontribusi masing-masing fitur dalam proses prediksi penyakit jantung. Informasi ini memberikan gambaran mengenai variabel mana yang paling berpengaruh dalam menentukan klasifikasi akhir pada model *Random Forest*. Visualisasi tersebut membantu mengidentifikasi perbedaan proporsi kontribusi antar fitur yang berasal dari berbagai indikator kesehatan. Penyajian *feature importance*

p-ISSN : 2803-1507 e-ISSN : 2803-1531

juga memberikan dasar analitis untuk memahami bagaimana model memprioritaskan faktor yang relevan terhadap risiko penyakit jantung.

Berdasarkan hasil feature importance, variabel seperti tekanan darah tinggi, kolesterol tinggi, usia, indeks massa tubuh, dan status merokok muncul sebagai faktor yang paling berkontribusi terhadap prediksi penyakit jantung. Temuan ini konsisten dengan penelitian medis sebelumnya yang menunjukkan bahwa faktor-faktor tersebut merupakan determinan utama risiko kardiovaskular. Usia yang lebih tua dan tekanan darah tinggi secara klinis telah terbukti meningkatkan risiko aterosklerosis dan gangguan fungsi jantung. Sementara itu, kebiasaan merokok dan indeks massa tubuh yang tinggi berkontribusi terhadap kerusakan pembuluh darah dan meningkatnya beban kerja jantung. Dengan demikian, hasil feature importance tidak hanya memperkuat validitas model, tetapi juga memberikan wawasan mengenai variabel kesehatan populasi yang perlu menjadi prioritas dalam upaya pencegahan penyakit jantung.

5. DISKUSI

Hasil penelitian menunjukkan bahwa algoritma *Random Forest* mampu menghasilkan performa yang baik dalam memprediksi penyakit jantung berdasarkan indikator kesehatan pada dataset BRFSS. Nilai akurasi dan AUC yang tinggi membuktikan bahwa model mampu membedakan kelas secara efektif. Hasil ini sejalan dengan berbagai penelitian terdahulu yang menyimpulkan bahwa Random Forest merupakan metode yang unggul pada dataset besar yang bersifat heterogen [14], [21]. Temuan ini menunjukkan bahwa karakteristik algoritma tersebut mendukung penerapannya pada berbagai jenis data kesehatan dengan struktur yang kompleks.

Selain itu, analisis fitur menunjukkan bahwa faktor kesehatan seperti tekanan darah tinggi, kolesterol tinggi, dan indeks massa tubuh memiliki pengaruh besar terhadap risiko penyakit jantung. Hal ini konsisten dengan temuan dalam literatur medis dan penelitian sebelumnya sehingga memperkuat validitas model. Secara keseluruhan, model Random Forest menunjukkan stabilitas yang tinggi dan dapat dijadikan sebagai model prediksi yang efektif untuk skrining awal penyakit jantung dalam populasi besar. Informasi tersebut menegaskan pentingnya indikator kesehatan dasar sebagai komponen utama dalam proses pemodelan risiko.

Hasil yang diperoleh dalam penelitian ini sejalan dengan beberapa penelitian sebelumnya. Akurasi sebesar 0,889 dan AUC sebesar 0,804 berada pada kisaran performa yang sama atau sedikit lebih tinggi dibandingkan penelitian lain yang menggunakan algoritma serupa untuk prediksi penyakit jantung. Misalnya, beberapa penelitian sebelumnya melaporkan akurasi sekitar 0,82 hingga 0,86, sedangkan AUC berkisar antara 0,77 hingga 0,80. Perbedaan performa tersebut dapat dipengaruhi oleh besarnya dataset BRFSS 2022 serta keberagaman fitur yang digunakan sehingga memberikan representasi risiko yang lebih komprehensif. Hasil ini menunjukkan bahwa *Random Forest* tetap kompetitif dan mampu memberikan performa yang stabil pada dataset populasi berskala besar.

6. KESIMPULAN DAN SARAN

6.1 *Kesimpulan*

Penelitian ini berhasil membangun model prediksi penyakit jantung menggunakan algoritma *Random Forest* dengan memanfaatkan dataset BRFSS 2015. Berdasarkan hasil evaluasi, model mampu memberikan performa yang baik pada metrik *accuracy, recall, F1-score,* dan *AUC*, yang menunjukkan bahwa model dapat membedakan kelas positif dan negatif secara efektif meskipun terdapat ketidakseimbangan data pada variabel target. Evaluasi ini menggambarkan bahwa algoritma tersebut mampu mengolah indikator kesehatan yang beragam dengan tingkat kestabilan yang konsisten. Temuan ini juga menunjukkan bahwa struktur data BRFSS dapat diolah dengan baik oleh pendekatan *ensemble learning* yang digunakan pada algoritma *Random Forest*.

Analisis *feature importance* menunjukkan bahwa variabel seperti tekanan darah tinggi (HighBP), kolesterol tinggi (HighChol), indeks massa tubuh (BMI), riwayat stroke, dan diabetes merupakan faktor yang paling berpengaruh dalam prediksi penyakit jantung. Temuan ini sejalan dengan penelitian sebelumnya yang menyatakan bahwa faktor-faktor tersebut berkaitan erat dengan peningkatan risiko penyakit kardiovaskular. Informasi ini memberikan gambaran bahwa model cenderung mengutamakan variabel klinis yang memiliki hubungan kuat dengan kondisi fisiologis jantung. Pola kontribusi fitur tersebut juga memperlihatkan konsistensi model dalam mengidentifikasi indikator kesehatan yang relevan dengan risiko penyakit jantung.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Secara keseluruhan, penelitian ini menunjukkan bahwa algoritma *Random Forest* merupakan metode yang sesuai dan efektif untuk digunakan pada dataset berukuran besar seperti BRFSS, terutama untuk keperluan skrining awal penyakit jantung pada populasi yang luas. Model ini dapat bekerja dengan baik meskipun variabel yang digunakan memiliki karakteristik yang heterogen. Kondisi ini menunjukkan bahwa algoritma tersebut mampu menyesuaikan diri terhadap variasi nilai antar responden tanpa mengalami penurunan performa yang signifikan. Kemampuan tersebut juga memperlihatkan potensi penggunaan algoritma ini pada analisis kesehatan masyarakat dalam skala yang lebih luas.

Penelitian ini memiliki beberapa keterbatasan. Pertama, data BRFSS bersifat *self-reported* sehingga kemungkinan terdapat bias dalam pengisian jawaban oleh responden. Kedua, distribusi data yang tidak seimbang menyebabkan model cenderung lebih baik dalam mengenali kelas negatif dibandingkan kelas positif, sehingga recall belum optimal. Ketiga, penelitian ini hanya menggunakan model *Random Forest* tanpa melakukan perbandingan dengan algoritma lain serta tanpa proses *hyperparameter tuning* sehingga performa model belum tentu merupakan performa terbaik yang dapat dicapai. Keempat, penelitian belum melakukan validasi eksternal menggunakan dataset dari tahun atau sumber lain, sehingga generalisasi model masih terbatas pada data BRFSS 2022. Keterbatasan-keterbatasan ini perlu diperhatikan untuk pengembangan penelitian selanjutnya.

6.2 *Saran*

Penelitian ini masih memiliki beberapa keterbatasan, terutama pada ketidakseimbangan kelas pada dataset yang berdampak pada hasil prediksi kelas minoritas. Penelitian selanjutnya disarankan untuk menerapkan teknik penyeimbangan data seperti *SMOTE* atau *SMOTE-ENN* agar performa model pada kelas positif semakin optimal. Selain itu, penggunaan algoritma lain seperti *XGBoost*, *Gradient Boosting*, atau *Deep Learning* dapat menjadi alternatif untuk dibandingkan dengan *Random Forest*. Penggunaan teknik ini juga dapat membantu menguji konsistensi hasil model ketika diterapkan pada struktur data yang lebih kompleks. Pendekatan tersebut berpotensi memberikan gambaran yang lebih luas mengenai efektivitas berbagai metode dalam memodelkan risiko penyakit jantung.

Pengembangan antarmuka berbasis aplikasi atau sistem pendukung keputusan juga dapat dilakukan agar model prediksi ini dapat digunakan oleh tenaga kesehatan atau instansi terkait dalam melakukan skrining awal terhadap risiko penyakit jantung di masyarakat. Penelitian selanjutnya dapat dikembangkan dengan melakukan perbandingan beberapa algoritma lain seperti *XGBoost*, *Gradient Boosting*, atau *Neural Network* untuk mengetahui model yang paling optimal pada data BRFSS. Penggunaan teknik penanganan data tidak seimbang seperti *SMOTE*, *ADASYN*, atau *class weighting* juga dapat meningkatkan nilai recall, terutama dalam mendeteksi kasus positif yang jumlahnya lebih sedikit. Selain itu, proses *hyperparameter tuning* pada *Random Forest* atau model lain dapat dilakukan untuk mendapatkan performa model yang lebih optimal. Penelitian berikutnya juga dapat menambahkan data dari tahun berbeda atau menggabungkan beberapa tahun BRFSS untuk melakukan validasi eksternal. Penggunaan metode *explainable AI* seperti *SHAP* dapat memberikan interpretasi yang lebih mendalam mengenai kontribusi masing-masing variabel dalam proses prediksi penyakit jantung.

DAFTAR PUSTAKA

[1] A. Javaid *et al.*, “Medicine 2032: The Future of Cardiovascular Disease Prevention with Machine Learning and digital Health Technology,” *Am. J. Prev. Cardiol.*, vol. 12, p. 100379, Dec. 2022, doi: 10.1016/j.ajpc.2022.100379.

[2] N. Sharma, L. Malviya, A. Jadhav, and P. Lalwani, “A Hybrid Deep Neural Net Learning Model for Predicting Coronary Heart Disease Using Randomized Search Cross-Validation Optimization,” *Decis. Anal. J.*, vol. 9, p. 100331, Dec. 2023, doi: 10.1016/j.dajour.2023.100331.

[3] K. Hacker, “The Burden of Chronic Disease,” *Mayo Clin. Proc. Innov. Qual. Outcomes*, vol. 8, no. 1, pp. 112–119, Feb. 2024, doi: 10.1016/j.mayocpiqo.2023.08.005.

[4] A. Pirillo and G. D. Norata, “The Burden of Hypercholesterolemia and Ischemic Heart Disease in an Ageing World,” *Pharmacol. Res.*, vol. 193, p. 106814, Jul. 2023, doi: 10.1016/j.phrs.2023.106814.

[5] A. Alanazi, “Using Machine Learning for Healthcare Challenges and Opportunities,” *Informatics Med. Unlocked*, vol. 30, p. 100924, Jan. 2022, doi: 10.1016/j.imu.2022.100924.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

[6] 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.

[7] O. Taylan, A. S. Alkabaa, H. S. Alqabbaa, E. Pamukçu, and V. Leiva, "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods," *Biology (Basel).*, vol. 12, no. 1, p. 117, Jan. 2023, doi: 10.3390/biology12010117.

[8] M. T. Ramakrishna, V. K. Venkatesan, I. Izonin, M. Havryliuk, and C. R. Bhat, "Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data," *Entropy*, vol. 25, no. 2, p. 245, Jan. 2023, doi: 10.3390/e25020245.

[9] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, p. 88, Feb. 2025, doi: 10.3390/technologies13030088.

[10] R. Iranzad and X. Liu, "A Review of Random Forest-Based Feature Selection Methods for Data Science Education and Applications," *Int. J. Data Sci. Anal.*, vol. 20, no. 2, pp. 197–211, Feb. 2024, doi: 10.1007/s41060-024-00509-w.

[11] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthcare*, vol. 11, no. 12, p. 1808, Jun. 2023, doi: 10.3390/healthcare11121808.

[12] S. M. Ganie and M. B. Malik, "An Ensemble Machine Learning Approach for Predicting Type-II Diabetes Mellitus Based on Lifestyle Indicators," *Healthc. Anal.*, vol. 2, p. 100092, Nov. 2022, doi: 10.1016/j.health.2022.100092.

[13] D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, "Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization," *Algorithms*, vol. 16, no. 6, p. 308, Jun. 2023, doi: 10.3390/a16060308.

[14] M. Imani and H. R. Arabnia, "Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis," *Technologies*, vol. 11, no. 6, p. 167, Nov. 2023, doi: 10.3390/technologies11060167.

[15] N. Moos, C. Juergens, and A. P. Redecker, "Combined Small- and Large-Scale Geo-Spatial Analysis of the Ruhr Area for an Environmental Justice Assessment," *Sustainability*, vol. 14, no. 6, p. 3447, Mar. 2022, doi: 10.3390/su14063447.

[16] M. Goyal and Q. H. Mahmoud, "A Systematic Review of Synthetic Data Generation Techniques Using Generative AI," *Electronics*, vol. 13, no. 17, p. 3509, Sep. 2024, doi: 10.3390/electronics13173509.

[17] S. Salvi, G. Vu, V. Gurupur, and C. King, "Classifying Tooth Loss and Assessing Risk Factors in U.S. Adults: A Machine Learning Analysis of BRFSS 2022 Data," *Electronics*, vol. 14, no. 17, p. 3559, Sep. 2025, doi: 10.3390/electronics14173559.

[18] J. Allgaier, L. Mulansky, R. L. Draelos, and R. Pryss, "How Does the Model Make Predictions? A Systematic Literature Review on the Explainability Power of Machine Learning in Healthcare," *Artif. Intell. Med.*, vol. 143, p. 102616, Sep. 2023, doi: 10.1016/j.artmed.2023.102616.

[19] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, p. 2333, Feb. 2023, doi: 10.3390/s23042333.

[20] J. H. Cabot and E. G. Ross, "Evaluating Prediction Model Performance," *Surgery*, vol. 174, no. 3, pp. 723–726, Sep. 2023, doi: 10.1016/j.surg.2023.05.023.

[21] N. Chotisarn, S. Gulyanon, T. Zhang, and W. Chen, "VISHIEN-MAAT: Scrollytelling Visualization Design for Explaining Siamese Neural Network Concept to Non-Technical Users," *Vis. Informatics*, vol. 7, no. 1, pp. 18–29, Mar. 2023, doi: 10.1016/j.visinf.2023.01.004.

[22] F. Corradini, L. Mozzoni, M. Piangerelli, B. Re, and L. Rossi, "A Framework for Rapidly Prototyping Data Mining Pipelines," *Big Data Cogn. Comput.*, vol. 9, no. 6, p. 150, Jun. 2025, doi: 10.3390/bdcc9060150.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

[23] E. Coronado, N. Yamanobe, and G. Venture, "NEP+: A Human-Centered Framework for Inclusive Human-Machine Interaction Development," *Sensors*, vol. 23, no. 22, p. 9136, Nov. 2023, doi: 10.3390/s23229136.

[24] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics*, vol. 14, no. 2, p. 144, Jan. 2024, doi: 10.3390/diagnostics14020144.

[25] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare Predictive Analytics Using Machine Learning and Deep Learning Techniques: A Survey," *J. Electr. Syst. Inf. Technol.*, vol. 10, no. 1, pp. 40-, Aug. 2023, doi: 10.1186/s43067-023-00108-y.

[26] F. Diapoldo Silalahi, T. Wijanarko, A. Putra, and E. Siswanto, "Machine Learning Technique for Credit Card Scam Detection," *J. Technol. Informatics Eng.*, vol. 1, no. 1, pp. 50–79, Apr. 2022, doi: 10.51903/jtie.v1i1.143.

[27] S. Wahyuning and S. K. Sudibyo, "Leveraging Machine Learning for Talent Acquisition: Predicting High-Performance Candidates in Human Resource Management," *J. Manag. Informatics*, vol. 3, no. 1, pp. 87–104, Apr. 2024, doi: 10.51903/jmi.v3i1.44.

[28] N. Nikhlis, M. Jamal, and U. Ghofur, "Deteksi Dini Pencurian Data pada Perangkat Seluler Menggunakan Machine Learning," *J. Ilm. Sist. Inf.*, vol. 4, no. 2, pp. 110–129, May 2025, doi: 10.51903/fz0xzt26.