Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

Identifikasi Faktor Risiko Serangan Jantung di Indonesia Menggunakan Model Prediktif LightGBM

Muhammad Fais Ramadhani, Amiq Fahmi*, Ramadhan Rakhmat Sani

Fakultas Ilmu Komputer, Sistem Informasi, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1]112202206959@mhs.dinus.ac.id , [2*]amiq.fahmi@dsn.dinus.ac.id, [3]ramadhan_rs@dsn.dinus.ac.id
Email Penulis Korespondensi: amiq.fahmi@dsn.dinus.ac.id*
Submitted: **03/09/2025**; Accepted: **26/12/2025**; Published: **31/12/2025**

Abstrak– Peningkatan prevalensi serangan jantung di Indonesia telah menjadi isu kesehatan publik yang signifikan karena penyakit ini konsisten termasuk penyebab kematian tertinggi secara nasional. Meskipun berbagai studi epidemiologis telah mengidentifikasi faktor risiko klinis maupun perilaku, pendekatan berbasis data untuk prediksi individual masih relatif terbatas, terutama pada konteks populasi Indonesia dengan karakteristik heterogen. Untuk menjawab kesenjangan tersebut, penelitian ini mengembangkan model prediktif serangan jantung menggunakan algoritma Light Gradient Boosting Machine (LightGBM) yang dikenal efisien pada data berukuran besar. Dataset terdiri dari 158.355 observasi dan 28 fitur demografis, gaya hidup, dan indikator medis. Prosedur prapemrosesan mencakup imputasi nilai hilang, pengkodean variabel kategorikal, seleksi fitur menggunakan Principal Component Analysis (PCA), serta penyeimbangan distribusi kelas melalui Synthetic Minority Over-Sampling Technique (SMOTE). Kinerja prediksi dievaluasi menggunakan metrik klasifikasi standar, di mana LightGBM mencapai akurasi 83,39% (train) dan 77,92% (test); presisi 85,67% dan 79,38%; recall 80,19% dan 75,44%; F1-score 82,84% dan 77,36%; serta AUC-ROC 91,84% dan 87,37%. Analisis komponen utama menunjukkan kontribusi varians yang tinggi pada fitur terkait pola konsumsi, penggunaan obat, stres, dan hipertensi. Hasil ini mengindikasikan bahwa LightGBM merupakan pendekatan yang menjanjikan untuk mendukung deteksi risiko serangan jantung secara lebih awal dan berpotensi meningkatkan strategi mitigasi penyakit kardiovaskular di Indonesia.

Kata Kunci: Serangan Jantung; LightGBM; PCA; SMOTE; AUC-ROC

Abstract– The increase in the prevalence of heart attacks in Indonesia has become a significant public health issue because this disease consistently ranks among the leading causes of death nationwide. Although various epidemiological studies have identified clinical and behavioral risk factors, data-based approaches to individual prediction are still relatively limited, especially in the context of the heterogeneous Indonesian population. To address this gap, this study developed a predictive model for heart attacks using the Light Gradient Boosting Machine (LightGBM) algorithm, which is known to be efficient with large data sets. The data set consisted of 158,355 observations and 28 demographic, lifestyle, and medical indicator features. The preprocessing procedure included imputation of missing values, coding of categorical variables, feature selection using Principal Component Analysis (PCA), and class distribution balancing through the Synthetic Minority Over-Sampling Technique (SMOTE). Prediction performance was evaluated using standard classification metrics, where LightGBM achieved an accuracy of 83.39% (train) and 77.92% (test); precision of 85.67% and 79.38%; recall of 80.19% and 75.44%; F1-score of 82.84% and 77.36%; and AUC-ROC of 91.84% and 87.37%. Principal component analysis showed a high variance contribution in features related to consumption patterns, medication use, stress, and hypertension. These results indicate that LightGBM is a promising approach to support earlier detection of heart attack risk and has the potential to improve cardiovascular disease mitigation strategies in Indonesia.

Keywords: Heart attack; LightGBM; PCA; SMOTE; AUC-ROC

1. PENDAHULUAN

Penyakit kardiovaskular (CVD) merupakan salah satu penyebab utama kematian di dunia, dengan jumlah kasus yang mencapai lebih dari 17,8 juta setiap tahunnya berdasarkan laporan *World Health Organization* (WHO). Salah satu bentuk CVD yang paling mematikan adalah serangan jantung, yaitu kondisi ketika aliran darah ke otot jantung tersumbat sehingga menyebabkan kerusakan jaringan secara permanen [1]. Di Indonesia, kasus serangan jantung menunjukkan tren peningkatan yang signifikan dan mengkhawatirkan, seiring dengan pergeseran gaya hidup masyarakat modern yang cenderung kurang sehat. Pola makan tinggi lemak, kebiasaan merokok, tingginya tingkat stres, serta kurangnya aktivitas fisik menjadi pemicu utama. Faktor risiko medis lain, seperti hipertensi, diabetes melitus, obesitas, dan paparan polusi udara, semakin memperbesar risiko terjadinya serangan jantung pada kelompok usia produktif maupun lanjut usia. Permasalahan ini semakin kompleks karena sebagian besar pasien baru menyadari gejala ketika kondisi klinis telah mencapai tahap kritis, sehingga deteksi dini menjadi aspek fundamental dalam upaya menurunkan angka mortalitas [2].

Di sisi lain, perkembangan teknologi digital dan ketersediaan data kesehatan yang besar membuka peluang baru dalam pencegahan penyakit kardiovaskular melalui pendekatan berbasis data, informasi, dan pengetahuan. Analisis data besar memungkinkan pengolahan rekam medis elektronik, catatan laboratorium, survei gaya hidup, hingga faktor lingkungan secara terpadu [3]. Salah satu dataset representatif yang relevan adalah *Heart Attack Prediction in Indonesia*, yang berisi lebih dari 158.355 catatan dengan 28 atribut mencakup data demografis, klinis,

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

gaya hidup, dan lingkungan. Kompleksitas dan kelengkapan data tersebut memiliki potensi tinggi untuk dimanfaatkan dalam pengembangan model prediksi berbasis machine learning. Dengan demikian, penerapan metode analisis canggih tidak hanya memungkinkan identifikasi individu dengan risiko tinggi, tetapi juga membentuk dasar bagi pengembangan sistem pendukung keputusan klinis yang dapat membantu tenaga kesehatan dalam melakukan intervensi secara lebih cepat, tepat, dan berbasis data, informasi dan pengetahuan [4].

Meskipun peluang tersebut terbuka lebar, terdapat sejumlah tantangan yang perlu diatasi. Pertama, data medis umumnya bersifat berdimensi tinggi, dengan kombinasi atribut numerik dan kategorikal yang kompleks, serta sering kali mengandung nilai yang hilang. Model konvensional seperti regresi logistik sering kali tidak cukup mampu menangkap interaksi non-linier antar variabel yang memengaruhi serangan jantung [5]. Kedua, adanya masalah ketidakseimbangan (*class imbalance*) karena jumlah pasien tanpa serangan jantung jauh lebih banyak daripada yang mengalami serangan. Dalam konteks medis, rendahnya kemampuan model mendeteksi pasien berisiko tinggi dapat menimbulkan konsekuensi fatal [6]. Ketiga, interpretabilitas model juga menjadi isu penting. Prediksi yang akurat tetapi sulit dipahami oleh tenaga medis justru menghambat penerapannya dalam praktik klinis. Oleh karena itu, diperlukan metode prediksi yang tidak hanya unggul dalam performa, tetapi juga transparan, efisien, dan relevan untuk digunakan pada kasus nyata [7].

Selain persoalan teknis, sejumlah penelitian terdahulu cenderung lebih menitikberatkan pada pencapaian akurasi tinggi, namun kurang mempertimbangkan relevansi dan keterterapan hasil dalam konteks praktik klinis dunia nyata. Banyak studi menghasilkan model prediksi dengan performa baik, namun tidak menyertakan analisis faktor risiko dominan yang esensial bagi tenaga kesehatan maupun pembuat kebijakan [8]. Padahal, identifikasi faktor-faktor seperti hipertensi, obesitas, diabetes, atau kebiasaan merokok sangat penting untuk menyusun strategi pencegahan yang lebih terarah, baik melalui program edukasi maupun kebijakan publik. Tanpa informasi ini, model prediksi hanya berfungsi sebatas alat teknis, tanpa memberikan dampak nyata terhadap upaya penurunan angka serangan jantung. Dengan demikian, penelitian yang mengintegrasikan klasifikasi dan identifikasi faktor risiko dominan berpotensi memberikan kontribusi positif serta nilai tambah yang signifikan dalam upaya peningkatan kualitas kesehatan masyarakat.

Sebagai solusi terhadap tantangan tersebut, penelitian ini mengusulkan penerapan algoritma *Light Gradient Boosting Machine* (LightGBM) untuk klasifikasi kasus serangan jantung. LightGBM merupakan algoritme berbasis *gradient boosting decision tree* yang dikenal sangat efisien, cepat, dan akurat dalam pengolahan data tabular berskala besar [9]. Keunggulan LightGBM terletak pada kemampuannya menangani data numerik dan kategorikal secara optimal, menangkap interaksi non-linier antar variabel, serta mengatasi ketidakseimbangan kelas melalui pengaturan bobot parameter. Lebih dari itu, LightGBM mendukung analisis fitur yang memungkinkan peneliti mengidentifikasi faktor risiko dominan secara transparan. Dengan kelebihan ini, LightGBM tidak hanya berperan sebagai model klasifikasi, tetapi juga sebagai alat untuk menghasilkan pemahaman mendalam pengetahuan klinis yang dapat dimanfaatkan oleh tenaga medis dalam pengambilan keputusan berbasis data pasien.

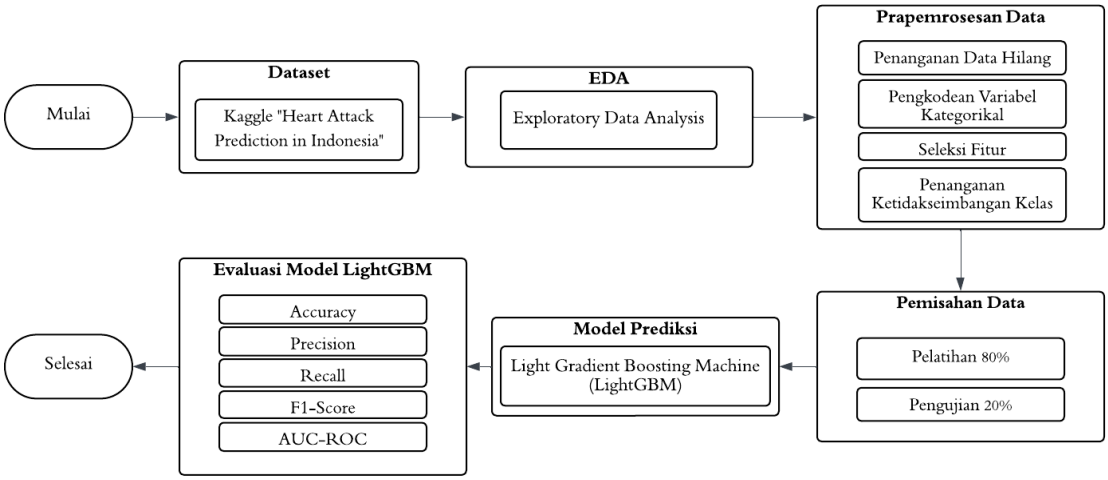
Sejumlah penelitian sebelumnya telah menunjukkan potensi algoritma boosting, termasuk *Light Gradient Boosting Machine* (LightGBM), dalam bidang prediksi kesehatan. Salah satu studi melaporkan bahwa LightGBM mampu mencapai skor F1 sebesar 91,2% dan akurasi sebesar 91,8% dalam memprediksi mortalitas akibat serangan jantung, dengan menggunakan dataset yang terdiri atas 1.700 instansi dan 124 fitur. Dataset tersebut mencakup informasi yang dikumpulkan saat pasien masuk rumah sakit, serta pada 24, 48, dan 72 jam setelah serangan jantung, dan mengandung sekitar 7,6% nilai yang hilang. Meskipun demikian, model LightGBM tetap mampu menangani nilai-nilai yang hilang tersebut secara efektif tanpa pra-pemrosesan tambahan, karena algoritma *boosted tree* secara inheren memiliki mekanisme untuk mengatasi data yang tidak lengkap [10]. Penelitian lain menunjukkan bahwa LightGBM secara konsisten mengungguli algoritme seperti *Support Vector Machine* (SVM) dan *regressi logistic* dalam menghadapi dataset kesehatan yang kompleks [11]. Kajian komparatif juga menemukan bahwa LightGBM memiliki performa mendekati *XGBoost* dengan keunggulan interpretabilitas yang lebih baik [12]. Dalam bidang kardiologi, LightGBM bahkan mampu menghasilkan AUC di atas 92% untuk prediksi hipertensi [13]. Sementara itu, algoritme berbasis *deep learning* seperti *TabNet* memang menawarkan interpretabilitas, tetapi terbukti kurang efisien dalam mengolah dataset besar [14]. Hal ini menegaskan bahwa LightGBM adalah pilihan tepat untuk kasus serangan jantung dengan dataset besar di Indonesia.

Berdasarkan tinjauan literatur sebelumnya, penelitian ini bertujuan untuk mengisi kesenjangan yang masih ada, yaitu terbatasnya studi yang secara spesifik mengaplikasikan algoritma *Light Gradient Boosting Machine* (LightGBM) dalam klasifikasi serangan jantung berbasis data berskala besar di Indonesia, sekaligus menyajikan interpretasi terhadap faktor risiko dominan yang memengaruhi hasil prediksi. Kebaruan penelitian ini terletak pada penerapan LightGBM pada dataset *Heart attack prediction in Indonesia* dengan tujuan menghasilkan model prediksi yang akurat, efisien, dan interpretatif. Selain itu, analisis dari LightGBM diharapkan mampu memberikan wawasan klinis yang bernilai dalam menyusun strategi pencegahan [15]. Penelitian ini juga menekankan pentingnya integrasi pendekatan berbasis data dengan praktik klinis untuk mendukung pengambilan keputusan yang lebih tepat. Dengan demikian, penelitian ini berkontribusi tidak hanya pada pengembangan metode klasifikasi, tetapi juga pada upaya nyata menurunkan angka kematian akibat serangan jantung melalui pemanfaatan teknologi analitik yang berbasis data.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

2. METODOLOGI PENELITIAN

Penelitian ini dilaksanakan melalui serangkaian tahapan sistematis yang bertujuan untuk menghasilkan model prediksi serangan jantung yang valid dan akurat. Proses dimulai dengan pengumpulan dataset dari Kaggle, yang menyediakan data kesehatan masyarakat Indonesia. Selanjutnya, dilakukan *Exploratory Data Analysis* (EDA) untuk memperoleh gambaran menyeluruh mengenai struktur data, sekaligus mendeteksi duplikasi, nilai hilang, serta distribusi variabel numerik dan kategorikal. Tahap berikutnya adalah data *preprocessing*, yang bertujuan menyiapkan dataset agar dapat diproses secara optimal oleh algoritma *machine learning*. Proses ini mencakup penanganan nilai hilang, transformasi variabel kategorikal ke dalam bentuk numerik, seleksi fitur yang relevan, serta penyeimbangan kelas menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Setelah data dibersihkan, dataset dibagi menjadi data latih (80%) dan data uji (20%) dengan menggunakan metode *stratified sampling* untuk menjaga distribusi kelas target secara proporsional. Pengembangan model dilakukan dengan menerapkan algoritma *Light Gradient Boosting Machine* (LightGBM), yang terbukti efisien dan unggul dalam pengolahan data tabular berskala besar. Model yang dihasilkan kemudian dievaluasi menggunakan berbagai metrik, seperti AUC-ROC, akurasi, presisi, recall, dan F1-Score, guna menilai performa prediksi secara komprehensif [16]. Alur penelitian secara visual ditampilkan pada Gambar 1.



Gambar 1. Alur Metodologi Penelitian Prediksi Serangan Jantung

2.1 Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari *platform Kaggle* dengan judul *Heart Attack Prediction in Indonesia*, yang memuat sebanyak 158.355 data individu. Dataset tersebut terdiri atas 28 variabel independen dan satu variabel target biner. Variabel target, yaitu *heart_attack*, bernilai 1 apabila individu mengalami serangan jantung, dan bernilai 0 jika tidak. Variabel independen mencakup beragam faktor, antara lain faktor demografi (usia, jenis kelamin, tingkat pendapatan), faktor gaya hidup (kebiasaan merokok, konsumsi alkohol, aktivitas fisik), serta faktor klinis (hipertensi, diabetes, obesitas, dan kadar kolesterol). Selain itu, terdapat pula hasil pemeriksaan medis seperti tekanan darah, gula darah puasa, trigliserida, kolesterol HDL dan LDL, serta rekam elektrokardiogram (EKG). Variabel lingkungan dan psikososial, seperti paparan polusi, tingkat stres, dan durasi tidur, turut melengkapi kompleksitas data. Dengan jumlah data yang besar dan keragaman atribut yang luas, dataset ini memberikan landasan yang komprehensif untuk pengembangan model prediksi serangan jantung yang lebih akurat dan aplikatif. Struktur sampel dataset ditampilkan pada Tabel 1.

Tabel 1. Struktur Dataset Penelitian

Fitur	Tipe Data	Nilai
Age	Numerik	25 – 90
Gender	Kategorikal	Male, Female
Region	Kategorikal	Urban, Rural
income_level	Kategorikal	Low, Middle, High

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

Fitur	Tipe Data	Nilai
hypertension	Biner	1=Ya, 0=Tidak
Diabetes	Biner	1=Ya, 0=Tidak
cholesterol_level	Numerik	mg/dL
Obesity	Biner	1=Ya, 0=Tidak
waist_circumference	Numerik	Cm
family_history	Biner	1=Ya, 0=Tidak
smoking_status	Kategorikal	Never, Past, Current
alcohol_consumption	Kategorikal	None, Moderate, High
physical_activity	Kategorikal	Low, Moderate, High
dietary_habits	Kategorikal	Healthy, Unhealthy
air_pollution_exposure	Kategorikal	Low, Moderate, High
stress_level	Kategorikal	Low, Moderate, High
sleep_hours	Numerik	3-9 jam
blood_pressure_systolic	Numerik	mmHg
blood_pressure_diastolic	Numerik	mmHg
fasting_blood_sugar	Numerik	mg/dL
cholesterol_hdl	Numerik	mg/dL
cholesterol_ldl	Numerik	mg/dL
triglycerides	Numerik	mg/dL
EKG_results	Kategorikal	Normal, Abnormal
previous_heart_disease	Biner	1=Ya, 0=Tidak
medication_usage	Biner	1=Ya, 0=Tidak
participated_in_free_screening	Biner	1=Ya, 0=Tidak
heart_attack (target)	Biner	1=Ya, 0=Tidak

2.2 *Exploratory Data Analysis* (EDA)

Pada tahap *Exploratory Data Analysis* (EDA), dilakukan analisis menyeluruh terhadap struktur dan kualitas data guna memastikan bahwa dataset yang digunakan layak untuk dianalisis lebih lanjut. Salah satu langkah penting dalam EDA adalah pemeriksaan terhadap keberadaan *missing values* dan duplikasi data. Hasil analisis menunjukkan bahwa sebagian besar variabel dalam dataset bebas dari nilai hilang, kecuali pada variabel *alcohol_consumption* yang mencatatkan 94.848 nilai hilang, mencakup hampir 60% dari total data. Temuan ini mengindikasikan adanya kekosongan informasi yang substansial terkait kebiasaan konsumsi alkohol sebagian besar individu dalam dataset. Di sisi lain, tidak ditemukan duplikasi data, yang menunjukkan bahwa setiap entri bersifat unik dan tidak terdapat pengulangan informasi yang dapat memengaruhi hasil analisis. Temuan awal ini memberikan gambaran mengenai kualitas data serta menjadi dasar dalam menentukan langkah korektif, khususnya dalam penanganan missing values pada variabel yang bersangkutan. Dengan demikian, EDA tidak hanya berfungsi sebagai tahap awal, tetapi juga merupakan komponen krusial dalam memastikan bahwa data yang digunakan telah terstruktur dengan baik dan siap untuk tahap pemodelan selanjutnya, sehingga mendukung pelaksanaan penelitian yang akurat dan efisien.

2.3 Prapemrosesan Data

Tahap *preprocessing* merupakan langkah fundamental dalam pembangunan model *machine learning* yang efektif, dengan tujuan menyiapkan dataset agar dapat diproses secara optimal oleh algoritma yang digunakan. Kualitas data pada tahap ini memiliki pengaruh yang sangat signifikan terhadap performa model prediksi. Tanpa proses *preprocessing* yang tepat, model yang dihasilkan berisiko tidak akurat, tidak efisien, atau bahkan tidak dapat digunakan sama sekali. Oleh karena itu, tahapan ini harus dilaksanakan secara cermat dan sistematis guna memastikan bahwa data yang digunakan memiliki kualitas terbaik. Proses utama dalam tahap preprocessing mencakup penanganan nilai hilang, transformasi variabel kategorikal ke dalam bentuk numerik, seleksi fitur yang relevan, serta penyeimbangan kelas untuk mengatasi ketimpangan distribusi data.

2.3.1 Penanganan Data Hilang

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

Penanganan *missing values* sangat penting untuk memastikan bahwa model yang dibangun tidak terganggu oleh ketidakhadiran data pada variabel-variabel tertentu. Salah satu metode yang digunakan untuk imputasi nilai hilang adalah *Multiple Imputation by Chained Equations* (MICE). MICE bekerja dengan memanfaatkan hubungan antarvariabel dalam dataset untuk memperkirakan nilai yang hilang dengan cara yang lebih canggih dibandingkan imputasi sederhana seperti rata-rata atau median. Imputasi pada MICE dilakukan dengan menduga nilai yang hilang berdasarkan informasi dari variabel lainnya. Setiap nilai hilang pada satu variabel diimputasi dengan menggunakan variabel lain yang ada dalam dataset yang memiliki hubungan kuat dengan variabel yang hilang. Proses ini dilakukan secara iteratif, di mana setiap variabel yang hilang diimputasi satu per satu, menggunakan nilai prediksi yang didasarkan pada hubungan antarvariabel yang ada di dalam data [9]. Rumus imputasi (1).

$$\hat{x}_j = f(x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_p) \quad (1)$$

Dimana:

1. $\hat{x}_j$x^j = nilai prediksi untuk variabel ke-jj yang hilang.
2. $f(\cdot)$f(·) = fungsi estimasi berdasarkan variabel lain.
3. $x_1, x_2, \dots, x_p$x1,x2,…,xp = himpunan variabel pada dataset.

2.3.2 Pengkodean Variabel Kategorikal

Salah satu teknik yang umum digunakan untuk mengonversi variabel kategorikal menjadi format numerik adalah *One-Hot Encoding*. Proses ini mengubah setiap kategori dalam variabel menjadi variabel biner yang terpisah. Setiap kategori yang ada akan dipetakan ke dalam kolom terpisah, di mana setiap baris data akan mendapatkan nilai 0 atau 1, tergantung pada kategori yang ada dalam data tersebut. Rumus *One-Hot Encoding*(2).

$$x_{i,k'} = \begin{cases} 1, & \text{jika } x_i = k \\ 0, & \text{jika } x_i \neq k \end{cases} \quad (2)$$

Dimana:

1 $x_{i,k'}$xi,k′ = variabel dummy untuk sampel ke-ii pada kategori ke-kk.
2 $x_i$xi = nilai asli dari sampel ke-ii.
3 kk = kategori unik dari variabel kategorikal.

2.3.3 Pemilihan Fitur

Pemilihan fitur adalah tahap penting dalam preprocessing data yang bertujuan untuk memilih variabel yang paling relevan dengan target prediksi, guna meningkatkan akurasi dan efisiensi model. Teknik seleksi fitur ini bertujuan untuk mengurangi kompleksitas model, mencegah overfitting, dan memastikan hanya fitur yang memberikan kontribusi signifikan yang dipertahankan dalam analisis. Salah satu teknik populer dalam seleksi fitur adalah *Principal Component Analysis* (PCA), yang digunakan untuk mereduksi dimensi data sambil mempertahankan sebagian besar informasi yang relevan. PCA adalah metode statistik yang digunakan untuk mengurangi dimensi data dengan memproyeksikan data asli ke dalam ruang dimensi yang lebih rendah. Dalam konteks seleksi fitur, PCA membantu mengidentifikasi fitur mana yang paling berkontribusi terhadap variasi dalam dataset. Dimana, alih-alih memilih fitur berdasarkan korelasi atau redundansi, PCA menyaring fitur-fitur yang tidak berkontribusi signifikan dan memilih komponen yang memiliki varians terbesar. PCA juga memungkinkan pengurangan dimensi yang membantu mempercepat proses pelatihan model, meningkatkan interpretabilitas, dan mengurangi risiko overfitting [17]. PCA sangat sensitif terhadap skala variabel, sehingga langkah pertama adalah menstandarisasi data sehingga setiap fitur memiliki mean 0 dan deviasi standar 1. Hal ini penting agar fitur dengan varians tinggi tidak mendominasi komponen utama. Rumus Standarisasi (*Z-score Standardization*) (3).

$$z_i = \frac{x_i - \mu}{\sigma} \text{zi=σxi−μ} \quad (3)$$

Dimana:

1 $z_i$zi adalah nilai standarisasi untuk fitur ke-ii,
2 $x_i$xi adalah nilai asli fitur ke-ii,
3 μμ adalah mean dari fitur,
4 σσ adalah deviasi standar fitur.

2.3.4 Penanganan Ketidakseimbangan Kelas

Dalam data prediksi serangan jantung di indonesia, kelas minoritas adalah pasien yang mengalami serangan jantung, yang jumlahnya jauh lebih sedikit dibandingkan dengan pasien yang tidak mengalami serangan jantung. Ketidakseimbangan ini dapat menyebabkan model lebih cenderung memprediksi kelas mayoritas, mengabaikan

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

kelas minoritas yang justru lebih penting untuk dideteksi. Untuk mengatasi masalah ini, *Synthetic Minority Oversampling Technique* (SMOTE) digunakan sebagai teknik oversampling untuk menyeimbangkan distribusi kelas. SMOTE bekerja dengan menciptakan sampel sintetis baru untuk kelas minoritas dengan cara menginterpolasi antara dua sampel yang ada. Proses ini membantu meningkatkan jumlah sampel dalam kelas minoritas, memungkinkan model untuk belajar dengan lebih baik mengenai pola yang terdapat dalam kelas minoritas [18], [19]. Rumus pembentukan sampel sintetis(4).

$$x_{baru} = x_i + \lambda(x_j - x_i), \lambda \sim U(0,1) \quad (4)$$

Dimana:

1 x_{baru}xbaru = sampel sintetis hasil interpolasi.
2 x_i, x_jxi,xj = dua sampel dari kelas minoritas.
3 $\lambda\lambda$ = bilangan acak dari distribusi uniform $U(0,1)$U(0,1).

Dengan SMOTE, distribusi kelas target lebih seimbang sehingga model dapat belajar secara adil.

2.4 Pemisahan Data

Dataset yang telah diproses dibagi menjadi dua subset, yaitu *training set* dan *testing set*. Teknik pembagian menggunakan *stratified sampling* agar proporsi kelas target pada kedua subset tetap seimbang. Data latih digunakan untuk melatih model [20], sementara data uji dipakai untuk mengukur kemampuan generalisasi. Secara matematis, pembagian dataset dituliskan dengan matematis (5).

$$| D_{train} | = 0.8n, | D_{test} | = 0.2n \quad (5)$$

Dimana:

1 D = himpunan seluruh data dengan ukuran nn.
2 D_{train}Dtrain = 80% dari data (data latih).
3 D_{test}Dtest = 20% dari data (data uji).
4 $\varnothing\emptyset$ = himpunan kosong, artinya tidak ada irisan antara data latih dan data uji.

Strategi ini memastikan model memiliki cukup data untuk belajar sekaligus dapat diuji secara objektif, serta mencegah *overfitting*.

2.5 Model Prediksi *Light Gradient Boosting Machine* (*LightGBM*)

Model studi ini menggunakan *LightGBM*, sebuah algoritma yang merupakan pengembangan dari *Gradient Boosting Decision Tree* (GBDT), yang dikenal karena keunggulannya dalam menangani data besar dengan efisien dan menghasilkan akurasi tinggi. *LightGBM* menggunakan strategi *leaf-wise tree growth*, yang memilih daun dengan pengurangan kerugian terbesar, berbeda dari metode boosting konvensional yang berbasis kedalaman pohon. Hal ini memungkinkan *LightGBM* menangkap pola data yang lebih kompleks secara lebih efektif. Agar model tidak overfit, *LightGBM* tetap membatasi kedalaman maksimum pohon. Proses pembangunan model dilakukan dengan mengubah data latih menjadi bentuk histogram untuk mempercepat pemisahan node, dan dilakukan secara iteratif dengan menambahkan pohon baru yang memperbaiki kesalahan dari iterasi sebelumnya [21]. Untuk meningkatkan performa model, dilakukan optimasi hyperparameter menggunakan Optuna, yang membantu memilih kombinasi parameter terbaik seperti jumlah pohon, laju pembelajaran, dan jumlah daun per pohon, untuk mendapatkan model dengan hasil terbaik. Optuna memungkinkan pencarian ruang parameter yang lebih efisien, sehingga model *LightGBM* dapat bekerja optimal dalam memprediksi serangan jantung [22]. Dengan mekanisme tersebut, *LightGBM* mampu menghasilkan model yang seimbang antara akurasi prediksi dan efisiensi komputasi, sehingga sangat sesuai untuk memproses dataset berskala besar seperti data prediksi serangan jantung dalam penelitian ini. *LightGBM* membangun model dengan menambahkan pohon keputusan secara bertahap untuk meminimalkan fungsi objektif. Rumus model *LightGBM*(6).

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \Omega(f) \quad (6)$$

Dimana:

1 $l(y_i, \hat{y}_i)$l(yi,y^i) = fungsi kerugian loss function, misalnya *binary cross-entropy*, yang menghitung selisih antara nilai aktual $y_i$yi dan prediksi $\hat{y}_i$y^i.
2 $\Omega(f)\Omega(f)$ = fungsi regularisasi untuk mengendalikan kompleksitas model agar tidak *overfitting*.
3 $\theta\theta$ = parameter model yang dioptimalkan.
4 nn = jumlah data latih.

2.6 Evaluasi Model

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

Evaluasi model dilakukan menggunakan lima metrik utama: akurasi, presisi, recall, skor F1, dan AUC-ROC. Metrik-metrik ini mengukur seberapa baik model dalam membuat prediksi yang benar, mendeteksi kasus positif, dan membedakan antara kelas positif dan negatif. Dengan menggunakan kelima metrik ini, kita dapat menilai kinerja model secara menyeluruh dan mengidentifikasi area yang perlu ditingkatkan [23].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Mengukur proporsi prediksi yang benar terhadap seluruh data.

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

Menunjukkan tingkat ketepatan prediksi positif.

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

Mengukur kemampuan model dalam menemukan semua kasus positif.

$$\text{F1} - Score = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (10)$$

Merupakan rata-rata harmonis presisi dan recall, cocok untuk data tidak seimbang.

$$AUC = \int_0^1 T\,PR(FPR)\,d(FPR) \quad (11)$$

Dimana:
$TPR = \frac{TP}{TP+FN}$TPR=TP+FNTP (*True Positive Rate*) dan $FPR = \frac{FP}{FP+TN}$FPR=FP+TNFP (*False Positive Rate*).

Metrik-metrik ini memberikan evaluasi menyeluruh: akurasi menunjukkan kebenaran prediksi secara umum, presisi menekankan keakuratan prediksi positif, recall menyoroti kemampuan mendeteksi kasus positif, F1-score menyeimbangkan keduanya, dan AUC-ROC mengukur kemampuan membedakan kelas positif dan negatif pada berbagai *threshold*.

3. HASIL DAN PEMBAHASAN

Temuan dalam penelitian ini disajikan secara sistematis, dimulai dari tahap preprocessing, hingga penerapan algoritma *LightGBM* untuk analisis prediktif. Setiap langkah dalam proses ini diuraikan dengan jelas untuk menunjukkan bagaimana setiap tahapan berkontribusi pada hasil yang diperoleh. Fokus utama evaluasi terletak pada metrik-metrik kinerja seperti akurasi, presisi, recall, F1-score, dan AUC-ROC, yang digunakan untuk menilai seberapa efektif dan kuat model prediksi yang dibangun. Dengan pendekatan yang transparan dan terstruktur ini, penelitian bertujuan memberikan pemahaman mendalam tentang faktor-faktor yang mempengaruhi prediksi serangan jantung serta menilai kinerja model dalam konteks aplikasi nyata. Penjelasan lebih rinci mengenai hasil ini akan dibahas pada bagian selanjutnya.

3.1 Data Prapemrosesan

3.1.1 Penanganan Data Hilang

Pada tahap data preprocessing, penanganan terhadap *missing values* dilakukan secara sistematis, terutama pada variabel "*alcohol_consumption*", yang memiliki jumlah missing values yang cukup besar, mencapai sekitar 60% dari total data. Hasil imputasi ini menghilangkan nilai hilang pada variabel tersebut, sehingga dataset menjadi lengkap dan siap untuk dianalisis lebih lanjut. Gambar 2 yang disajikan di bawah ini menunjukkan perbandingan jumlah missing values pada variabel "*alcohol_consumption*" sebelum dan setelah proses imputasi, yang menggambarkan pengurangan signifikan dalam nilai hilang, memastikan data yang lebih bersih dan siap digunakan untuk model prediksi selanjutnya. Gambar 2 dan Gambar 3 berikut memperlihatkan kondisi sebelum dan sesudah proses imputasi, yang menunjukkan perubahan jumlah missing values pada variabel terkait.

alcohol_consumption	94848

Gambar 2. Missing Values Sebelum Imputasi

alcohol_consumption	0

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

Gambar 3. Missing Values Setelah Imputasi

3.1.2 Pengkodean Variabel Kategorikal

Setelah melakukan encoding variabel kategorikal menggunakan *One-Hot Encoding*, hasilnya menunjukkan bahwa setiap kategori dalam variabel seperti *gender*, *region*, *income_level*, *smoking_status*, *physical_activity*, *dietary_habits*, *air_pollution_exposure*, *stress_level*, dan *EKG_results* telah berhasil diubah menjadi kolom baru dengan nilai biner (0 atau 1). Hasil encoding ini memudahkan model untuk menganalisis data, memastikan tidak ada informasi yang hilang atau terdistorsi dalam proses pengolahan data. Gambar 4 dan 5 menunjukkan hasil encoding variabel kategrikal tersebut.

	gender	region	income_level	smoking_status	physical_activity	dietary_habits	air_pollution_exposure	stress_level	EKG_results
0	Male	Rural	Middle	Never	High	Unhealthy	Moderate	Moderate	Normal
1	Female	Urban	Low	Past	Moderate	Healthy	High	High	Normal
2	Female	Urban	Low	Past	Moderate	Healthy	Low	Low	Abnormal
3	Male	Urban	Low	Never	Moderate	Unhealthy	Low	High	Normal
4	Male	Urban	Middle	Current	Moderate	Unhealthy	High	Moderate	Normal

Gambar 4. Sebelum Encoding

	gender	region	income_level	smoking_status	physical_activity	dietary_habits	air_pollution_exposure	stress_level	EKG_results
0	1	0	2	1	0	1	2	2	1
1	0	1	1	2	2	0	0	0	1
2	0	1	1	2	2	0	1	1	0
3	1	1	1	1	2	1	1	0	1
4	1	1	2	0	2	1	0	2	1

Gambar 5. Sesudah Encoding

3.1.3 Pemilihan Fitur

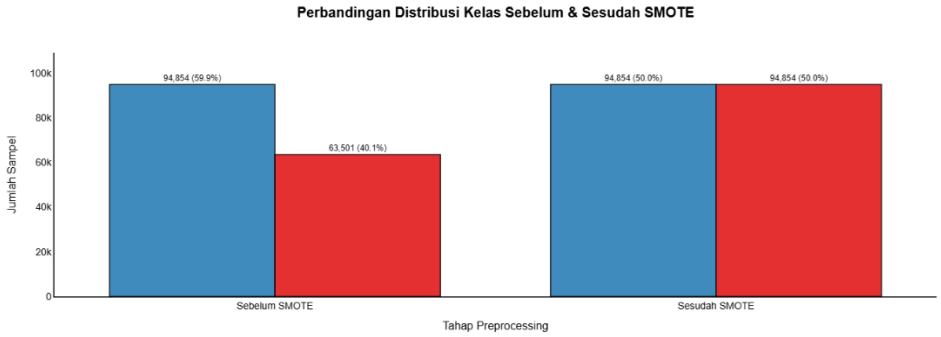
Hasil PCA pada tahap pemilihan fitur memberikan pemahaman yang jelas tentang fitur mana yang memberikan kontribusi paling besar dalam menjelaskan variansi data terkait prediksi serangan jantung. PCA digunakan untuk mengurangi dimensi data tanpa kehilangan informasi penting, dengan fokus pada komponen utama yang memiliki kontribusi varians terbesar. Hasil PCA pada Tabel 2. menunjukkan bahwa beberapa fitur memiliki kontribusi yang signifikan terhadap variabilitas dataset, sementara fitur lainnya memiliki kontribusi yang lebih rendah. Berikut adalah hasil table kontribusi varians yang menampilkan 10 teratas masing-masing fitur berdasarkan hasil PCA:

Tabel 2. Hasil PCA 10 teratas

No	Fitur Terbaik	Kontribusi Varian (%)
1	dietary_habits	9.96
2	medication_usage	9.93
3	obesity	9.60
4	family_history	9.33
5	region	8.89
6	stress_level	8.50
7	Hypertension	8.42
8	participated_in_free_screening	8.22
9	Gender	7.91
10	previous_heart_disease	4.64

3.1.4 Penanganan Ketidakseimbangan Kelas

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065



Gambar 6. Sebelum dan Sesudah SMOTE

Distribusi kelas sebelum penerapan SMOTE menunjukkan ketidakseimbangan yang jelas, dengan kelas mayoritas berjumlah 94.854 sampel (59,9%) dan kelas minoritas hanya 63.501 sampel (40,1%). Kondisi ini menyebabkan model cenderung lebih akurat dalam memprediksi kelas mayoritas, namun memiliki kemampuan deteksi yang rendah untuk kasus-kasus kelas minoritas yang lebih penting, terutama dalam konteks prediksi serangan jantung.

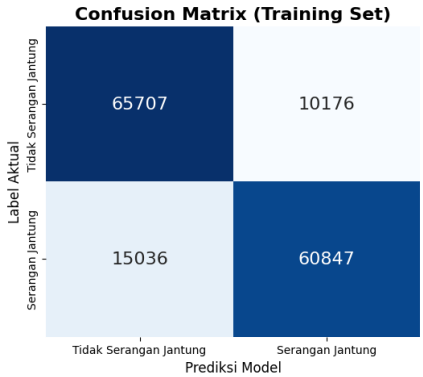
Setelah penerapan SMOTE, distribusi kelas diperbaiki dengan menambahkan sampel sintetis pada kelas minoritas. Hasilnya, distribusi kedua kelas menjadi lebih seimbang, dengan masing-masing kelas memiliki 94.854 sampel (50%). Hal ini menunjukkan bahwa SMOTE berhasil meningkatkan kesetaraan antara kelas mayoritas dan minoritas, sehingga model tidak lagi terdistorsi oleh ketidakseimbangan kelas. Gambar 6. Menunjjukan jumlah instan kelas sebelum dan sesudah penyeimbangan data menggunakan SMOTE.

3.2 Model *Light Gradient Boosting Machine* (LightGBM)

Hasil model algoritma *LightGBM* mampu menangani dataset berukuran besar, serta akurasi prediksi yang tinggi. Dataset terdiri atas 158.355 sampel dengan 10 variabel independen terbaik, dengan pendekatan pemodelan yang mampu mengolah data kompleks secara optimal. Proses pencarian parameter terbaik dilakukan melalui teknik optimisasi hiperparameter berbasis Optuna, dengan menjalankan 30 kali percobaan untuk mengeksplorasi kombinasi parameter yang berbeda, seperti *n_estimators*, *learning_rate*, *num_leaves*, *max_depth*, *colsample_bytree*, *subsample*, *reg_alpha*, dan *reg_lambda*. Hasil pencarian menunjukkan bahwa model terbaik diperoleh pada percobaan ke-20, dengan nilai AUC sebesar 0.9184, yang menandakan kemampuan diskriminasi model dalam membedakan kasus serangan jantung dan non-serangan jantung sangat baik. Parameter optimal yang diperoleh antara lain *n_estimators* = 1089, *learning_rate* = 0.001762, *num_leaves* = 28, *max_depth* = 13, *colsample_bytree* = 0.6688, *subsample* = 0.6612, *reg_alpha* = 3.2677e-07, dan *reg_lambda* = 1.9482e-06.

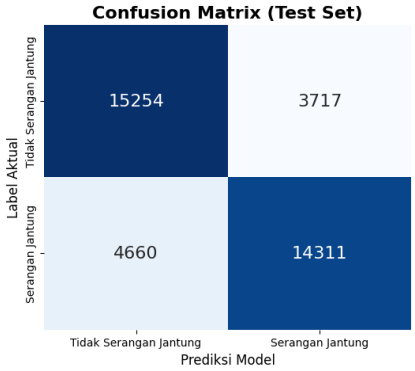
3.3 Evaluasi Model

Evaluasi kinerja model dilakukan untuk menilai sejauh mana algoritma *LightGBM* mampu memprediksi risiko serangan jantung dengan baik pada data latih maupun data uji. Proses evaluasi ini menggunakan lima metrik utama, yaitu akurasi, presisi, recall, F1-score, dan AUC-ROC, yang dipilih untuk memberikan gambaran menyeluruh terhadap kemampuan generalisasi model.



Gambar 7. *Convusion Matrix (Training Set)*

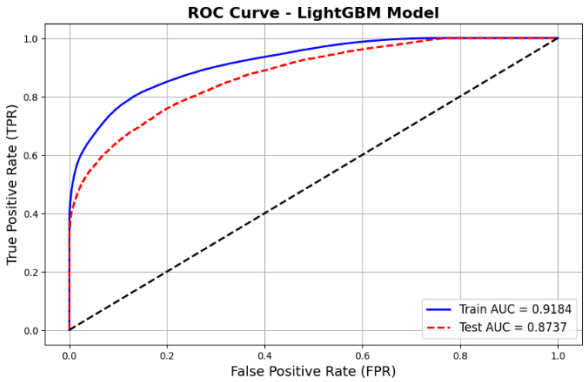
Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065



Gambar 8. *Convusion Matrix (Test Set)*

Gambar 7 menampilkan *confusion matrix* pada data latih dan data uji. Pada *training set* , model mampu mengklasifikasikan sebagian besar sampel dengan benar, ditunjukkan oleh nilai *True Positive* 60.847 dan *True Negative* 65.707 yang dominan. Meskipun masih terdapat kesalahan prediksi *False Positive* 10.176 dan *False Negative* 15.036, model tetap menunjukkan akurasi yang tinggi pada data latih.

Sementara itu, pada testing set gambar 8, pola yang sama juga terlihat. Model dapat mengidentifikasi 14.311 kasus positif serangan jantung secara benar *True Positive* dan 15.254 kasus negatif *True Negative*. Namun, terdapat sejumlah kesalahan prediksi, yaitu 3.717 *False Positive* dan 4.660 *False Negative*. Hasil ini menegaskan bahwa performa model pada data uji relatif konsisten, meskipun sedikit lebih rendah dibanding data latih, yang merupakan indikasi bahwa model memiliki kemampuan generalisasi yang baik.



Gambar 9. ROC Curve

Gambar 9 menampilkan kurva ROC pada data latih dan data uji. Nilai AUC pada data latih sebesar 0.9184, yang mengindikasikan kemampuan diskriminatif model sangat tinggi dalam membedakan antara kelas positif dan negatif. Pada data uji, AUC sebesar 0.8737, yang tetap menunjukkan performa prediksi yang kuat meskipun terjadi sedikit penurunan dibanding data latih. Perbedaan ini wajar, karena data uji digunakan sebagai representasi kasus baru yang tidak pernah dilihat model sebelumnya.

Tabel 3. Hasil Evaluasi Model LightGBM pada Data Latih dan Data Uji

Dataset	Akurasi	Presisi	Recall	F1-Score	AUC
Training Set	83,39%	85,67%	80,19%	82,84%	91,84%
Testing Set	77,92%	79,38%	75,44%	77,36%	87,37%

Tabel 3 menyajikan ringkasan metrik evaluasi model dalam bentuk persentase, yang menunjukkan bahwa algoritma *LightGBM* berhasil membangun model prediksi serangan jantung dengan performa yang solid. Pada data pelatihan, model mencapai akurasi sebesar 83,39%, presisi 85,67%, recall 80,19%, skor F1 82,84%, dan AUC 91,84%. Sementara itu, pada data pengujian, akurasi tercatat sebesar 77,92%, presisi 79,38%, recall 75,44%, skor F1 77,36%, dan AUC 87,37%. Nilai AUC yang tinggi pada kedua dataset menunjukkan kemampuan model dalam

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

membedakan kelas secara konsisten, sedangkan keseimbangan antara presisi dan recall yang tercermin dari skor F1 di atas 77% pada data uji mengindikasikan bahwa model tidak hanya akurat, tetapi juga andal dalam mendeteksi kasus positif suatu aspek yang krusial dalam konteks medis. Perbedaan performa antara data pelatihan dan pengujian dapat dijelaskan oleh karakteristik dataset yang digunakan dalam studi ini, yang berskala besar dan bersifat populatif dengan fitur yang lebih sederhana. Meskipun hasilnya sedikit lebih rendah dibandingkan studi klinis sebelumnya [10], yang menggunakan data dengan kompleksitas dan kedalaman fitur yang lebih tinggi, model yang dikembangkan tetap relevan dan bermanfaat untuk prediksi risiko secara luas di masyarakat, serta berpotensi mendukung sistem pendukung keputusan dalam upaya pencegahan penyakit kardiovaskular secara lebih inklusif dan aplikatif.

4. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa algoritma *Light Gradient Boosting Machine* (*LightGBM*) mampu membangun model prediksi serangan jantung dengan performa yang kuat pada data kesehatan di Indonesia. Evaluasi model yang disajikan menegaskan bahwa kinerja *LightGBM* cukup konsisten baik pada data latih maupun data uji. Pada data latih, model mencapai akurasi sebesar 83,39%, presisi 85,67%, recall 80,19%, F1-score 82,84%, dan AUC 91,84%, yang menunjukkan kemampuan model dalam mempelajari pola data dengan baik tanpa mengalami overfitting yang berlebihan. Sementara itu, pada data uji, performa model tetap kompetitif dengan akurasi 77,92%, presisi 79,38%, recall 75,44%, F1-score 77,36%, dan AUC 87,37%. Nilai AUC yang tinggi pada kedua subset menegaskan kemampuan *LightGBM* dalam membedakan pasien yang berisiko mengalami serangan jantung dan yang tidak. Selain performa model, penelitian ini juga mengidentifikasi faktor risiko signifikan yang berkontribusi terhadap kemungkinan terjadinya serangan jantung. Berdasarkan analisis PCA, variabel dengan kontribusi terbesar antara lain pola makan 9,96%, penggunaan obat-obatan 9,93%, obesitas 9,60%, riwayat keluarga 9,33%, wilayah tempat tinggal 8,89%, serta tingkat stress 8,50%. Faktor klinis seperti hypertensi 8,42% dan riwayat penyakit jantung 4,64% juga terbukti memiliki pengaruh besar dalam menjelaskan variansi data, sehingga menegaskan relevansi faktor gaya hidup, riwayat kesehatan, dan kondisi lingkungan terhadap risiko penyakit kardiovaskular. Secara keseluruhan, model *LightGBM* dalam penelitian ini dapat dijadikan dasar untuk sistem deteksi dini risiko serangan jantung. Namun, untuk meningkatkan performa, terutama pada aspek recall, penelitian lanjutan disarankan untuk mengeksplorasi kombinasi metode optimisasi lain serta mempertimbangkan integrasi variabel klinis tambahan. Dengan pengembangan lebih lanjut, model ini berpotensi besar dalam mendukung strategi pencegahan serangan jantung dan pengambilan keputusan di bidang kesehatan masyarakat.

REFERENCES

[1] E. F. Laili, Z. Alawi, R. Rohmah, and M. A. Barata, “komparasi algoritma decision tree dan support vector machine (svm) dalam klasifikasi serangan jantung,” *J. Sist. Inf. Dan Inform. Simika*, vol. 8, no. 1, pp. 67–76, Jan. 2025, doi: 10.47080/simika.v8i1.3683.

[2] M. A. Sembiring, “analisis faktor prediksi diagnosa tingkat serangan jantung menggunakan metode regression,” *J. Tek.*, vol. 4, no. 1, p. 16, Feb. 2024, doi: 10.54314/teknisi.v4i1.1800.

[3] H. Yang, Z. Chen, H. Yang, and M. Tian, “Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison,” *IEEE Access*, vol. 11, pp. 23366–23380, 2023, doi: 10.1109/ACCESS.2023.3253885.

[4] N. Nandal, L. Goel, and R. Tanwar, “Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis,” *F1000Research*, vol. 11, p. 1126, Sep. 2022, doi: 10.12688/f1000research.123776.1.

[5] Y. Xue *et al.*, “The Prediction Models for High-Risk Population of Stroke Based on Logistic Regressive Analysis and Lightgbm Algorithm Separately,” *Iran. J. Public Health*, May 2022, doi: 10.18502/ijph.v51i5.9415.

[6] D. Dablain, B. Krawczyk, and N. V. Chawla, “DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: 10.1109/TNNLS.2021.3136503.

[7] T. Md. N. U. Akhund and W. M. Al-Nuwaiser, “Improving Prediction Efficiency of Machine Learning Models for Cardiovascular Disease in IoST-Based Systems through Hyperparameter Optimization,” *Comput. Mater. Contin.*, vol. 80, no. 3, pp. 3485–3506, 2024, doi: 10.32604/cmc.2024.054222.

[8] M. H. Al-Adhaileh, M. I. Ahmed Al-mashhadani, E. M. Alzahrani, and T. H. H. Aldhyani, “Improving Heart Attack Prediction Accuracy Performance Using Machine Learning and Deep Learning Algorithms,” *Iraqi J. Comput. Sci. Math.*, vol. 6, no. 2, Apr. 2025, doi: 10.52866/2788-7421.1239.

[9] T. O. Omotehinwa, D. O. Oyewola, and E. G. Moung, “Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease,” *Inform. Health*, vol. 1, no. 2, pp. 70–81, Sep. 2024, doi: 10.1016/j.infoh.2024.06.001.

[10] A. L. G. Vicente, R. D. M. Junior, and R. A. F. Romero, “Explainable LightGBM Approach for Predicting Myocardial Infarction Mortality,” Apr. 23, 2024, *arXiv*: arXiv:2404.15029. doi: 10.48550/arXiv.2404.15029.

[11] R. Han, R. Meng, and Q. Zhu, “Predictive Analytics in Heart Disease: Leveraging LightGBM for Improved Diagnostic Accuracy”.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9065

[12] J. Miah, D. M. Ca, M. A. Sayed, E. R. Lipu, F. Mahmud, and S. M. Y. Arafat, "Improving Cardiovascular Disease Prediction Through Comparative Analysis of Machine Learning Models: A Case Study on Myocardial Infarction," in *2023 15th International Conference on Innovations in Information Technology (IIT)*, Al Ain, United Arab Emirates: IEEE, Nov. 2023, pp. 49–54. doi: 10.1109/IIT59782.2023.10366476.

[13] X. Ji *et al.*, "Prediction Model of Hypertension Complications Based on GBDT and LightGBM," *J. Phys. Conf. Ser.*, vol. 1813, no. 1, p. 012008, Feb. 2021, doi: 10.1088/1742-6596/1813/1/012008.

[14] S. Rao *et al.*, "An explainable Transformer-based deep learning model for the prediction of incident heart failure".

[15] N. M. K. Ramalingamsakthivelan, V. Silambarasan, S. Thavasi, and P. V. Shankar, "Heart Disease Risk Assessment by Using LightGBM Technique," vol. 5, no. 2, 2023.

[16] N. A. Baghdadi, S. M. Farghaly Abdelaliem, A. Malki, I. Gad, A. Ewis, and E. Atlam, "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis," *J. Big Data*, vol. 10, no. 1, p. 144, Sep. 2023, doi: 10.1186/s40537-023-00817-1.

[17] M. A. Siddiqi and W. Pak, "Optimizing Filter-Based Feature Selection Method Flow for Intrusion Detection System," *Electronics*, vol. 9, no. 12, p. 2114, Dec. 2020, doi: 10.3390/electronics9122114.

[18] S. Zhang, Y. Yuan, Z. Yao, J. Yang, X. Wang, and J. Tian, "Coronary Artery Disease Detection Model Based on Class Balancing Methods and LightGBM Algorithm," *Electronics*, vol. 11, no. 9, p. 1495, May 2022, doi: 10.3390/electronics11091495.

[19] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 273, Sep. 2024, doi: 10.1007/s10462-024-10884-2.

[20] A. Akshay, M. Katoch, N. Shekarchizadeh, and M. Abedi, "Machine Learning Made Easy (MLme): A Comprehensive Toolkit for Machine Learning-Driven Data Analysis".

[21] Y. Wang and T. Wang, "Application of Improved LightGBM Model in Blood Glucose Prediction," *Appl. Sci.*, vol. 10, no. 9, p. 3227, May 2020, doi: 10.3390/app10093227.

[22] L. Sari, A. Romadloni, R. Lityaningrum, and H. D. Hastuti, "Implementation of LightGBM and Random Forest in Potential Customer Classification," *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 43–55, Jun. 2023, doi: 10.38043/tiers.v4i1.4355.

[23] C. G. L. Pringandana, "A Comparative Analysis of Hyperparameter-Tuned XGBoost and LightGBM for Multiclass Rainfall Classification in Jakarta," vol. 6, no. 4, 2025.