Terbit pada laman web jurnal : **http://ojsamik.amikmitragama.ac.id**

JURNAL JARINGAN SISTEM INFORMASI ROBOTIK (JSR)
Vol. 10 No. 1 TAHUN 2026 E - ISSN : 2579-373X

ALGORITMA LIGHTGBM DENGAN SMOTE DAN ADASYN UNTUK PREDIKSI RISIKO SERANGAN JANTUNG

Nanda Putri Sugianto[1,*], Ade Irma Purnamasari[2], Denni Pratama[3], Puji Pramudya Marta[4],Yudhistira Arie Wijaya[5]

[1,2,3,4,5]STMIK IKMI Cirebon, Cirebon, Indonesia
Email: [1,*]nandaputri042@gmail.com, [2]irma2974@yahoo.com, [3]pratamadenni@gmail.com, [4]prammarta88@gmail.com, [5]yudhistira010471@gmail.com

Abstrak
Ketidakseimbangan data merupakan tantangan utama dalam pemodelan prediksi medis, termasuk prediksi serangan jantung, karena jumlah kasus positif jauh lebih sedikit dibandingkan kasus negatif sehingga menurunkan kemampuan model dalam mendeteksi pasien berisiko tinggi. Penelitian ini bertujuan untuk membandingkan efektivitas dua teknik *oversampling*, yaitu *Synthetic Minority Oversampling Technique (SMOTE)* dan *Adaptive Synthetic Sampling (ADASYN),* dalam meningkatkan performa algoritma *Light Gradient Boosting Machine (LightGBM)* untuk prediksi risiko serangan jantung. Dataset berjumlah 1.319 sampel dengan sembilan fitur klinis dan dianalisis melalui tahapan pra-pemrosesan, normalisasi, penanganan *class imbalance*, pembangunan model, serta evaluasi menggunakan *Accuracy, Precision, Recall, F1-Score*, dan *AUC-ROC.* Hasil menunjukkan bahwa model baseline memiliki akurasi tinggi namun sensitivitas terhadap kelas positif masih rendah. Setelah diterapkan oversampling, model mengalami peningkatan signifikan. LightGBM-SMOTE memperoleh F1-Score terbesar (0.9876) dan AUC-ROC 0.9853, sedangkan LightGBM-ADASYN mencapai F1-Score 0.9855 dan AUC-ROC 0.9861. Temuan ini menunjukkan bahwa SMOTE memberikan peningkatan performa yang lebih stabil dalam mendeteksi kelas minoritas. Dengan demikian, teknik oversampling khususnya SMOTE terbukti efektif untuk meningkatkan akurasi dan sensitivitas model prediksi serangan jantung.

Kata Kunci: LightGBM, SMOTE, ADASYN, Serangan Jantung, BuktiKetidakseimbangan Kelas.

Abstract
Imbalancedata is a major challenge in medical prediction modeling, includingheart attack prediction, because the number of positive cases is much smallerthan negative cases, thereby reducing the model's ability to detecthigh-risk patients. This study aims to comparethe effectiveness of two oversampling techniques, namely Synthetic Minority OversamplingTechnique (SMOTE) and Adaptive Synthetic Sampling (ADASYN), in improving theperformance of the Light Gradient Boosting Machine (LightGBM) algorithm for predictingthe risk of heart attack. The dataset consisted of 1,319 samples with nine clinical featuresand was analyzed through pre-processing, normalization, class imbalance handlingclass imbalance, model building, and evaluation using Accuracy,Precision, Recall, F1-Score, and AUC-ROC. The results show that the baseline modelhas high accuracy but low sensitivity to the positive class. After oversampling was applied, the model experienced a significant improvement. LightGBM-SMOTE obtained the highest F1-Score (0.9876) and AUC-ROC0.9853, while LightGBM-ADASYN achieved an F1-Score of 0.9855 and an AUC-ROC of 0.9861.These findings indicate that SMOTE provides a more stable performance improvementin detecting minority classes. Thus, the oversampling technique, especially SMOTE, is proven to be effective in improving the accuracy and sensitivityof heart attack prediction models.

Keywords: LightGBM, SMOTE, ADASYN, Heart Attack, Class Imbalance.

1. PENDAHULUAN

Perkembangan pesat dalam bidang Informatika dan Kecerdasan Buatan (AI) telah memberikan dampak besar terhadap sektor kesehatan, terutama melalui pemanfaatan machine learning (ML) untuk meningkatkan akurasi diagnosis, deteksi dini penyakit, serta mendukung pengambilan keputusan klinis yang lebih objektif dan efisien [1]. Integrasi teknologi informasi memungkinkan analisis data medis dalam skala besar dan real-time, sehingga memperkuat kemampuan organisasi kesehatan dalam memprediksi risiko dan merancang intervensi

yang tepat [2]. Berbagai kemajuan tersebut berkontribusi pada berkembangnya *health informatics*, yang memfokuskan pemanfaatan algoritma ML dalam mengolah data klinis secara cerdas dan terukur, termasuk untuk prediksi penyakit kardiovaskular seperti serangan jantung yang masih menjadi penyebab utama kematian global [3].

Algoritma modern seperti Light Gradient Boosting Machine (LightGBM) mulai banyak digunakan dalam prediksi medis karena kemampuannya mengolah data tabular kompleks secara efisien, performa klasifikasi yang tinggi, serta fleksibilitas dalam menangani interaksi non-linier antarvariabel [4]. Studi-studi terbaru menunjukkan bahwa LightGBM mampu memberikan hasil prediksi kardiovaskular yang kompetitif bahkan bila dibandingkan model-model ensemble lainnya seperti XGBoost dan CatBoost [5]. Namun, penerapan model prediktif dalam domain klinis masih menghadapi tantangan signifikan terkait heterogenitas fitur medis, seperti variasi biomarker CK-MB, troponin, tekanan darah, dan faktor demografis yang memerlukan prapemrosesan ketat agar model dapat menangkap pola dengan benar [6].

Masalah yang lebih dominan dalam pemodelan risiko serangan jantung adalah ketidakseimbangan data, di mana jumlah kasus positif jauh lebih sedikit dibandingkan kasus negatif, sehingga model sering kali bias terhadap kelas mayoritas dan gagal mendeteksi pasien berisiko tinggi [7]. Untuk mengatasi hal tersebut, teknik resampling seperti Synthetic Minority Oversampling Technique (SMOTE) dan Adaptive Synthetic Sampling (ADASYN) banyak digunakan dalam domain medis karena terbukti dapat meningkatkan sensitivitas dan F1-score model, meskipun efektivitasnya tergantung pada karakteristik dataset [9].Penelitian terbaru menyoroti bahwa meskipun SMOTE dan ADASYN sama-sama meningkatkan representasi kelas minoritas, kedua metode tersebut memiliki perilaku berbeda dalam mengatasi kompleksitas data klinis. SMOTE menghasilkan sampel sintetis berbasis interpolasi yang cenderung stabil [10], sedangkan ADASYN lebih adaptif namun berpotensi menambah noise pada area keputusan yang sulit [11]. Hal ini menegaskan perlunya evaluasi komparatif yang ketat pada model boosting modern khususnya LightGBM karena hingga kini belum banyak penelitian yang menguji kedua teknik tersebut secara langsung dalam konteks prediksi serangan jantung [12].

Berdasarkan gap penelitian tersebut, penelitian ini bertujuan untuk menganalisis dan membandingkan efektivitas teknik oversampling SMOTE dan ADASYN dalam meningkatkan performa model LightGBM menggunakan dataset medis publik. Fokus utama penelitian ini adalah mengevaluasi perubahan metrik penting seperti sensitivitas, F1-score, dan AUC-ROC antara kondisi baseline dan dua skenario oversampling [13]. Hasil penelitian diharapkan dapat memberikan kontribusi empiris terhadap pengembangan model prediksi medis yang lebih akurat dan responsif terhadap kelas minoritas, sekaligus memberikan pedoman praktis bagi peneliti dan praktisi dalam merancang pipeline pembelajaran mesin yang lebih adil dan andal untuk prediksi risiko serangan jantung.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini merupakan penelitian kuantitatif eksperimental yang menggunakan pendekatan eksperimental komputasional berbasis machine learning untuk menguji performa algoritma Light Gradient Boosting Machine (LightGBM) dalam prediksi risiko serangan jantung. Melalui tahapan penelitian, eksperimen digunakan untuk menguji dan membandingkan hasil dari dua teknik untuk peningkatan model terhadap objek penelitian. Tahapan dalam penelitian tersaji pada Gambar 1.

Gambar 1. Tahapan penelitian

2.1.1 Pengumpulan Data dan Eksplorasi Data Awal

Tahap awal penelitian ini dimulai dengan pengumpulan dan eksplorasi data menggunakan dataset *Medicaldataset.csv* yang berisi variabel klinis terkait kondisi kesehatan jantung pasien yang diambil pada data public Kaggle. Proses eksplorasi data awal (Exploratory Data Analysis/EDA) dilakukan untuk memahami struktur dan tipe variabel, menganalisis distribusi nilai, serta mengidentifikasi ketidakseimbangan kelas pada variabel target. Selain itu, dilakukan pula pemeriksaan terhadap data hilang, deteksi *outlier*, dan analisis deskriptif dasar untuk memastikan kualitas serta kesiapan dataset sebelum digunakan dalam tahap pemodelan *machine learning*.

2.1.2 Pra-Pemprosesan Data

Setelah tahap eksplorasi data, dilakukan pra-pemrosesan untuk menyiapkan dataset sebelum pemodelan. Proses ini meliputi pembersihan data (menangani nilai hilang dan inkonsistensi), pengkodean variabel kategorikal menjadi numerik, serta normalisasi jika diperlukan. Selanjutnya, data dibagi menggunakan metode *train-test split* dengan proporsi 70% data latih dan 30% data uji serta *random state* untuk memastikan replikasi hasil. Tahap ini bertujuan memastikan data siap digunakan dalam pelatihan dan evaluasi model secara optimal.

2.1.3 Penanganan ketidakseimbangan

Data set umumnya mengalami ketidakseimbangan kelas, di mana jumlah data kelas mayoritas lebih dominan dibandingkan kelas minoritas. Untuk mengatasi hal tersebut, penelitian ini menerapkan teknik oversampling *Synthetic Minority Oversampling Technique* (SMOTE) dan *Adaptive Synthetic Sampling* (ADASYN) pada data pelatihan. SMOTE menghasilkan data sintetis melalui interpolasi pada kelas minoritas, sedangkan ADASYN menambahkan sampel pada area yang sulit diklasifikasikan. Hasilnya, diperoleh dua dataset pelatihan baru yang digunakan untuk melatih model Light Gradient Boosting Machine (LightGBM) dan dibandingkan dengan model baseline tanpa penyeimbangan kelas, guna meningkatkan akurasi prediksi pada kelas minoritas.

2.1.4 Pembangunan Model

Pembangunan model dalam penelitian ini dilakukan melalui tiga skenario eksperimen yang dirancang untuk mengevaluasi pengaruh penyeimbangan data terhadap kinerja algoritma LightGBM. Skenario pertama menggunakan data pelatihan asli tanpa perlakuan penyeimbangan, sementara skenario kedua dan ketiga masing-masing menerapkan teknik SMOTE dan ADASYN untuk mengatasi ketidakseimbangan kelas.

Pada setiap skenario, proses hyperparameter tuning dilakukan secara sistematis untuk mengoptimalkan kinerja model. Langkah ini bertujuan tidak hanya untuk meningkatkan akurasi prediksi, tetapi juga untuk memastikan stabilitas model dan efisiensi proses pelatihan. Dengan pendekatan ini, setiap model yang dihasilkan dapat dibandingkan secara adil dan mencerminkan standar metodologis dalam penelitian machine learning di bidang kesehatan.

2.1.5 Evaluasi Model

Tahap evaluasi model dalam penelitian ini dilakukan dengan menggunakan dua kategori metrik, yaitu metrik berbasis prediksi kelas yang terdiri atas *precision*, *recall*, dan *F1-score*, serta metrik berbasis probabilitas berupa *Area Under the Receiver Operating Characteristic Curve* (AUC-ROC). Metrik AUC-ROC tidak dihasilkan secara otomatis oleh fungsi *classification_report* pada pustaka Scikit-Learn, sehingga perhitungannya dilakukan secara terpisah menggunakan fungsi *roc_auc_score* yang memanfaatkan probabilitas prediksi (*y_pred_proba*) sebagai input. Perhitungan AUC-ROC didasarkan pada hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai nilai ambang batas (*threshold*), dengan rumus integral numerik menggunakan pendekatan *trapezoidal rule*.

$$AUC = \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \times \frac{TPR_i + TPR_{i+1}}{2}$$

Dengan:

FPR= False Positive Rate pada titik ke-i
TPR= True Positive Rate pada titik ke-i

Nilai AUC-ROC yang diperoleh dimasukkan ke dalam tabel hasil untuk membandingkan performa model pada tiga skenario, yaitu *baseline*, SMOTE, dan ADASYN. Variasi nilai AUC antar skenario mencerminkan perbedaan distribusi probabilitas yang dihasilkan masing-masing model, di mana nilai AUC yang mendekati 1 menunjukkan kemampuan model yang lebih baik dalam membedakan kelas positif dan negatif secara akurat, sesuai dengan prinsip evaluasi model prediktif berbasis *machine learning* dalam penelitian medis.

2.1.6 Analisis Hasil

Analisis hasil penelitian ini dilakukan secara komparatif untuk mengevaluasi efektivitas tiga skenario model *Light Gradient Boosting Machine* (LightGBM) dalam menangani ketidakseimbangan data pada prediksi risiko serangan jantung. Ketiga skenario meliputi model baseline tanpa *oversampling*, serta model dengan penerapan *Synthetic Minority Oversampling Technique* (SMOTE) dan *Adaptive Synthetic Sampling* (ADASYN). Evaluasi menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, *AUC-ROC*, dan *confusion matrix* menunjukkan bahwa penerapan SMOTE dan ADASYN mampu meningkatkan sensitivitas model terhadap kelas minoritas, sehingga memperbaiki kemampuan prediksi kasus positif secara lebih akurat dan seimbang.

3. HASIL DAN PEMBAHASAN

3.1 Eksplorasi Data Awal (EDA)

Tahap pertama sebelum melakukan permodelan perlu dilakukan tahap eksplorasi data Awal atau EDA . Ini dilakukan untuk memahami struktur dan tipe variabel, menganalisis distribusi nilai, serta mengidentifikasi ketidakseimbangan kelas pada variabel target. Selain itu, dilakukan pula pemeriksaan terhadap data hilang, deteksi nilai, *outlier*, dan analisis deskriptif dasar untuk memastikan kualitas serta kesiapan dataset.

```
RangeIndex: 1319 entries, 0 to 1318
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1319 non-null   int64
 1   Gender                    1319 non-null   int64
 2   Heart rate                1319 non-null   int64
 3   Systolic blood pressure   1319 non-null   int64
 4   Diastolic blood pressure  1319 non-null   int64
 5   Blood sugar               1319 non-null   float64
 6   CK-MB                     1319 non-null   float64
 7   Troponin                  1319 non-null   float64
 8   Result                    1319 non-null   object
dtypes: float64(3), int64(5), object(1)
memory usage: 92.9+ KB
```

Gambar 2. Analisis Data

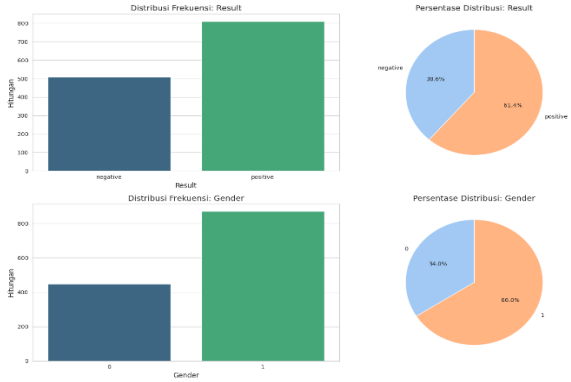
Gambar 2 menunjukkan ringkasan struktur dataset yang terdiri atas 1.319 entri dan 9 variabel, seluruhnya tanpa nilai hilang (*non-null*). Delapan variabel bertipe numerik (integer dan float), yaitu *Age*, *Gender*, *Heart rate*, *Systolic blood pressure*, *Diastolic blood pressure*, *Blood sugar*, *CK-MB*, dan *Troponin*. Sementara itu, variabel *Result* bertipe *object* karena masih berupa data kategorikal sebelum dilakukan proses encoding. Informasi ini menegaskan bahwa dataset berada dalam kondisi bersih dan siap untuk tahap pra-pemrosesan serta pembangunan model prediksi.

```
D. Statistik Deskriptif (df.describe()):
                          count        mean         std     min      25%  \
Age                      1319.0   56.191812   13.647315  14.000   47.000
Gender                   1319.0    0.659591    0.474027   0.000    0.000
Heart rate               1319.0   78.336619   51.630270  20.000   64.000
Systolic blood pressure  1319.0  127.170584   26.122720  42.000  110.000
Diastolic blood pressure 1319.0   72.269143   14.033924  38.000   62.000
Blood sugar              1319.0  146.634344   74.923045  35.000   98.000
CK-MB                    1319.0   15.274306   46.327083   0.321    1.655
Troponin                 1319.0    0.360942    1.154568   0.001    0.006

                            50%       75%     max
Age                      58.000   65.0000   103.0
Gender                    1.000    1.0000     1.0
Heart rate               74.000   85.0000  1111.0
Systolic blood pressure  124.000  143.0000   223.0
Diastolic blood pressure  72.000   81.0000   154.0
Blood sugar              116.000  169.5000   541.0
CK-MB                      2.850    5.8050   300.0
Troponin                   0.014    0.0855    10.3
```

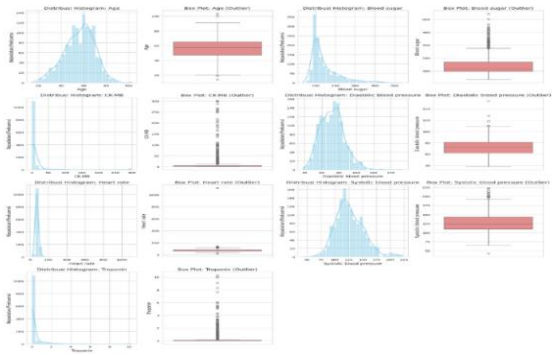
Gambar 1. Analisis Statistik Deskriptif

Gambar 3 merupakan statistik deskriptif yang menunjukkan karakteristik dasar dari delapan variabel numerik pada dataset yang terdiri dari 1.319 sampel. Rata-rata usia pasien adalah 56,19 tahun dengan rentang 14–103 tahun. Detak jantung memiliki nilai rata-rata 78,33 bpm tetapi menampilkan nilai maksimum 1.111 yang mengindikasikan keberadaan outlier. Tekanan darah sistolik dan diastolik masing-masing memiliki nilai rata-rata 127 mmHg dan 80 mmHg, sesuai rentang fisiologis umum. Variabel *Blood sugar*, CK-MB, dan Troponin juga menunjukkan variasi yang cukup lebar, mencerminkan heterogenitas kondisi klinis pasien. Secara keseluruhan, tabel ini memberikan gambaran distribusi awal data sebagai dasar untuk proses eksplorasi dan pra-pemrosesan selanjutnya.



Gambar 2. Variabel Kategorikal

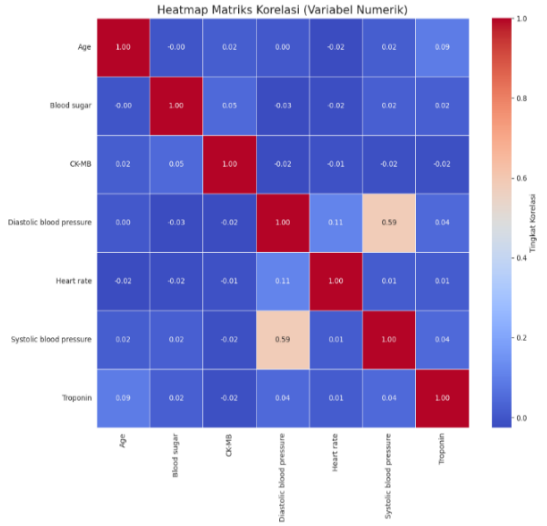
Gambar 4 merupakan Distribusi kelas *Result* menunjukkan ketidakseimbangan data yang cukup jelas, dengan 61,4% pasien berada pada kategori *positive* dan 38,6% pada kategori *negative*, sehingga diperlukan penanganan imbalance agar model tidak bias terhadap kelas mayoritas. Distribusi *Gender* juga memperlihatkan bahwa 66% pasien adalah laki-laki dan 34% perempuan, selaras dengan literatur yang menyebutkan risiko kardiovaskular lebih tinggi pada laki-laki. Temuan ini memberikan gambaran awal mengenai karakteristik populasi dan penting untuk diperhatikan dalam proses pemodelan agar hasil prediksi tetap representatif.



Gambar 3. Variabel Numerik

Gambar 5 menunjukkan distribusi variabel numerik melalui histogram dan boxplot, yang mengungkap pola *right-skewed* pada variabel Blood Sugar, CK-MB, Heart Rate, dan Troponin, disertai banyak outlier bernilai tinggi. Heart Rate bahkan

menampilkan outlier ekstrem yang mengindikasikan kemungkinan kesalahan pencatatan. Sementara itu, Age, Systolic, dan Diastolic Blood Pressure memiliki distribusi yang lebih stabil meskipun tetap mengandung outlier. Secara keseluruhan, visualisasi ini menegaskan perlunya penanganan outlier dan normalisasi data sebagai tahap penting dalam preprocessing sebelum pemodelan dilakukan.



Gambar 4. Heatmap Korelasi

Pada Gambar 6 menunjukkan bahwa seluruh variabel numerik memiliki hubungan linear yang rendah, sehingga tidak terdapat indikasi multikolinearitas yang signifikan dalam dataset. Satu-satunya hubungan yang tampak menonjol adalah korelasi sedang antara *Systolic Blood Pressure* dan *Diastolic Blood Pressure* (≈0,59), yang secara klinis wajar karena keduanya merupakan komponen tekanan darah. Variabel lainnya seperti *Age*, *Blood Sugar*, *CK-MB*, *Heart Rate*, dan *Troponin*—memiliki korelasi mendekati nol, menandakan bahwa masing-masing fitur membawa informasi yang unik. Kondisi ini menguntungkan bagi pemodelan LightGBM karena tidak ada fitur yang saling tumpang tindih secara signifikan, sehingga seluruh variabel layak digunakan sebagai prediktor tanpa risiko distorsi akibat korelasi tinggi.

3.2 Pra-Pemprosesan Data

Pada tahap pra-pemrosesan, nilai outlier pada variabel *Heart Rate* (1111 bpm) diidentifikasi sebagai kesalahan input dan diganti dengan nilai median sehingga distribusi kembali berada pada rentang fisiologis normal. Variabel target *Result* kemudian dikonversi dari kategori "positive" dan "negative" menjadi numerik biner (1 dan 0), yang menunjukkan adanya ketidakseimbangan kelas sehingga perlu diterapkan teknik penyeimbangan pada tahap modeling. Selanjutnya, data dibagi menjadi *training set* (70%) dan *testing set* (30%) menggunakan stratifikasi berdasarkan kelas target. Seluruh fitur numerik (Age, Heart Rate, Systolic/Diastolic Blood Pressure, Blood Sugar, CK-MB, dan Troponin) distandardisasi dengan *StandardScaler* yang di-fit hanya pada data latih dan diaplikasikan pada data uji untuk mencegah *data leakage*, sementara fitur kategorikal (Gender) dibiarkan apa adanya. Proses ini menghasilkan skala fitur yang seragam serta mengungkap adanya observasi dengan pola nilai identik yang mengindikasikan potensi duplikasi data, sekaligus memperlihatkan variasi klinis yang relevan untuk pembangunan model prediksi.

Tabel 1. Pembagian Data

Dataset	Jumlah Sampel (Baris)	Jumlah Fitur (Kolom)
X_train (Fitur Latih)	78	8
X_test (Fitur Uji)	34	8
y_train (Target Latih)	78	1
y_test (Target Uji)	34	1

3.3 Tahap Penanganan Ketidakseimbangan Data (*Resampling Stage*)

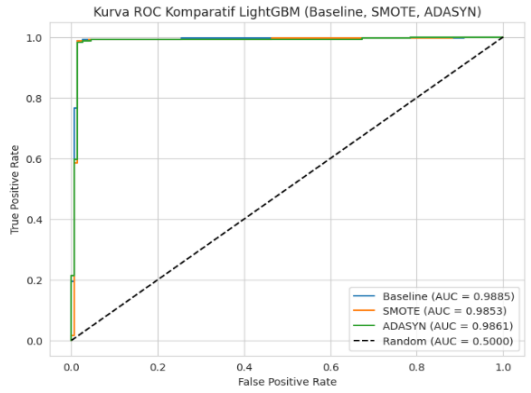
Setelah diketahui bahwa distribusi kelas pada variabel target tidak seimbang, dilakukan proses oversampling pada data latih menggunakan dua pendekatan sintetis untuk meningkatkan proporsi kelas minoritas agar model tidak bias terhadap kelas mayoritas. Teknik pertama, SMOTE (*Synthetic Minority Oversampling Technique*), menghasilkan sampel sintetis melalui interpolasi antar data minoritas sehingga jumlah kelas menjadi seimbang. Teknik kedua, ADASYN (*Adaptive Synthetic Sampling*), juga membangkitkan data sintetis namun bersifat adaptif, dengan menambah lebih banyak sampel pada area yang sulit diklasifikasikan sehingga memperbaiki sensitivitas model terhadap kasus minoritas. Hasil kedua metode menunjukkan peningkatan signifikan pada proporsi kelas, di mana SMOTE menghasilkan distribusi seimbang 0:567 dan 1:567, sedangkan ADASYN menghasilkan 0:609 dan 1:567. Proses ini memastikan bahwa model memperoleh representasi data yang memadai dari kedua kelas sehingga meningkatkan kemampuan prediksi terhadap kelas minoritas.

Tabel 2. Hasil Resampling

Metode	Metode Resampling	Distribusi Setelah Resampling
(**SMOTE)**	Synthetic Minority Oversampling Technique	0: 567, 1: 567
(ADASYN)	Adaptive Synthetic Sampling	0: 609, 1: 567

3.4 Tahap Pembangunan Model (*Model Building*)

Tiga skenario eksperimen diterapkan untuk mengevaluasi kinerja model LightGBM. Pada skenario baseline, model dilatih menggunakan data asli yang tidak seimbang dan menghasilkan performa awal yang tinggi. Skenario kedua menggunakan SMOTE untuk menyeimbangkan kelas melalui interpolasi data minoritas, yang meningkatkan sensitivitas dan akurasi model. Skenario ketiga menggunakan ADASYN, yang secara adaptif menambah sampel pada area sulit sehingga representasi kelas minoritas lebih baik, meskipun performanya sedikit di bawah SMOTE. Hasil keseluruhan menunjukkan bahwa teknik oversampling mampu meningkatkan stabilitas dan efektivitas model dibandingkan penggunaan data original.

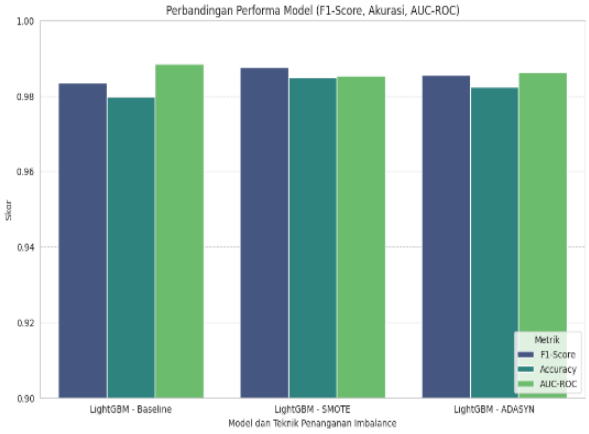


Gambar 5. Kurva ROC

Gambar 7 ROC menampilkan perbandingan performa diskriminasi antara tiga model LightGBM pada skenario baseline, SMOTE, dan ADASYN. Kurva ROC menunjukkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) di berbagai ambang prediksi. Semua model memiliki kurva yang sangat dekat dengan titik ideal (TPR = 1; FPR = 0), menunjukkan bahwa model mampu mengenali kasus positif dengan tingkat kesalahan yang sangat rendah. Nilai AUC untuk ketiga model berada di atas 0.98, yakni Baseline (0.9885), SMOTE (0.9853), dan ADASYN (0.9861), mencerminkan performa diskriminasi yang hampir sempurna. Garis diagonal hitam melambangkan model acak (AUC = 0.5). Perbandingan ini menunjukkan bahwa meskipun baseline memiliki AUC paling tinggi, model yang diterapkan dengan teknik oversampling juga menunjukkan performa yang sangat kompetitif, bahkan lebih stabil dalam mendeteksi kelas minoritas.

3.5 Analisis Hasil

Hasil analisis menunjukkan bahwa teknik oversampling memberikan peningkatan performa dibandingkan model baseline. SMOTE menghasilkan keseimbangan prediksi terbaik dengan F1-Score tertinggi, sementara ADASYN tetap meningkatkan kinerja namun sedikit di bawah SMOTE karena variasi sintetis yang lebih agresif. Secara keseluruhan, SMOTE terbukti sebagai metode paling efektif dalam menangani ketidakseimbangan data tanpa menurunkan kemampuan diskriminasi model.



Gambar 6. Perbandingan Model

3.6 Pembahasan

Hasil penelitian menunjukkan bahwa penanganan ketidakseimbangan kelas berperan krusial dalam meningkatkan kinerja LightGBM untuk prediksi risiko serangan jantung. Meskipun model baseline menghasilkan akurasi dan AUC yang tinggi, bias terhadap kelas mayoritas menyebabkan sensitivitas terhadap kelas positif belum optimal, sebagaimana umum terjadi pada dataset klinis yang tidak seimbang.

Penerapan SMOTE dan ADASYN terbukti meningkatkan performa model, khususnya pada metrik recall dan F1-score, sejalan dengan temuan studi sebelumnya. Di antara keduanya, SMOTE memberikan performa paling stabil dengan F1-score dan recall tertinggi, sementara ADASYN menunjukkan peningkatan yang lebih fluktuatif. Hal ini menunjukkan bahwa SMOTE mampu menghasilkan representasi kelas minoritas yang lebih konsisten, sedangkan ADASYN cenderung menambah variasi yang berpotensi meningkatkan noise.

Secara keseluruhan, meskipun baseline memiliki kemampuan diskriminasi probabilitas yang baik (AUC tinggi), penyeimbangan kelas tetap diperlukan untuk meningkatkan sensitivitas model. Temuan ini menegaskan pentingnya teknik oversampling dalam pengembangan model prediksi kardiovaskular yang lebih akurat dan responsif terhadap pasien berisiko tinggi.

4. KESIMPULAN

Berdasarkan proses analisis Bab ini menyajikan kesimpulan akhir dari penelitian yang telah dilakukan.Pada skenario baseline tanpa oversampling, LightGBM menunjukkan performa tinggi dengan akurasi 97,98% dan F1-score 0,9834, namun ketidakseimbangan kelas (61,4% positif vs 38,6% negatif) menyebabkan bias terhadap kelas mayoritas, sejalan dengan temuan [15]bahwa LightGBM unggul dalam akurasi awal tetapi kurang sensitif terhadap kelas minoritas. Setelah penerapan SMOTE dan ADASYN, performa model meningkat signifikan. SMOTE menghasilkan F1-score tertinggi (0,9876), sementara ADASYN juga efektif dengan F1-score 0,9855; hasil ini konsisten dengan [16] yang melaporkan peningkatan recall dan F1-score pada dataset medis imbalanced. Perbandingan kedua teknik menunjukkan bahwa SMOTE lebih optimal dibanding ADASYN karena lebih stabil dan menghasilkan trade-off TPR-FPR terbaik pada kurva ROC, mendukung ulasan [17] bahwa SMOTE unggul dalam dataset medis berukuran sedang. Secara keseluruhan, penelitian ini menegaskan bahwa penanganan ketidakseimban gan data berperan penting dalam meningkatkan performa LightGBM dan mengurangi bias, sehingga meningkatkan reliabilitas sistem prediksi klinis berbasis AI.

5. REFERENCES

[1] E. Feigerlova, H. Hani, and E. Hothersall-Davies, “A systematic review of the impact of artificial intelligence on educational outcomes in health professions education,” *BMC Med. Educ.*, vol. 25, no. 1, 2025, doi: 10.1186/s12909-025-06719-5.

[2] Y. Cai *et al.*, “Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review,” *BMC Med.*, vol. 22, no. 1, pp. 1–18, 2024, doi: 10.1186/s12916-024-03273-7.

[3] D. I. Kasartzian and T. Tsiampalis, “Transforming Cardiovascular Risk Prediction: A Review of Machine Learning and Artificial Intelligence Innovations,” *Life*, vol. 15, no. 1, 2025, doi: 10.3390/life15010094.

[4] W. Alsabhan and A. Alfadhly, “Effectiveness of machine learning models in diagnosis of heart disease : a comparative study,” pp. 1–19, 2025.

[5] P. Shah and et al., “Predicting cardiovascular risk with hybrid ensemble approaches,” *Sci. Rep.*, 2025, [Online]. Available: https://www.nature.com/articles/s41598-025-01650-7

[6] N. A. Hussain and A. A. Mohammed, “Early Heart Attack Detection Using Hybrid Deep Learning Techniques,” *Inf.*, vol. 16, no. 5, 2025, doi: 10.3390/info16050334.

[7] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, *Handling imbalanced medical datasets: review of a decade of research*, vol. 57, no. 10. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10884-2.

[8] M. Carvalho, A. J. Pinho, and S. Brás, “Resampling approaches to handle class imbalance: a review from a data perspective,” *J. Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01119-4.

[9] A. H. Putra and A. Salam, “A Comparative Performance of SMOTE, ADASYN and Random Oversampling in Machine Learning Models on Prostate Cancer Dataset,” *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 603–610, 2025, doi: 10.30871/jaic.v9i3.9308.

[10] M. U. Rehman, S. Naseem, A. Ur, R. Butt, and T. Mahmood, “Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment,” pp. 1–15, 2025.

[11] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, “Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified SMOTE Strategies,” *JOIV Int. J. Informatics Vis.*, 2024, doi: 10.62527/joiv.8.3.2283.

[12] J. Zhu, S. Pu, J. He, D. Su, W. Cai, and X. Xu, “Processing imbalanced medical data at the data level with assisted-reproduction data as an example,” *BioData Min.*, vol. 17, no. 1, 2024, doi: 10.1186/s13040-024-00352-4.

[13] A. Masruriyah, H. Novita, C. Sukmawati, A. Ramadhan, S. Arif, and B. Dermawan, “Pengukuran Kinerja Model Klasifikasi dengan Data Oversampling pada Algoritma Supervised Learning untuk Penyakit Jantung,” *Comput. Sci.*, vol. 4, no. 1, pp. 62–70, 2024, doi: 10.31294/coscience.v4i1.2389.

[14] M. Kannan, D. Umamaheswari, B. Manimekala, I. P. S. Mary, and P. M. Savitha, “An enhancement of machine learning model performance in disease prediction with synthetic data generation,” pp. 1–21, 2025.

[15] T. O. Omotehinwa, “Optimizing the Light Gradient-Boosting Machine algorithm for an efficient early detection of coronary heart disease,” *J. / Conf. Proc.*, 2024, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949953424000122

[16] J. Zhu *et al.*, “Processing imbalanced medical data at the data level with assisted-reproduction data as an example,” *BioData Min.*, vol. 17, no. 1, 2024, doi: 10.1186/s13040-024-00384-y.

[17] M. Salmi, “Handling imbalanced medical datasets: review of a decade,” *Springer*, 2024, doi: 10.1007/s10462-024-10884-2.