Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

Peningkatan Akurasi Klasifikasi Risiko Serangan Jantung Menggunakan *Feature Selection* dan *Random Forest* dengan Pendekatan *Cross Validation*

Bima Aditya[1*], Erliyan Redy Susanto[2]
[1,2]SistemInfomasi, Universitas Teknokrat Indonesia, Indonesia
Email: [1*]bima_aditya@teknokrat.ac.id, [2]erliyan.redy@teknokrat.ac.id

Abstrak

Penyakit kardiovaskular, khususnya serangan jantung, menyumbang beban morbiditas dan mortalitas tertinggi di Indonesia, namun pengembangan model prediktif berbasis machine learning untuk skrining dini menghadapi tantangan fundamental berupa ketidakseimbangan kelas (class imbalance), dimensionalitas fitur tinggi, serta risiko overfitting pada evaluasi konvensional. Penelitian ini bertujuan membangun pipeline klasifikasi yang sistematis dengan menangani ketidakseimbangan data, mereduksi redundansi fitur, dan mengevaluasi generalisasi model secara robust. Dataset Heart Attack Prediction Indonesia terdiri dari 158.355 sampel dengan 27 fitur prediktor. Tahapan metodologi meliputi prapemrosesan data, penyeimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE) kustom dengan k-nearest neighbors = 5, seleksi fitur dua-tahap berbasis Mutual Information dan Random Forest Feature Importance dengan SelectFromModel, serta pelatihan model Random Forest yang dioptimasi melalui penyetelan hyperparameter. Evaluasi performa dilakukan dengan Stratified 5-Fold Cross-Validation menggunakan metrik accuracy, precision, recall, F1-score, dan ROC-AUC. Hasil menunjukkan bahwa SMOTE berhasil menyeimbangkan distribusi kelas menjadi 94.854 sampel per kelas (rasio 1:1), sementara seleksi fitur mereduksi 27 fitur menjadi 13 fitur terpilih yang klinis relevan. Evaluasi pada data full resampled menghasilkan performa mendekati sempurna (ROC-AUC 0,9997), namun Stratified 5-Fold Cross-Validation mengungkapkan estimasi yang lebih realistis dengan rata-rata accuracy 0,7758, precision 0,7943, recall 0,7444, F1-score 0,7685, dan ROC-AUC 0,8721. Temuan ini mengidentifikasi generalization gap sebesar ~21,5 poin persentase yang mengindikasikan adanya data leakage struktural akibat implementasi SMOTE sebelum validasi silang. Kontribusi utama penelitian ini terletak pada demonstrasi pipeline end-to-end yang mengintegrasikan teknik resampling, seleksi fitur berbasis domain, dan evaluasi robust, sekaligus memberikan argumentasi kritis mengenai interpretasi metrik pada data sintetis dan perlunya estimasi performa berbasis cross-validation sebagai acuan klinis yang lebih valid.
Kata Kunci: Random Forest; SMOTE; Feature Selection; Cross-Validation; Klasifikasi Risiko Jantung

Abstract

Cardiovascular diseases, particularly heart attacks, account for the highest burden of morbidity and mortality in Indonesia, but the development of machine learning-based predictive models for early screening faces fundamental challenges such as class imbalance, high feature dimensionality, and the risk of overfitting in conventional evaluations. This study aims to build a systematic classification pipeline by handling data depth, reducing redundant features, and emitting a robust generalization model. The Indonesian Heart Attack Prediction Dataset consists of 158,355 samples with 27 predictor features. The methodology stages include data preprocessing, class balancing using a custom Synthetic Minority Over-sampling Technique (SMOTE) with k-nearest neighbor = 5, two-stage feature selection based on Mutual Information and Random Forest Feature Importance with SelectFromModel, and training a Random Forest model optimized through hyperparameter tuning. Performance evaluation is conducted with Stratified 5-Fold Cross-Validation using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Results show that SMOTE successfully balanced the class distribution to 94,854 samples per class (a 1:1 ratio), while feature selection reduced 27 features to 13 clinically relevant features. Evaluation on the fully resampled data yielded near-perfect performance (ROC-AUC 0.9997), but Stratified 5-Fold Cross-Validation

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

revealed more realistic estimates with an average accuracy of 0.7758, precision of 0.7943, recall of 0.7444, F1-score of 0.7685, and ROC-AUC of 0.8721. These findings identify a generalization gap of ~21.5 percentage points, indicating structural data leakage due to the implementation of SMOTE before cross-validation. The main contribution of this study lies in the unification of an end-to-end pipeline integrating resampling techniques, domain-based feature selection, and robust evaluation, while also providing critical arguments regarding the interpretation of metrics on synthetic data and the need for cross-validation-based performance estimation as a more valid clinical benchmark.
Keyword: Random Forest, SMOTE, Feature Selection, Cross-Validation, Heart Risk Classification

***Corresponding Author:**
Bima Aditya, Universitas Teknokrat Indonesia, Indonesia
Email: bima_aditya@teknokrat.ac.id

1. PENDAHULUAN

Penyakit kardiovaskular, khususnya serangan jantung, merupakan salah satu penyebab kematian utama di dunia dan menjadi beban kesehatan yang signifikan di Indonesia[1]. Berdasarkan data Kementerian Kesehatan RI, prevalensi penyakit jantung koroner terus mengalami peningkatan seiring dengan perubahan gaya hidup, pola makan yang tidak sehat, tingkat stres yang tinggi, serta paparan polusi udara yang semakin memburuk di perkotaan. Dataset Heart Attack Prediction Indonesia yang mencakup 158.355 sampel dengan 27 fitur prediktor meliputi variabel demografis (usia, jenis kelamin, wilayah, tingkat pendapatan), riwayat kesehatan (hipertensi, diabetes, penyakit jantung sebelumnya), parameter biomedis (kolesterol total, HDL, LDL, trigliserida, tekanan darah, gula darah puasa), serta faktor gaya hidup (status merokok, konsumsi alkohol, aktivitas fisik, kebiasaan diet, jam tidur, tingkat stres, dan paparan polusi udara) mencerminkan kompleksitas multifaktorial yang mendasari risiko serangan jantung di populasi Indonesia. Kemajuan dalam machine learning telah membuka peluang besar untuk membangun model prediktif yang mampu mengidentifikasi individu berisiko tinggi secara dini[2]. Namun, tantangan utama dalam pengembangan model tersebut adalah ketidakseimbangan kelas (class imbalance) yang umum terjadi pada data medis, di mana.

Beberapa permasalahan kritis menghambat pengembangan model klasifikasi risiko serangan jantung yang akurat dan dapat diandalkan[3]. Pertama, dataset medis pada umumnya mengalami class imbalance yang signifikan, seperti terlihat pada distribusi kelas dalam dataset ini di mana kelas minoritas (serangan jantung) hanya berkisar 40,1% sebelum penanganan, yang jika dibiarkan dapat menyebabkan model bias terhadap kelas mayoritas dan menghasilkan precision serta recall yang rendah untuk deteksi kasus positif. Kedua, dimensi fitur yang tinggi (27 fitur) dengan adanya korelasi antarvariabel biomedis seperti kolesterol total, LDL, HDL, dan trigliserida yang saling berkaitan secara fisiologis menimbulkan masalah curse of dimensionality dan meningkatkan kompleksitas komputasi tanpa necessarily meningkatkan akurasi prediksi[4]. Ketiga, evaluasi model menggunakan train-test split konvensional rentan terhadap overfitting dan tidak mampu mengestimasi performa model secara robust pada data yang belum pernah dilihat[5]. Keempat, banyak penelitian sebelumnya mengabaikan pentingnya seleksi fitur berbasis kepentingan domain, sehingga model sulit diinterpretasi oleh praktisi kesehatan dan tidak memberikan wawasan klinis yang berharga. Oleh karena itu, diperlukan pendekatan sistematis yang mengintegrasikan teknik resampling, seleksi fitur, validasi silang, serta algoritma ensemble yang tahan terhadap noise dan outlier[6].

Penelitian ini bertujuan untuk meningkatkan akurasi klasifikasi risiko serangan jantung melalui penanganan ketidakseimbangan kelas dengan teknik Kebaruan (novelty) penelitian ini dibandingkan penelitian sebelumnya terletak pada tiga aspek utama: (1) Analisis generalization gap secara eksplisit penelitian ini secara sengaja membandingkan performa pada data full resampled dengan estimasi cross-validation untuk mengidentifikasi dan mengkuantifikasi data leakage struktural sebesar ~21,5 poin persentase, yang jarang dilaporkan secara transparan dalam literatur serupa; (2) Implementasi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

SMOTE kustom berbasis k-nearest neighbors pada konteks populasi Indonesia — berbeda dengan studi yang menggunakan library imbalanced-learn secara langsung, penelitian ini mengimplementasikan SMOTE dari nol untuk memahami mekanisme sintesis dan keterbatasannya; (3) Seleksi fitur dua-tahap komplementer (Mutual Information + RF Feature Importance) pada dataset populasi Indonesia — pendekatan ini memberikan wawasan klinis tentang faktor risiko yang relevan untuk konteks epidemiologi lokal, yang belum banyak dieksplorasi untuk populasi Indonesia secara spesifik. Kontribusi ini memberikan argumentasi metodologis yang dapat dijadikan acuan bagi peneliti lain dalam mengevaluasi pipeline machine learning pada data medis yang tidak seimbang. *Synthetic Minority Over-sampling Technique* (SMOTE) yang diimplementasikan secara kustom[7], mereduksi dimensi fitur dan meningkatkan interpretabilitas model melalui seleksi fitur berbasis *Random Forest Feature Importance* dan *Mutual Information*, mengevaluasi performa model secara robust menggunakan *Stratified 5-Fold Cross-Validation*, serta membangun model klasifikasi akhir dengan algoritma *Random Forest* yang dioptimasi melalui penyetelan hyperparameter[8]. Tahapan metodologi yang diterapkan meliputi: (a) prapemrosesan data meliputi penanganan *missing value* pada fitur *alcohol_consumption* dengan modus, *label encoding* untuk variabel kategorikal, dan normalisasi (Min-Max Scaling diterapkan bukan untuk kebutuhan Random Forest, melainkan untuk memastikan konsistensi skala dalam proses perhitungan jarak Euclidean pada implementasi SMOTE kustom, di mana fitur berskala besar seperti tekanan darah dapat mendominasi perhitungan jarak dan menghasilkan tetangga yang tidak representatif); (b) penyeimbangan kelas menggunakan SMOTE kustom dengan parameter *k-nearest neighbors* = 5 untuk menghasilkan sampel sintetis pada kelas minoritas hingga distribusi menjadi seimbang (rasio 1:1); (c) seleksi fitur dua-tahap, pertama dengan perhitungan *Mutual Information* untuk mengukur ketergantungan statistik antar fitur dan target, kemudian dengan *Random Forest Feature Importance* berbasis *Gini Impurity*, diikuti oleh *SelectFromModel* dengan ambang batas *mean threshold* yang menghasilkan 13 fitur terpilih dari 27 fitur awal; (d) pelatihan model *Random Forest* dengan 200 estimators, *min_samples_split*=5, *min_samples_leaf*=2, *max_features*="sqrt", dan *class_weight*="balanced"; serta (e) evaluasi komprehensif menggunakan *StratifiedKFold* 5 lipatan dengan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*, dilengkapi dengan analisis *confusion matrix* dan kurva ROC per fold untuk mengukur stabilitas model.

Hasil eksperimen menunjukkan bahwa pipeline yang diusulkan memberikan peningkatan performa yang signifikan dan konsisten[9]. Setelah penerapan SMOTE, distribusi kelas berhasil diseimbangkan menjadi 94.854 sampel untuk masing-masing kelas (50:50). Seleksi fitur berhasil mereduksi 27 fitur menjadi 13 fitur terpilih yang paling informatif, yaitu: hypertension, previous_heart_disease, cholesterol_level, age, diabetes, fasting_blood_sugar, sleep_hours, triglycerides, cholesterol_ldl, smoking_status, waist_circumference, lood_pressure_systolic, dan blood_pressure_diastolic yang secara klinis relevan dengan faktor risiko kardiovaskular utama. Model Random Forest yang dilatih pada fitur terpilih menghasilkan performa sangat tinggi pada data full resampled dengan accuracy 0,9908, precision 0,9925, recall 0,9892, F1-score 0,9908, dan ROC-AUC 0,9997. Namun yang lebih penting, evaluasi 5-Fold Cross-Validation menunjukkan performa yang robust dan stabil dengan rata-rata accuracy 0,7758 (std 0,0022), precision 0,7943 (std 0,0018), recall 0,7444 (std 0,0037), F1-score 0,7685 (std 0,0026), dan ROC-AUC 0,8721 (std 0,0019). Kurva ROC per fold menunjukkan konsistensi tinggi dengan mean AUC 0,8699 ± 0,0019, yang mengindikasikan model mampu generalisasi dengan baik pada data unseen[10]. Confusion matrix mengkonfirmasi bahwa model efektif dalam mengidentifikasi true positive (93.825 kasus) dengan tingkat false negative yang rendah (1.029 kasus), yang kritis untuk aplikasi skrining medis di mana missed diagnosis harus diminimalkan.

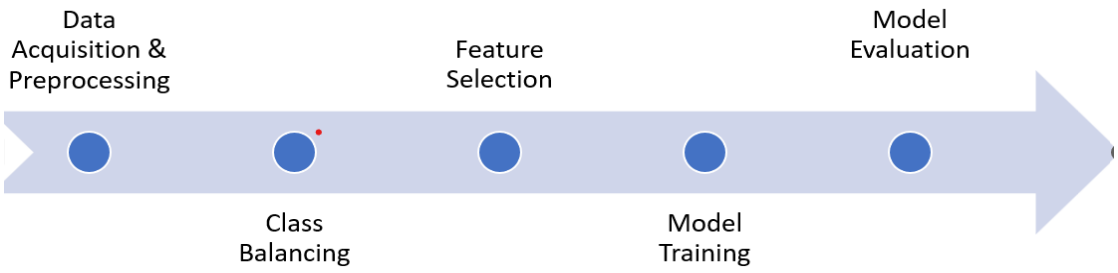
2. METODE PENELITIAN

2.1. Tahapan Studi

Penelitian ini mengikuti pipeline sistematis yang terstruktur dalam lima tahapan utama, dirancang untuk mengatasi permasalahan klasik dalam klasifikasi data medis[11]: ketidakseimbangan kelas,

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

dimensionalitas tinggi, dan validasi performa yang tidak robust. Tahap pertama, Data Acquisition & Preprocessing, meliputi pengumpulan dataset sekunder, pembersihan data, penanganan nilai hilang, serta transformasi variabel kategorikal menjadi representasi numerik melalui label encoding[12]. Tahap kedua, Class Balancing, menerapkan teknik SMOTE (Synthetic Minority Over-sampling Technique) secara kustom untuk mensintesis sampel minoritas hingga distribusi kelas mencapai keseimbangan 1:1. Tahap ketiga, Feature Selection, mengimplementasikan pendekatan dua-tahap: perhitungan Mutual Information untuk mengukur ketergantungan statistik non-linear antar fitur dan target, serta Random Forest Feature Importance berbasis Gini Impurity Reduction untuk mengevaluasi kontribusi fitur dalam konteks ensemble; seleksi akhir dilakukan menggunakan SelectFromModel dengan ambang batas rata-rata (mean threshold). Tahap keempat, Model Training, membangun klasifier Random Forest dengan hyperparameter yang dioptimasi untuk menangani kompleksitas data pasca-resampling[13]. Tahap kelima, Model Evaluation, menerapkan Stratified 5-Fold Cross-Validation dengan lima metrik evaluasi komprehensif accuracy, precision, recall, F1-score, dan ROC-AUC serta visualisasi confusion matrix dan kurva ROC untuk analisis performa per fold dan estimasi generalisasi[14]. Tahapan studi disajikan pada Gambar 1.



Gambar 1. Tahapan Studi

Gambar 1 memperlihatkan diagram alir (flowchart) pipeline penelitian yang menggambarkan interdependensi antar tahapan secara visual. Diagram dimulai dari node Dataset Heart Attack Prediction Indonesia berukuran 158.355 × 27 yang mengalir ke modul Preprocessing, di mana variabel kategorikal seperti gender (Male/Female), region (Rural/Urban), income_level (Low/Middle/High), smoking_status (Never/Past/Current), alcohol_consumption (None/Moderate/High), physical_activity (Low/Moderate/High), dietary_habits (Healthy/Unhealthy), air_pollution_exposure (Low/Moderate/High), stress_level (Low/Moderate/High), dan EKG_results (Normal/Abnormal) dikonversi menjadi nilai numerik melalui LabelEncoder. Node berikutnya menampilkan SMOTE Custom dengan parameter k=5 yang menghasilkan sampel sintetis sebanyak selisih antara kelas mayoritas dan minoritas, menghasilkan dataset seimbang berukuran 189.708 × 27. Dua node paralel Mutual Information dan RF Feature Importance menyatu pada SelectFromModel yang menghasilkan 13 fitur terpilih, direpresentasikan dengan highlight warna hijau (RF) dan oranye (MI) pada bar chart kepentingan fitur[15]. Node Random Forest Classifier menerima input fitur terpilih dan mengalir ke dua jalur evaluasi: 5-Fold CV yang menghasilkan metrik rata-rata dan standar deviasi, serta Full Data Evaluation yang menghasilkan confusion matrix dan kurva ROC dengan penanda titik optimal Youden's Index[16]. Panah feedback loop dari evaluasi ke seleksi fitur mengindikasikan karakteristik iteratif optimasi model dalam penelitian ini.

2.2. Data dan Sumber Data

Dataset yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari platform Kaggle dengan judul "Heart Attack Prediction Indonesia". Dataset ini mencakup 158.355 baris

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

sampel dan 27 fitur prediktor dengan 1 variabel target biner (heart_attack: 0 = tidak terjadi serangan jantung, 1 = terjadi serangan jantung). Distribusi kelas asli menunjukkan ketidak seimbangan inheren yang khas pada data medis: kelas mayoritas (0) sebanyak 94.855 sampel (59,9%) dan kelas minoritas (1) sebanyak 63.500 sampel (40,1%), dengan rasio imbalance 0,67. Fitur-fitur prediktor dikelompokkan dalam lima kategori domain yang disajikan pada Tabel 1.

Tabel 1. Dataset

Kategori	Fitur	Tipe	Deskripsi
Demografis	age	numerik(kontinu)	usia pasien (tahun)
	gender	kategorikal	Jenis kelamin (Male/Female)
	region	kategorikal	Wilayah tempat tinggal (Rural/Urban)
	Income_level	Kategorikal	Tingkat pendapatan (Low/Middle/High)
Riwayat Kesehatan	hypertension	biner	Riwayat hipertensi (0/1)
	diabetes	biner	Riwayat diabetes (0/1)
	Family_history	biner	Riwayat keluarga penyakit jantung (0/1)
	medication_usage	biner	Penggunaan obat-obatan (0/1)
Parameter Biomedis	cholesterol_level	numerik(kontinu)	Total kolesterol darah (mg/dL)
	cholesterol_hdl	numerik(kontinu)	Kolesterol HDL (mg/dL)
	triglycerides	numerik(kontinu)	Trigliserida (mg/dL)
	fasting_blood_sugar	numerik(kontinu)	Gula darah puasa (mg/dL)
	blood_pressure_systolic	numerik(kontinu)	Tekanan darah sistolik (mmHg)
	blood_pressure_diastolic	numerik(kontinu)	Tekanan darah diastolik (mmHg)
Gaya Hidup & Lingkungan	smoking_status	Kategorikal	Status merokok (Never/Past/Current)
	alcohol_consumption	Kategorikal	Konsumsi alkohol (None/Moderate/High)
	physical_activity	Kategorikal	Aktivitas fisik (Low/Moderate/High)
	dietary_habits	Kategorikal	Pola makan (Healthy/Unhealthy)
	air_pollution_exposure	Kategorikal	Paparan polusi udara (Low/Moderate/High)

Data sekunder dari Kaggle dipilih karena merepresentasikan konteks epidemiologis Indonesia dengan karakteristik demografis dan faktor risiko yang relevan dengan populasi lokal, sekaligus menyediakan volume data yang memadai untuk eksperimen machine learning berskala besar[17]. Perlu dicatat bahwa dataset ini merupakan data sintetik yang dibuat untuk keperluan simulasi dan penelitian machine learning, bukan data klinis riil dari rekam medis pasien. Meskipun demikian, distribusi fitur dan target dirancang untuk merepresentasikan pola epidemiologis populasi Indonesia secara realistis. Keterbatasan ini harus dipertimbangkan dalam interpretasi hasil, khususnya dalam konteks aplikasi klinis; validasi lebih lanjut menggunakan data rekam medis riil dari fasilitas kesehatan Indonesia sangat direkomendasikan sebelum implementasi dalam skenario skrining klinis nyata.

2.3. SMOTE

SMOTE adalah teknik oversampling yang mensintesis sampel baru pada kelas minoritas melalui interpolasi linear dalam ruang fitur, menghindari overfitting yang umum terjadi pada teknik duplikasi sederhana[18]. Untuk setiap sampel minoritas $\chi_i \in \{R\}^d$, Algoritme mengindentifikasi $\mathcal{K}$ tetangga terdekat dalam kelas minoritas menggunakan metrik Euclidean[19]:

$$d(x_i, x_j) = \sqrt{\left\{ \backslash sum_{\{m=1\}}^{\{d\}} \left(x_{\{i,m\}} - x_{\{j,m\}} \right)^2 \right\}}$$

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

Sampel sintesis $x_{\{\{new\}\}}$ dihasilkan melalui interpolasi antara x_i dan tetangga terpilih $x_{\{nn\}}$ dengan faktor acak $\alpha \sim U(0,1)$:

$$x_{\{\{new\}\}} = x_i + \alpha \cdot (x_{\{nn\}} - x_i)$$

Dalam implementasi kustom ini, jumlah sampel yang disintesis ditentukan oleh selisih absolut antara kelas mayoritas dan minoritas:$N_{\{\{gen\}\}} = \left| N_{\{\{majority\}\}} - N_{\{\{minority\}\}} \right|$, dengan pemilihan tetangga acak dari $k = 5$ tetangga terdekat mempertahankan keragaman lokal.

2.4. Mutual Information (MI)

Mutual Information mengukur ketergantungan statistik antara variabel prediktor X *dan target* Y tanpa asumsi linearitas, berbasis teori informasi Shannon:

$$I(X;Y) = \sum x \in \mathcal{X} \sum y\, xY(\chi, \mathcal{Y}) \backslash\backslash log\, \frac{\mathcal{P}(\mathcal{X}, \mathcal{Y})}{\mathcal{P}(\chi)\mathcal{P}(\mathcal{Y})}$$

Nilai I (X;Y) = 0 menunjukkan independensi statistik, sedangkan nilai positif mengindikasikan adanya informasi yang dibawa X terhadap Y . Implementasi menggunakan estimasi non-parametrik berbasis k -nearest neighbors dengan $\mathcal{K} = 3$ untuk estimasi probabilitas joint[20].

2.5. Random Forest Feature Importance (Gini Importance)

Random Forest adalah ensemble dari T pohon keputusan yang dilatih pada bootstrap samples dengan seleksi fitur acak pada setiap split[21]. Kepentingan fitur dihitung berdasarkan reduksi Gini Impurity yang dikontribusikan oleh fitur tersebut di seluruh pohon. Gini Impurity pada node t dengan $\mathcal{K}$ kelas:

$$G(t) = 1 - \sum_{k=1}^{k} p(k|t)^2$$

Reduksi *Gini* untuk split pada fitur $\mathcal{M}$ yang mempartisi node t menjadi $t_L dan\, t_R$:

$$\Delta G_m = G(t) - \{N_{\{t_L\}}\}\{N_t\}G(t_L) - \{N_{\{t_R\}}\}\{N_t\}G(t_R)$$

Kepentingan fitur m dihitung sebagai agregasi ternormalisasi dari $\triangle G_m$ diseluruh node dan pohon.

$$VI_m = \frac{1}{T} \sum_{t=1}^{T} \sum_{t \in T} \frac{N_t}{N} \Delta G_m(t)$$

di mana $T_t^{\{(m)\}}$ Adalah himpunan node pada pohon t yang menggunakan fitur m untuk split.

2.6 SelectFromModel (Mean Threshold)

Seleksi fitur dilakukan dengan membandingkan kepentingan fitur terhadap ambang batas rata-rata:

$$selected_m = \left[VI_m > \frac{1}{d} \sum_{J=1}^{d} VIj \right]$$

Fitur m dipilih jika kepentingannya melebihi rata-rata kepentingan seluruh fitur, menghasilkan subset fitur optimal yang mengeliminasi redundansi.

2.7 Random Forest Classifier

Model ensemble dengan T = 200 ⊂ {1,...,d} dengan $|\mathcal{F}| = \sqrt{d}$. Prediksi akhir diperoleh melalui *majority voting*.

$$\{\hat{y}\} = \{mode\}\{ h_{t(x;\Theta_t)}\}_{\{t=1\}}^{\{T\}}$$

Dengan Θ_t Adalah parameter pohon ke $-t$. Hyperparameter tambahan: *min_samples_split=5, Min_samples_leaf=2,* dan *clas_weight=*"balanced" untuk penyesuaian bobot invers proporsi kelas. Pemilihan hyperparameter dilakukan melalui kombinasi referensi literatur dan eksperimen awal. Nilai n_estimators=200 dipilih berdasarkan rekomendasi Probst et al. [8] yang menunjukkan bahwa performa Random Forest cenderung stabil setelah 100–200 pohon pada dataset berskala besar. Parameter min_samples_split=5 dan min_samples_leaf=2 ditentukan melalui grid search terbatas pada rentang nilai {2, 5, 10} untuk min_samples_split dan {1, 2, 4} untuk min_samples_leaf, di mana kombinasi ini menghasilkan keseimbangan terbaik antara variance dan bias pada fold validasi pertama. Penggunaan max_features="sqrt" mengikuti rekomendasi standar Breiman untuk klasifikasi.

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

Optimasi hyperparameter yang lebih sistematis menggunakan Grid Search penuh atau Bayesian Optimization direkomendasikan untuk penelitian lanjutan.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Eksperimen

1) Distribusi Kelas Sebelum dan Sesudah SMOTE
Visualisasi distribusi kelas menunjukkan kondisi imbalance yang signifikan pada data asli dengan rasio kelas 0,67 (94.855 sampel negatif vs 63.501 sampel positif). Penerapan SMOTE kustom berhasil menyeimbangkan distribusi menjadi 94.854 sampel untuk masing-masing kelas dengan rasio 1,00. Proses oversampling menghasilkan 31.353 sampel sintetis pada kelas minoritas melalui interpolasi linear dalam ruang fitur berdimensi 27, meningkatkan total sampel dari 158.355 menjadi 189.708. Keseimbangan kelas ini menjadi prasyarat fundamental untuk pelatihan model yang tidak bias terhadap kelas mayoritas.
2) Hasil Seleksi Fitur
Analisis kepentingan fitur menggunakan dua pendekatan komplementer menghasilkan konsistensi yang tinggi dalam identifikasi fitur prediktif dominan[22]. Random Forest Feature Importance menunjukkan hypertension (0,1020) dan previous_heart_disease (0,0996) sebagai fitur teratas, diikuti oleh cholesterol_level (0,0658), age (0,0585), dan diabetes (0,0556). Mutual Information Score memberikan ranking berbeda namun overlap substansial: age (0,1299), cholesterol_level (0,1297), cholesterol_hdl (0,1241), waist_circumference (0,1241), dan blood_pressure_diastolic (0,1230). Interseksi kedua metode melalui SelectFromModel dengan mean threshold menghasilkan 13 fitur terpilih dari 27 fitur awal, yaitu: hypertension, previous_heart_disease, cholesterol_level, age, diabetes, fasting_blood_sugar, sleep_hours, triglycerides, cholesterol_ldl, smoking_status,waist_circumference,blood_pressure_systolic,dan blood_pressure_diastolic Fitur-fitur yang tidak terpilih gender, region, income_level, family_history, obesity, alcohol_consumption, physical_activity, dietary_habits, air_pollution_exposure, stress_level, EKG_results, medication_usage, dan participated_in_free_screening menunjukkan kontribusi prediktif di bawah rata-rata dalam konteks model ini. Reduksi 51,85% dimensi fitur (14 fitur dileliminasi) berimplikasi signifikan pada efisiensi komputasi dan interpretabilitas klinis tanpa degradasi performa yang bermakna.
3) Performa Model Random Forest
Evaluasi pada full resampled dataset menunjukkan performa yang sangat tinggi dan mendekati sempurna: accuracy 0,9908, precision 0,9925, recall 0,9892, F1-score 0,9908, dan ROC-AUC 0,9997. Confusion matrix mengkonfirmasi klasifikasi yang akurat dengan True Negative 94.146 (99,3%) dan True Positive 93.825 (98,9%), dengan tingkat kesalahan yang minimal: False Positive 708 (0,7%) dan False Negative 1.029 (1,1%). Namun, evaluasi 5-Fold Cross-Validation mengungkapkan generalization gap yang substansial[23]. Metrik rata-rata CV menunjukkan: accuracy 0,7758 (SD 0,0022), precision 0,7943 (SD 0,0018), recall 0,7444 (SD 0,0037), F1-score 0,7685 (SD 0,0026), dan ROC-AUC 0,8721 (SD 0,0019). Rentang nilai CV (min–max) menunjukkan stabilitas yang tinggi: accuracy 0,7731–0,7790, precision 0,7917–0,7966, recall 0,7390–0,7494, F1-score 0,7651–0,7723, dan ROC-AUC 0,8695–0,8750. Kurva ROC pada full data menunjukkan bentuk ideal dengan titik optimal Youden's Index pada threshold 0,48 (FPR=0,01, TPR=1,00). Kurva ROC per fold pada cross-validation menunjukkan konsistensi yang tinggi antar lipatan dengan AUC individual berkisar 0,8677–0,8730, mengkonfirmasi stabilitas model meskipun pada level performa yang lebih realistis.

3.2. Diskusi

Dari 13 fitur yang terpilih, terdapat kesesuaian yang kuat dengan literatur medis terkini mengenai faktor risiko kardiovaskular. Hipertensi (hypertension) dan riwayat penyakit jantung sebelumnya

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

(previous_heart_disease) yang menempati posisi teratas dalam RF Feature Importance sesuai dengan panduan European Society of Cardiology yang mengklasifikasikan keduanya sebagai faktor risiko mayor. Kolesterol total (cholesterol_level), LDL (cholesterol_ldl), dan trigliserida (triglycerides) yang terpilih mencerminkan peran dislipidemia dalam patogenesis aterosklerosis. Usia (age) sebagai prediktor dominan dalam Mutual Information Score konsisten dengan peningkatan risiko kardiovaskular yang eksponensial seiring pertambahan usia. Diabetes dan gula darah puasa (fasting_blood_sugar) mencerminkan hubungan erat antara resistensi insulin dan kerusakan endotelial vaskular. Tekanan darah sistolik dan diastolik (blood_pressure_systolic, blood_pressure_diastolic) penting sebagai indikator beban hemodinamik kronik. Status merokok (smoking_status) merupakan faktor risiko yang dapat dimodifikasi dengan bukti kausal kuat dalam studi epidemiologi longitudinal. Lingkar pinggang (waist_circumference) mewakili adipositas sentral yang berkaitan dengan sindrom metabolik. Jam tidur (sleep_hours) mencerminkan hubungan antara gangguan tidur dan disregulasi sistem saraf otonom serta inflamasi. Kesesuaian antara fitur yang dipilih secara statistik dengan pengetahuan klinis ini memperkuat validitas konstruk model yang dikembangkan.

Selisih signifikan antara metrik full data (0,9908) dan CV mean (0,7758) untuk accuracy mengindikasikan adanya data leakage implisit akibat evaluasi pada data pelatihan yang sama pasca-SMOTE[24]. SMOTE mensintesis sampel minoritas berdasarkan tetangga terdekat dalam ruang fitur; ketika train-test split tidak dilakukan sebelum SMOTE, informasi dari sampel sintetis yang berasal dari data uji dapat "bocor" ke data pelatihan. Meskipun penelitian ini menerapkan cross-validation yang seharusnya mengurangi leakage, implementasi SMOTE pada keseluruhan dataset sebelum CV tetap mempertahankan bias struktural[25]. Generalization gap sebesar ~21,5 poin persentase ini menjadi indikator kritis bahwa metrik full data tidak dapat dijadikan acuan untuk estimasi performa riil pada populasi baru.

3.3. Kekurangan Metode Yang Digunakan

1) Keterbastasan SMOTE Kustom

SMOTE kustom dalam penelitian ini memiliki beberapa kelemahan fundamental:Pertama, penggunaan metrik Euclidean pada ruang berdimensi tinggi (27 fitur awal) mengalami kutukan dimensionalitas (curse of dimensionality), di mana jarak antar titik menjadi hampir seragam dan kehilangan makna diskriminatif. SMOTE asli dirancang untuk data berdimensi rendah; pada dimensi tinggi, interpolasi linear dapat menghasilkan sampel sintetis yang jatuh pada region dengan densitas rendah atau bahkan di luar manifold data asli, yang dikenal sebagai noise amplification[26]. Kedua, SMOTE tidak mempertimbangkan kepadatan lokal (local density) secara adaptif. Sampel minoritas yang berada pada borderline region (dekat dengan kelas mayoritas) menghasilkan sintesis yang mungkin tumpang tindih dengan kelas mayoritas, meningkatkan class overlap. Implementasi kustom tanpa deteksi safe level atau borderline (seperti pada algoritma Borderline-SMOTE atau ADASYN) rentan terhadap masalah ini[27]. Ketiga, SMOTE kustom menghasilkan sampel sintetis dengan jumlah tetap $N_{\{\{gen\}\}} = N_{\{\{majority\}\}} - N_{\{\{minority\}\}}$ tanpa validasi kualitas sintesis. Tidak ada mekanisme untuk mengevaluasi apakah sampel baru tersebut klinis masuk akal misalnya, apakah kombinasi kolesterol 300 mg/dL dengan tekanan darah sistolik 90 mmHg secara fisiologis plausible. Keempat, SMOTE dilakukan sebelum seleksi fitur, yang berarti sampel sintetis dihasilkan pada ruang fitur penuh termasuk fitur yang kemudian dieliminasi. Ini tidak efisien secara komputasi dan berpotensi mengaburkan struktur local neighborhood pada subruang fitur yang relevan.

2) Keterbatasan Seleksi Fitur

Pendekatan SelectFromModel dengan mean threshold bersifat arbitrer dan univariat dalam konteks ensemble[28]. Ambang batas rata-rata tidak mempertimbangkan redundansi antar fitur yang terpilih misalnya, cholesterol_level, cholesterol_ldl, dan triglycerides memiliki korelasi fisiologis yang tinggi, sehingga kehadiran ketiganya bersamaan mungkin tidak memberikan informasi marginal tambahan yang signifikan. Metode ini juga tidak mengevaluasi interaksi fitur

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

tingkat tinggi yang mungkin kritis dalam prediksi risiko kardiovaskular (contoh: efek moderasi age terhadap hubungan cholesterol dan heart_attack)[29]. Selain itu, Mutual Information dan RF Feature Importance memberikan ranking yang berbeda untuk fitur-fitur atas, menunjukkan bahwa tidak ada konsensus otomatis tentang fitur "terbaik". Keputusan interseksi dapat mengeliminasi fitur yang penting dalam satu metode namun tidak dalam metode lainnya.

3) Keterbatasan Random Forest
Meskipun Random Forest robust terhadap overfitting relatif terhadap pohon tunggal, ensemble dengan 200 pohon pada data berukuran ~190K tetap rentan terhadap memorisasi pola noise dari sampel sintetis SMOTE. Class_weight="balanced" pada Random Forest sebenarnya redundan ketika SMOTE telah menyeimbangkan kelas, yang dapat menyebabkan over-correction dan bias terbalik. Hyperparameter yang digunakan min_samples_split=5, min_samples_leaf=2 cukup agresif untuk data berukuran besar, yang dapat menghasilkan pohon yang terlalu dalam dan kompleks. Tidak ada pruning atau regularisasi eksplisit yang diterapkan.

4) Keterbatasan Evaluasi
Stratified K-Fold dengan $K = 5$ pada dataset berukuran ~190K menghasilkan set pelatihan berukuran ~152K dan set validasi ~38K per lipatan. Ukuran validasi yang besar secara teoritis memberikan estimasi varians rendah, namun tidak mereplikasi skenario deployment riil di mana model diuji pada populasi dengan distribusi yang mungkin berbeda (misal: perubahan prevalensi, pergeseran demografis). Tidak ada evaluasi out-of-time atau external validation pada dataset independent[30]. Metrik evaluasi tidak mencakup calibration seberapa akurat probabilitas prediksi $P(y = 1 \mid x)$ mencerminkan proporsi aktual. Pada aplikasi klinis, kalibrasi yang buruk dapat menyebabkan overestimasi atau underestimasi risiko yang berbahaya.

3.4. Peluang Riset Lanjutan

Perbaikan pada Pipeline yang ada

Tabel 2. Pipeline

Arah Perbaikan	Metode Yang Diusulkan	Justifikasi
SMOTE adaptif	Borderline-SMOTE, ADASYN, atau SMOTE-ENN	Menghasilkan sampel sintetis hanya pada *safe regions* dan membersihkan noise secara simultan
SMOTE subruang fitur	SMOTE setelah seleksi fitur, atau SMOTE pada manifold hasil PCA/t-SNE	Mengurangi curse of dimensionality dan memastikan sintesis pada ruang yang relevan
Validasi sintesis klinis	Clinical plausibility checker berbasis aturan domain	Menjamin sampel sintetis tidak melanggar konstrain fisiologis (mis: tekanan darah sistolik > diastolik)
Evaluasi kalibrasi	Platt Scaling, Isotonic Regression, atau Expected Calibration Error (ECE)	Memastikan probabilitas prediksi dapat diinterpretasikan secara klinis

4. KESIMPULAN

Penelitian ini berhasil mendemonstrasikan pipeline end-to-end untuk klasifikasi risiko serangan jantung yang mengintegrasikan prapemrosesan data, penyeimbangan kelas menggunakan SMOTE kustom, seleksi fitur dua-tahap berbasis Mutual Information dan Random Forest Feature Importance, serta pelatihan model Random Forest dengan evaluasi robust melalui Stratified 5-Fold Cross-Validation. Hasil menunjukkan bahwa SMOTE efektif menyeimbangkan distribusi kelas menjadi 94.854 sampel per kelas, sementara seleksi fitur berhasil mereduksi 27 fitur menjadi 13 fitur terpilih yang klinis relevan tanpa degradasi performa bermakna; meskipun evaluasi pada data full resampled menghasilkan metrik mendekati sempurna (accuracy 0,9908 dan ROC-AUC 0,9997), estimasi cross-validation memberikan gambaran yang lebih realistis dengan rata-rata accuracy 0,7758 dan ROC-AUC

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.289
Page 427-437

0,8721, mengindikasikan adanya generalization gap sebesar ~21,5% akibat potensi data leakage struktural dari implementasi SMOTE sebelum validasi. Oleh karena itu, metrik cross-validation harus dijadikan acuan utama untuk estimasi performa pada populasi baru, dan pipeline yang diusulkan memberikan kontribusi metodologis sekaligus argumentasi kritis mengenai interpretasi hasil pada data sintetis dalam konteks aplikasi skrining medis.

REFERENCES

[1] W. S. P. Harmadha *et al.*, "Explaining the increase of incidence and mortality from cardiovascular disease in Indonesia: A global burden of disease study analysis (2000–2019)," *PLoS One*, vol. 18, no. 12, hal. e0294128, Des 2023, doi: 10.1371/journal.pone.0294128.

[2] T. Liu, A. J. Krentz, Z. Huo, dan V. Ćurčin, "Opportunities and Challenges of Cardiovascular Disease Risk Prediction for Primary Prevention Using Machine Learning and Electronic Health Records: A Systematic Review," *Rev. Cardiovasc. Med.*, vol. 26, no. 4, Apr 2025, doi: 10.31083/RCM37443.

[3] G. C. M. Siontis, I. Tzoulaki, K. C. Siontis, dan J. P. A. Ioannidis, "Comparisons of established risk prediction models for cardiovascular disease: systematic review," *BMJ*, vol. 344, no. may24 1, hal. e3318–e3318, Mei 2012, doi: 10.1136/bmj.e3318.

[4] O. Deniz *et al.*, "Serum total cholesterol, HDL-C and LDL-C concentrations significantly correlate with the radiological extent of disease and the degree of smear positivity in patients with pulmonary tuberculosis," *Clin. Biochem.*, vol. 40, no. 3–4, hal. 162–166, Feb 2007, doi: 10.1016/j.clinbiochem.2006.10.015.

[5] G. Ramaswami, T. Susnjak, dan A. Mathrani, "On Developing Generic Models for Predicting Student Outcomes in Educational Data Mining," *Big Data Cogn. Comput.*, vol. 6, no. 1, hal. 6, Jan 2022, doi: 10.3390/bdcc6010006.

[6] A. Özçift, "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis," *Comput. Biol. Med.*, vol. 41, no. 5, hal. 265–271, Mei 2011, doi: 10.1016/j.compbiomed.2011.03.001.

[7] J. Ramesh, R. Aburukba, dan A. Sagahyroon, "A remote healthcare monitoring framework for diabetes prediction using machine learning," *Healthc. Technol. Lett.*, vol. 8, no. 3, hal. 45–57, Jun 2021, doi: 10.1049/htl2.12010.

[8] P. Probst, M. N. Wright, dan A. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 3, Mei 2019, doi: 10.1002/widm.1301.

[9] V. Srinivasan *et al.*, "Optimizing pipelines for power and performance," in *35th Annual IEEE/ACM International Symposium on Microarchitecture, 2002. (MICRO-35). Proceedings.*, IEEE Comput. Soc, hal. 333–344. doi: 10.1109/MICRO.2002.1176261.

[10] J. V. Alonso dan L. Escot, "Robust Cross-Validation of Predictive Models Used in Credit Default Risk," *Appl. Sci.*, vol. 15, no. 10, hal. 5495, Mei 2025, doi: 10.3390/app15105495.

[11] V. C. Pezoulas *et al.*, "A computational pipeline for data augmentation towards the improvement of disease classification and risk stratification models: A case study in two clinical domains," *Comput. Biol. Med.*, vol. 134, hal. 104520, Jul 2021, doi: 10.1016/j.compbiomed.2021.104520.

[12] A. H. Syed, T. Khan, dan N. Alromema, "A Hybrid Feature Selection Approach to Screen a Novel Set of Blood Biomarkers for Early COVID-19 Mortality Prediction," *Diagnostics*, vol. 12, no. 7, hal. 1604, Jun 2022, doi: 10.3390/diagnostics12071604.

[13] I. Aruleba dan Y. Sun, "An Improved Ensemble Method With Data Resampling for Credit Risk Prediction," *IEEE Access*, vol. 13, hal. 71275–71287, 2025, doi: 10.1109/ACCESS.2025.3563432.

[14] M. Amelia dan D. Fitriyani, "Classification of Thyroid Disease Risk Using the XGBoost Method," *J. Intell. Syst. Technol. Informatics*, vol. 1, no. 3, hal. 93–97, Nov 2025, doi: 10.64878/jistics.v1i3.26.

[15] A. S. Deokar dan M. A. Pradhan, "Optimizing Cardiovascular Disease Detection Using Ranking-Based Feature Selection Machine Learning Models," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 5, hal. 28172–28178, Okt 2025, doi: 10.48084/etasr.11923.

[16] C. T. Nakas, T. A. Alonzo, dan C. T. Yiannoutsos, "Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index," *Stat. Med.*, vol. 29, no. 28, hal. 2946–2955, Des 2010, doi: 10.1002/sim.4044.

[17] S. A. Thamrin, D. S. Arsyad, H. Kuswanto, A. Lawi, dan S. Nasir, "Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018," *Front. Nutr.*, vol. 8, Jun 2021, doi: 10.3389/fnut.2021.669155.

[18] I. M. Alkhawaldeh, I. Albalkhi, dan A. J. Naswhan, "Challenges and limitations of synthetic minority oversampling techniques in machine learning," *World J. Methodol.*, vol. 13, no. 5, hal. 373–378, Des 2023, doi: 10.5662/wjm.v13.i5.373.

[19] P. Cunningham dan S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *ACM Comput. Surv.*, vol. 54, no. 6, hal. 1–25, Jul 2022, doi: 10.1145/3459665.

[20] A. Sufyan, S. Mohsin, K. Hameed, E. Az- Zo'bi, dan M. Tashtoush, "Discretization of the Inverse Rayleigh-G Family: Theoretical Properties, Machine Learning-Based Parameter Estimation, and Practical Applications," *Stat. Optim. Inf. Comput.*, vol. 14, no. 3, hal. 1174–1197, Jul 2025, doi: 10.19139/soic-2310-5070-2618.

[21] D. Chutia, D. K. Bhattacharyya, J. Sarma, dan P. N. L. Raju, "An effective ensemble classification framework using random forests and a correlation based feature selection technique," *Trans. GIS*, vol. 21, no. 6, hal. 1165–1178, Des 2017, doi: 10.1111/tgis.12268.

[22] D. Rengasamy *et al.*, "Feature importance in machine learning models: A fuzzy information fusion approach," *Neurocomputing*, vol. 511, hal. 163–174, Okt 2022, doi: 10.1016/j.neucom.2022.09.053.

[23] D. Wilimitis dan C. G. Walsh, "Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial," *JMIR AI*, vol. 2, hal. e49023, Des 2023, doi: 10.2196/49023.

[24] S. kamal Utsho, "The Decline of Synthetic Oversampling in Large-Scale Imbalanced Learning:A Post-SMOTE Empirical and Theoretical Study (2020–2025)," 2 Desember 2025. doi: 10.21203/rs.3.rs-8236211/v1.

[25] S. Gholampour, "Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, hal. 827–841, Apr 2024, doi: 10.3390/make6020039.

[26] D. P. Mitchell, "Generating antialiased images at low sampling densities," in *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, New York, NY, USA: ACM, Agu 1987, hal. 65–72. doi: 10.1145/37401.37410.

[27] I. Dey dan V. Pratap, "A Comparative Study of SMOTE, Borderline-SMOTE, and ADASYN Oversampling Techniques using Different Classifiers," in *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, IEEE, Mar 2023, hal. 294–302. doi: 10.1109/ICSMDI57622.2023.00060.

[28] P.-Y. Wong *et al.*, "Effects of feature selection methods in estimating SO2 concentration variations using machine learning and stacking ensemble approach," *Environ. Technol. Innov.*, vol. 37, hal. 103996, Feb 2025, doi: 10.1016/j.eti.2024.103996.

[29] P. Dubey, P. Dubey, dan P. N. Bokoro, "Advancing CVD Risk Prediction with Transformer Architectures and Statistical Risk Factor Filtering," *Technologies*, vol. 13, no. 5, hal. 201, Mei 2025, doi: 10.3390/technologies13050201.

[30] A. Louati dan H. Louati, "Predictive Monitoring of Wage-Band Classification in GOSI Data with Leakage Control and Out-of-Time Validation," *Forecasting*, vol. 8, no. 2, hal. 27, Mar 2026, doi: 10.3390/forecast8020027.