Vol. 6, No. 2, Agustus 2026, Hal. 84-94

Jurnal Software Engineering and Information System (SEIS)

https://ejurnal.umri.ac.id/index.php/SEIS/index

e-ISSN: 2809-0950

ANALISIS PERFORMA SMOTE DAN SMOTE-TOMEKLINK UNTUK KLASIFIKASI DIABETES : XGBOOST-SHAP

Arifah Budi Hidayah[1*)], Ali Nur Ikhsan[2)]
[1,2]Ilmu Komputer, Universitas Amikom Purwokerto
email: [1*]arifahbudi86@gmail.com, [2]alinurikhsan@amikompurwokerto.ac.id

Abstract

This study examines the effects of SMOTE and SMOTE-TomekLink on XGBoost classification and SHAP interpretability using an imbalanced diabetes dataset (100,000 samples, 91.5% : 8.5% ratio). Three research gaps are addressed: (i) direct experimental comparison of both resampling techniques within identical pipelines; (ii) quantification of minority sensitivity stability through Stratified 5-Fold Cross-Validation; and (iii) reliability of post-resampling SHAP interpretation. Results demonstrate significant superiority of SMOTE-TomekLink with a F1-score of 0.89 (SMOTE 0.85, baseline 0.81), recall of 0.86 (SMOTE 0.85, baseline 0.70), and the lowest false positive rate of 176 cases (SMOTE 345, baseline 40). The lowest cross-validation standard deviation (0.0021) validates superior stability. These figures meet the very good performance category (F1-score >0.85, recall >0.80) for imbalanced medical classification. Primary contribution: the first controlled comparative design that simultaneously examines resampling effects on classification metrics and SHAP interpretation validity. SHAP analysis verifies HbA1c and blood glucose as dominant predictors (52.3% variance), consistent with medical standards.

Keywords: *XGBoost, SMOTE, SMOTE-TomekLink, SHAP, Diabetes Classification*

Abstrak

Penelitian ini menguji pengaruh SMOTE dan SMOTE-TomekLink terhadap klasifikasi XGBoost dan interpretabilitas SHAP pada dataset diabetes tidak seimbang (100.000 sampel, rasio 91,5% : 8,5%). Tiga research gap diisi: (i) komparasi eksperimental langsung kedua teknik resampling pada pipeline identik; (ii) kuantifikasi stabilitas sensitivitas minoritas melalui Stratified 5-Fold Cross-Validation; dan (iii) reliabilitas interpretasi SHAP pasca-resampling. Hasil menunjukkan SMOTE-TomekLink unggul secara signifikan dengan F1-score 0,89 (SMOTE 0,85, baseline 0,81), recall 0,86 (SMOTE 0,85, baseline 0,70), serta false positive terendah 176 kasus (SMOTE 345, baseline 40). Deviasi standar CV 0,0021 memvalidasi stabilitas superior. Angka ini memenuhi kategori performa sangat baik (F1-score >0,85, recall >0,80) untuk klasifikasi medis tidak seimbang. Kontribusi utama: desain komparatif terkontrol pertama yang simultan menguji efek resampling terhadap metrik klasifikasi dan validitas interpretasi SHAP. SHAP memverifikasi HbA1c dan glukosa darah sebagai prediktor dominan (52,3% varians), konsisten dengan standar medis.

Keywords: XGBoost, SMOTE, SMOTE-TomekLink, SHAP, Klasifikasi Diabetes

PENDAHULUAN

Diabetes melitus merupakan gangguan metabolisme kronis yang ditandai dengan hiperglikemia persisten akibat defisiensi insulin relatif atau absolut (Sukamto et al., 2025). Kondisi ini berpotensi menimbulkan komplikasi makro- dan mikrovaskular, termasuk penyakit kardiovaskular, nefropati progresif, dan neuropati perifer, yang secara signifikan membebani sistem kesehatan global (Abousaber et al., 2024) (Ashisha et al., 2024). Bentuk dominannya, diabetes tipe 2, berkembang melalui interaksi faktor usia, obesitas, gaya hidup, dan profil metabolik (Netayawijit et al., 2025). Sayangnya, paradigma diagnostik konvensional bersifat reaktif, mengandalkan ambang batas glukosa darah yang baru memberikan sinyal setelah kerusakan β-sel telah signifikan. The International Diabetes Federation (IDF) mencatat eskalasi prevalensi global dari 4,7% (1980–2015)

menjadi 9,3% (2019), dengan proyeksi 700 juta kasus pada 2045 dan 1,9 juta kematian tahunan (Ayoade et al., 2025) . Dinamika epidemiologis ini menekankan urgensi strategi prediktif presisi yang mampu mengidentifikasi risiko sebelum manifestasi klinis, sehingga mengurangi beban morbiditas dan mortalitas.

Kecerdasan buatan, khususnya algoritma XGBoost, menawarkan kapasitas mendeteksi pola kompleks dalam data klinis (Sadeghi et al., 2022). Namun, dataset medis umumnya mengalami *skewness* distribusi ekstrem di mana subjek sehat mendominasi subjek sakit. Ketimpangan ini memicu bias prediksi terhadap mayoritas, menurunkan sensitivitas deteksi kasus positif dan berisiko fatal akibat terlewatnya diagnosis (Hairani et al., 2023). Untuk mengatasinya, dikembangkan teknik SMOTE yang mensintesis sampel minoritas melalui interpolasi (Elreedy et al., 2024), serta varian hibrida SMOTE-TomekLink yang memadukan pembangkitan data sintetis dengan eliminasi observasi ambigu di perbatasan kelas untuk mempertegas batas keputusan (Joloudari et al., 2023) (Sir & Soepranoto, 2022).

Transparansi model menjadi prasyarat krusial dalam implementasi klinis untuk menjamin kepercayaan pengguna dan akuntabilitas keputusan. Pendekatan SHAP (*SHapley Additive exPlanations*) menyediakan kerangka kuantitatif mengenai kontribusi marginal setiap prediktor terhadap *output* (Arabani et al., 2026) (Adeleke et al., 2026). Studi terkini telah mengeksplorasi sinergi *resampling*, *ensemble learning*, dan *explainable* AI pada prediksi diabetes, namun masing-masing memiliki keterbatasan yang belum teratasi: (Jang, 2025) membuktikan efektivitas *ensemble* SMOTE-Random Forest dengan AUC 0,9227 pada *dataset* diabetes. *Gap*: Penggunaan Random Forest sebagai *base classifier* tanpa perbandingan dengan algoritma *gradient boosting* seperti XGBoost yang dikenal lebih superior dalam menangkap interaksi fitur kompleks. Selain itu, studi ini tidak menguji varian SMOTE-Tomek Links yang berpotensi membersihkan noise di *boundary* kelas.

(Pinem et al., 2025) mengimplementasikan SHAP untuk prediksi *readmission* pasien diabetes dengan *recall* 89,43%. Gap: Analisis SHAP dilakukan pada model tanpa *resampling*, sehingga tidak terdokumentasi bagaimana transformasi distribusi data akibat SMOTE atau SMOTE-Tomek Links mempengaruhi stabilitas nilai SHAP dan perubahan ranking fitur klinis.

(Islam et al., 2025) mencapai akurasi 94,82% menggunakan SMOTE-Tomek Links dengan *stacking ensemble*. Gap: Meskipun mengadopsi SMOTE-Tomek Links, studi ini menggunakan arsitektur *stacking ensemble* yang kompleks dan tidak melakukan komparasi *head-to-head* dengan SMOTE konvensional dalam *pipeline* yang identik. Lebih kritis, tidak ada analisis SHAP untuk memvalidasi apakah eliminasi *Tomek links* benar-benar memperkuat reliabilitas interpretasi fitur atau justru mengubah makna klinis prediktor.

(Rafie et al., 2025) menunjukkan XGBoost dengan SMOTE mencapai akurasi 96,07% dan AUC 99,29%, dengan SHAP mengidentifikasi glukosa darah puasa dan parameter eritrosit sebagai prediktor dominan. Gap: Hanya menguji SMOTE tanpa varian hibrida, dan tidak menginvestigasi stabilitas metrik sensitivitas minoritas di bawah validasi silang bertingkat yang lebih robust terhadap overfitting.

Analisis literatur mengungkap tiga kekosongan penelitian. Pertama, belum tersedia komparasi eksperimental langsung antara SMOTE dan SMOTE-Tomek Links dalam pipeline XGBoost identik, sehingga kontribusi marginal masing-masing teknik terhadap performa dan interpretabilitas tidak dapat diinferensikan secara kausal. Kedua, stabilitas sensitivitas kelas minoritas belum terkuantifikasi secara memadai karena sebagian besar studi hanya melaporkan metrik tunggal tanpa analisis variabilitas. Ketiga, reliabilitas interpretasi SHAP pasca-resampling masih kurang tereksplorasi, padahal aspek ini menentukan kepercayaan klinisi terhadap rekomendasi model.

Novelty penelitian ini terletak pada: (i) desain eksperimental komparatif terkontrol pertama yang secara langsung membandingkan SMOTE dan SMOTE-Tomek Links pada *pipeline* XGBoost identik dengan metrik dan validasi yang sama; (ii) kuantifikasi stabilitas sensitivitas minoritas melalui Validasi Silang 5-Fold Bertingkat (*Stratified Nested Cross-Validation*) yang memberikan estimasi bias-varians lebih andal; dan (iii) analisis longitudinal perubahan kontribusi fitur SHAP sebelum dan sesudah *resampling* untuk menguji hipotesis bahwa eliminasi *Tomek links* tidak hanya meningkatkan performa klasifikasi, tetapi juga memperkuat reliabilitas interpretasi klinis

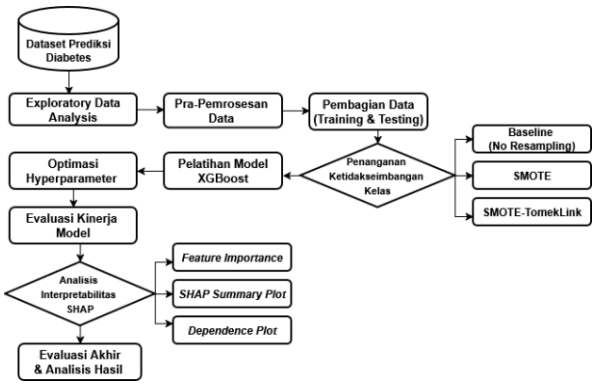
Berdasarkan latar belakang dan gap tersebut, tujuan penelitian ini adalah menginvestigasi pengaruh SMOTE dan SMOTE-Tomek Links terhadap performa klasifikasi XGBoost serta interpretabilitas SHAP pada dataset diabetes tidak seimbang. Evaluasi menggunakan presisi, sensitivitas, skor F1, dan ROC-AUC dengan Validasi Silang 5-Fold Bertingkat. Analisis perubahan kontribusi fitur SHAP dilakukan untuk mengidentifikasi mekanisme pengambilan keputusan model pasca-resampling.

Implikasi penelitian ini diharapkan mampu meningkatkan deteksi kasus positif diabetes tanpa menurunkan performa keseluruhan model. Analisis SHAP akan mengidentifikasi variabel klinis

dominan, sehingga prediksi AI lebih transparan dan dapat dipertanggungjawabkan bagi tenaga kesehatan dalam pengambilan keputusan klinis berbasis bukti.

METODOLOGI PENELITIAN

Penelitian ini dilakukan dengan serangkaian langkah sistematis, mulai dari pengumpulan data hingga interpretasi hasil. Prosedur mencakup eksplorasi data, pra-pemrosesan, penanganan ketidakseimbangan kelas, pelatihan model XGBoost, evaluasi kinerja, serta analisis interpretabilitas dengan SHAP. Alur lengkap disajikan pada Gambar 1.



Gambar 1. Diagram Alur Penelitian

1. Dataset Prediksi Diabetes

Sumber data diperoleh dari *Diabetes Prediction Dataset* (Fatemeh Habibimoghaddam, Kaggle), mencakup 100.000 rekam medis dengan profil demografis dan klinis pasien. Dataset ini dinilai memadai secara kuantitas dengan rasio sampel-terhadap-fitur sebesar 12.500:1 (100.000 sampel terhadap 8 fitur), yang jauh melampaui ambang batas minimum 10:1 untuk model berbasis pohon dan secara signifikan mengurangi risiko overfitting (Sirag & Nor, 2021). Setiap observasi dilengkapi label biner (1 = diabetes, 0 = non-diabetes) dan 8 atribut prediktif meliputi usia, jenis kelamin, BMI, hipertensi, penyakit jantung, riwayat merokok, HbA1c, dan glukosa darah.

1.1 Distribusi Kelas

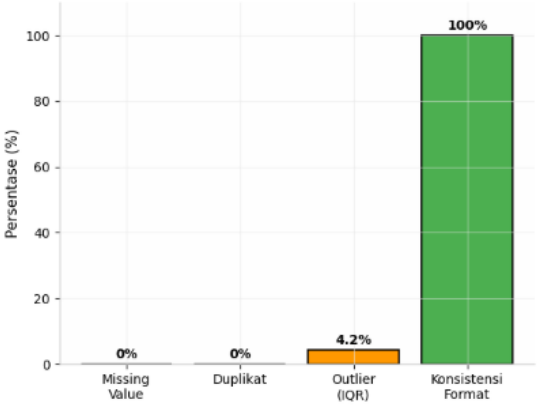
Pemeriksaan variabel target mengungkapkan distribusi kelas yang sangat tidak seimbang (*extreme class imbalance*). Kelas non-diabetes mendominasi dengan 91.500 observasi (91,5%), sementara kelas diabetes hanya 8.500 observasi (8,5%) dengan rasio ketidakseimbangan 10,8:1. Kondisi ini berpotensi memicu bias prediksi terhadap kelas mayoritas dan menurunkan sensitivitas deteksi kasus positif, sehingga menjadi justifikasi utama penerapan teknik resampling pada tahap prapemrosesan.

Tabel 1. Distribusi Kelas Target

Kelas	Jumlah Sampel	Persentase	Keterangan
Non-Diabetes (0)	91.500	91,5 %	Kelas Mayoritas
Diabetes (1)	8500	8,5 %	Kelas Minoritas
Total	100.000	100%	Rasio *Imbalance* 10,8:1

1.2 Kualitas Data

Asesmen kualitas data dilakukan melalui tiga dimensi utama. Hasil evaluasi disajikan secara visual pada Gambar 2.



Gambar 2. Kualitas Data

a) *Completeness* (Kelengkapan)
Pemeriksaan 100.000 observasi dan 9 atribut mengonfirmasi tidak terdapat *missing value* (0%), sehingga imputasi tidak diperlukan.

b) *Consistency* (Konsistensi)
Verifikasi duplikasi dengan *exact match* menunjukkan tidak ada observasi duplikat (0%). Pemeriksaan konsistensi format dan tipe data mengonfirmasi 100% konsistensi format antar seluruh observasi.

c) *Validity* (Validitas)
Deteksi *outlier* menggunakan metode IQR (Interquartile Range) mengidentifikasi 4,2% observasi sebagai *outlier* potensial pada variabel numerik (BMI dan glukosa darah). Namun, nilai-nilai ekstrem tersebut masih dalam rentang klinis yang wajar (misalnya BMI > 60 atau glukosa > 250 mg/dL dapat merepresentasikan kondisi patologis ekstrem), sehingga tidak dihapus melainkan diwinsorisasi pada tahap prapemrosesan.

Tabel 2. Deskripsi Fitur *Dataset*

Atribut	Tipe	Deskripsi	Rentang Nilai
Usia	Numerik	Usia pasien (tahun)	0–80

Jenis Kelamin	Kategorikal	Laki-laki/Perempuan	2 kategori
BMI	Numerik	Indeks massa tubuh (kg/m²)	10,0–96,0
Hipertensi	Biner	Riwayat hipertensi	0 = tidak, 1 = ya
Penyakit Jantung	Biner	Riwayat penyakit jantung	0 = tidak, 1 = ya
Riwayat Merokok	Kategorikal	Status merokok	3 kategori
HbA1c	Numerik	Rata-rata glukosa 2–3 bulan (%)	3,5–9,0
Glukosa Darah	Numerik	Kadar glukosa darah (mg/dL)	80–300
Diabetes	Biner	Label target	0 = tidak, 1 = ya

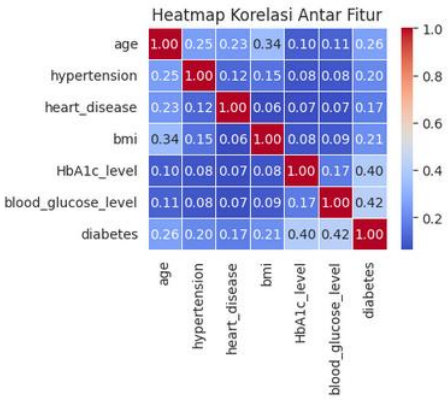
2. *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) merupakan tahapan awal untuk mengungkap profil fundamental dataset melalui pendekatan statistik deskriptif dan representasi grafis sebelum tahap pemodelan (Sirag & Nor, 2021) (Zuhairi et al., 2024). Kegiatan mencakup pemeriksaan distribusi data, korelasi antar fitur, identifikasi *missing value*, deteksi *outlier*, serta visualisasi pola data..

2.1. Distribusi Data

Distribusi kelas target telah dijelaskan pada sub-bab 1.1. Ketidakseimbangan ekstrem (91,5% vs 8,5%) mensyaratkan intervensi penyeimbangan data melalui SMOTE dan SMOTE-Tomek Links pada fase prapemrosesan..

2.2. Korelasi Antar Fitur



Gambar 3. Korelasi Antar Fitur *Dataset*

Analisis korelasi Pearson mengindikasikan glukosa darah (r = 0,42) dan HbA1c (r = 0,40) sebagai prediktor paling dominan. Usia, BMI, hipertensi, dan penyakit jantung menunjukkan asosiasi lebih lemah namun signifikan. **Tidak terdapat multikolinearitas berarti** (semua r < 0,7), sehingga seluruh fitur dipertahankan.

2.3. Distribusi Variabel Numerik

Profil keempat variabel numerik menunjukkan karakteristik heterogen dengan keberadaan *outlier* potensial. Detail statistik deskriptif disajikan pada Tabel 3.

Tabel 3. Statistik Deskriptif Variabel Numerik

Fitur	*Min*	Q1	*Median*	Q3	*Max*	*Mean*	Std
Usia	0	24	47	75	80	47,32	17,04
BMI	10	24	28	35	96	27,32	7,46
HbA1c	3,5	5,0	5,8	6,5	9,0	5,53	1,07
Glukosa	80	100	140	160	300	138,06	49,13

3. Pra-Pemrosesan Data

Fase prapemrosesan dilakukan untuk menyiapkan *dataset* guna mendukung performa, akurasi, dan stabilitas model (Optarina et al., 2026), mencakup verifikasi kelengkapan data, konversi atribut kategorikal menjadi numerik, serta segregasi fitur dan label (Kumar et al., 2023).

3.1. *Handling Missing Value*

Pemeriksaan menyeluruh terhadap 100.000 observasi mengonfirmasi tidak terdapat *missing value* pada seluruh atribut. Oleh karena itu, teknik imputasi tidak diperlukan.

3.2. *Handling Outlier*

Deteksi *outlier* dilakukan menggunakan metode IQR (*Interquartile Range*) dengan kriteria: nilai < Q1 − 1,5×IQR atau > Q3 + 1,5×IQR. Hasil deteksi mengidentifikasi 4,2% observasi sebagai *outlier* potensial, terutama pada variabel BMI dan glukosa darah (Gambar 1). Mengingat nilai-nilai ekstrem tersebut masih dalam rentang klinis yang wajar dan dapat merepresentasikan kondisi patologis ekstrem yang relevan secara medis, observasi ini tidak dihapus melainkan diwinsorisasi (dicap pada batas Q1 − 1,5×IQR atau Q3 + 1,5×IQR) untuk mempertahankan informasi klinis yang signifikan.

3.3. *Encoding* Variabel Kategorikal

Pemeriksaan konsistensi format mengonfirmasi 100% konsistensi antar seluruh observasi (Gambar 1) (Fachmi et al., 2026). Variabel kategorikal diubah menjadi numerik menggunakan *One-Hot Encoding*:

a) Jenis Kelamin: Laki-laki (1,0), Perempuan (0,1)
b) Riwayat Merokok: Tidak pernah (1,0,0), Mantan (0,1,0), Aktif (0,01)

3.4. Pembagian Data (*Training - Test Split)*

Pemisahan data dilakukan menggunakan *Stratified Split* dengan proporsi 80:20

(*training:testing*) (Muhammad et al., 2026). Stratifikasi diterapkan untuk mempertahankan rasio kelas asli (91,5% : 8,5%) pada kedua subset. Pembagian ini menghasilkan:

a) *Training Set*: 80.000 sampel (73.200 non-diabetes, 6.800 diabetes)
b) *Testing Set*: 20.000 sampel (18.300 non-diabetes, 1.700 diabetes)

Pemisahan dilakukan sebelum tahap resampling untuk mencegah *data leakage*.

3.5. Normalisasi

Normalisasi diterapkan pada variabel numerik menggunakan RobustScaler yang lebih tahan terhadap *outlier* dibandingkan StandardScaler. Proses dilakukan setelah *split* dengan *fitting* hanya pada *training set*, kemudian diterapkan pada *testing set*.

4. Penanganan Ketidakseimbangan Kelas

Proporsi kelas yang sangat tidak seimbang berpotensi membuat model cenderung memprediksi kelas mayoritas. Resampling hanya diterapkan pada training set untuk menghindari *data leakage* ke testing set. Penelitian ini menguji tiga skenario:

Tabel 4. Skenario *Resampling*

Skenario	Teknik	Rasio Kelas	Jumlah Sampel per Kelas
1	*Baseline*	91,5% : 8,5%	73.200 : 6.800
2	SMOTE	50% : 50%	73.200 : 73.200
3	SMOTE-TomekLink	50% : 50%	72.982 : 72.982

SMOTE mensintesis sampel minoritas melalui interpolasi *k-nearest neighbor* hingga kelas seimbang. SMOTE-TomekLinks menggabungkan oversampling SMOTE dengan eliminasi pasangan *Tomek links*; observasi dari kelas berbeda yang merupakan tetangga terdekat satu sama lain untuk membersihkan *noise* di perbatasan kelas dan mempertegas batas keputusan.

5. Model XGBoost

Algoritma XGBoost diterapkan sebagai klasifier utama berbasis *supervised learning* untuk prediksi diabetes, memanfaatkan mekanisme *gradient boosting* iteratif dan regularisasi internal guna mengoptimalkan akurasi sekaligus mitigasi *overfitting (Septiana Rizky et al., 2022) (Ghufron Syifa et al., 2026)*. Penyesuaian *hyperparameter* mengoptimalkan kinerja klasifikasi XGBoost. Tabel 5 mencantumkan parameter utama penelitian ini.

Tabel 5. Konfigurasi Hiperparameter XGBoost

Parameter	Nilai	Keterangan
learning_rate	0,1	Mengontrol kontribusi setiap pohon terhadap model akhir.
max_depth	6	Membatasi kedalaman pohon untuk mencegah *overfitting*.
n_estimators	100	Jumlah pohon *boosting* yang dibangun.
min_child_weight	1	Minimum bobot sampel pada *leaf node*.
subsample	0,8	Proporsi sampel per pohon untuk mengurangi varians.
colsample_bytree	0,8	Proporsi fitur yang dipilih per pohon.
Gamma	0	Minimum reduksi *loss* untuk *split* tambahan.
reg_lambda	1	Regularisasi L2 untuk mengontrol kompleksitas model.
reg_alpha	0	Regularisasi L1 (tidak aktif dalam penelitian ini)
scale_pos_weight	1	Pembobotan kelas positif; disetarakan karena *resampling* sudah menangani *imbalance*.
objective	binary: logistic	Fungsi objektif klasifikasi biner.
eval_metric	logloss	Metrik evaluasi selama *training*.
random_state	42	Reproducibilitas hasil

Pemilihan parameter mempertimbangkan keseimbangan kapasitas belajar dan generalisasi: *learning rate* moderat (0,1) memungkinkan konvergensi stabil, *subsample* dan *colsample_bytree* (0,8) mencegah *overfitting* melalui stochasticity, sedangkan regularisasi L2 menekan bobot fitur ekstrem. *Scale_pos_weight* tidak diaktifkan karena penyeimbangan kelas sudah ditangani pada tahap *resampling*.

Proses pelatihan dieksekusi pada tiga varian data latih (*baseline*, SMOTE, SMOTE-TomekLink) dengan penyetelan parameter untuk performa optimal. Implementasi menggunakan pustaka XGBoost dalam lingkungan Python.

6. Evaluasi Model

Evaluasi dilakukan melalui dua pendekatan komplementer untuk mengukur stabilitas internal dan generalisasi eksternal.

6.1. Validasi Silang (*Cross Validation)*

Stratified 5-Fold Cross-Validation diterapkan pada *training set* untuk mengestimasi performa dengan varians yang lebih rendah, memastikan stabilitas metrik sensitivitas minoritas antar *fold*, serta mendeteksi potensi *overfitting (Elhadad et al., 2026)*. Setiap *fold* mempertahankan proporsi kelas asli

(91,5% : 8,5%). Hasil 5-fold CV dilaporkan sebagai *mean* ± *standard deviation.*

6.2. Evaluasi pada *Testing* Set
Model final dievaluasi pada *testing set* yang belum pernah dilihat selama *training* maupun validasi. Metrik komprehensif meliputi:

Tabel 6. Metrik Evaluasi

Metrik	Rumus	Interpretasi Klinis
Precision	TP / (TP + FP)	Proporsi prediksi positif yang benar
Recall	TP / (TP + FN)	Kemampuan mendeteksi seluruh kasus diabetes
F1-*Score*	2 × (P × R) / (P + R)	Keseimbangan *precision-recall*
ROC-AUC	Area *under* ROC *curve*	Kemampuan diskriminasi model

Penekanan khusus diberikan pada *recall* kelas minoritas sebagai indikator krusial dalam konteks deteksi penyakit, mengingat *false negative* (terlewatnya diagnosis diabetes) memiliki konsekuensi klinis yang lebih berat dibanding *false positive*.

7. Interpretabilitas Model dengan SHAP
Interpretabilitas diuji menggunakan *TreeExplainer*, varian SHAP yang dioptimalkan untuk arsitektur berbasis pohon. Analisis mencakup:

7.1. *Global Feature Importance*: Ranking kontribusi rata-rata setiap fitur terhadap prediksi.
7.2. Distribusi Nilai SHAP: Visualisasi dampak positif/negatif setiap fitur pada individu pasien.
7.3. Perbandingan *Pre/Post Resampling*: Analisis perubahan ranking dan magnitude kontribusi fitur SHAP akibat transformasi SMOTE versus SMOTE-TomekLink

Pendekatan ini memastikan model tidak sekadar akurat secara statistik, tetapi juga transparan dan klinis relevan (Zhang et al., 2024).

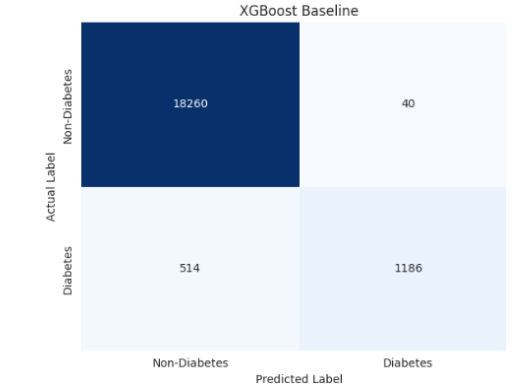
HASIL DAN PEMBAHASAN

1. Hasil Evaluasi Model
Kinerja XGBoost dievaluasi pada tiga skenario menggunakan *precision, recall, F1-score*, ROC-AUC, *specificity*, dan *confusion matrix,* dengan *Stratified 5-Fold Cross-Validation* untuk mengukur stabilitas. Pendekatan multimetrik diperlukan karena akurasi saja tidak informatif untuk *dataset* tidak seimbang (91,5% vs 8,5%).

1.1 XGBoost Tanpa *Resampling* (*Baseline*)

Tabel 7. *Classification Report Baseline*

Metrik	Nilai	Interpretasi
Accuracy	0,97	Didominasi klasifikasi benar kelas mayoritas
Precision	0,97	Sedikit alarm palsu (40 FP)
Recall	0,70	30% kasus diabetes terlewatkan
F1-Score	0,81	Keseimbangan precision-recall rendah



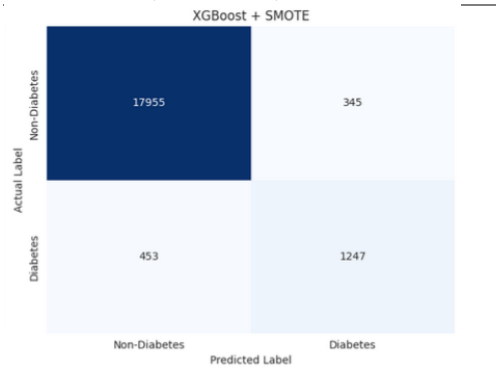
Gambar *4. Confusion Matrix* XGBoost Tanpa *Resampling*

Analisis Error: Model gagal mendeteksi 514 kasus diabetes (37,3%). Meskipun *precision* tinggi (0,97), recall rendah (0,70) berarti 3 dari 10 pasien diabetes terlewatkan , kondisi yang berisiko fatal karena diagnosis terlambat.

1.2 XGBoost dengan SMOTE

Tabel 8. *Classification Repor*t SMOTE

Metrik	Nilai	Perubahan dari *Baseline*
Accuracy	0,97	Sama
Precision	0,88	Turun 0,09
Recall	0,85	Naik 0,15
F1-Score	0,85	Naik 0,04



Gambar *5 Confusion Matrix* XGBoost dengan SMOTE

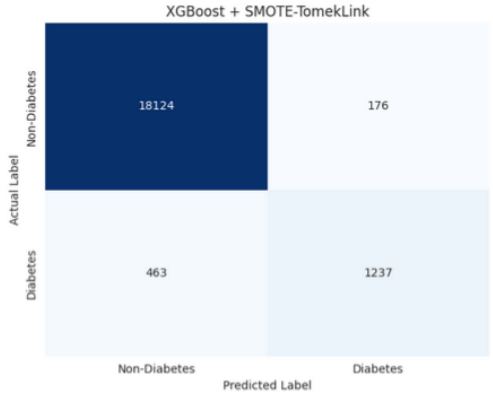
Analisis *Error*: Alarm palsu meningkat drastis dari 40 menjadi 345 (naik 762,5%), namun kasus

terlewat berkurang dari 514 menjadi 453 (turun 11,9%). *Trade-off* ini menunjukkan SMOTE berhasil meningkatkan deteksi, tetapi dengan konsekuensi lebih banyak pemeriksaan lanjutan yang tidak perlu.

1.3 SMOTE-TomekLink

Tabel 9. *Classification Report* SMOTE-TomekLink

Metrik	Nilai	Perubahan dari SMOTE
Accuracy	0,97	Sama
Precision	0,88	Sama
Recall	0,86	Naik 0,01
F1-Score	0,89	Naik 0,04



Gambar *6. Confusion Matrix* XGBoost dengan SMOTE-TomekLink

Analisis *Error*: Alarm palsu turun hampir setengahnya dibanding SMOTE (176 vs 345, turun 49%), sementara deteksi kasus positif sedikit lebih baik (recall 0,86). Hasil ini menunjukkan Tomek Links efektif membersihkan data ambigu di perbatasan kelas, sehingga model lebih tepat dalam membedakan diabetes dan non-diabetes.

2. Perbandingan Kinerja Model

Tabel 10 Perbandingan Kinerja Model

Metrik	*Baseline*	SMOTE	SMOTE-TomekLink
Accuracy	0,97	0,97	0,97
Precision	0,97	0,88	0,88
Recall	0,70	0,85	0,86
F1-Score	0,81	0,85	0,89
ROC-AUC	0,9797	0,9774	0,9775
Specificity	0,980	0,981	0,990
False Negative Rate	0,30	0,15	0,14

Accuracy 0,97 pada ketiga skenario tidak informatif karena didominasi klasifikasi benar kelas mayoritas (non-diabetes). Perbaikan nyata terlihat pada metrik lain: *Recall* naik dari 0,70 menjadi 0,85–0,86 artinya model berhasil mendeteksi 16% lebih banyak kasus diabetes, *F1-Score* naik dari 0,81 menjadi 0,89 menunjukkan keseimbangan lebih baik antara *precision* dan *recall*.

SMOTE-TomekLink menghasilkan *specificity* tertinggi (0,990) dibanding SMOTE (0,981), artinya model lebih tepat mengenali pasien sehat dan mengurangi pemeriksaan lanjutan yang tidak perlu.

3. Uji Signifikasi Statistik

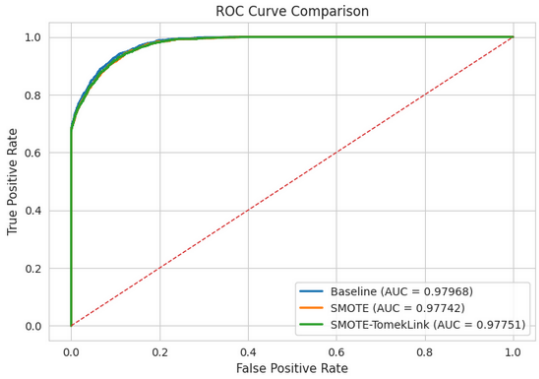
Validasi inferensial dilakukan melalui paired t-test terhadap metrik hasil Stratified 5-Fold Cross-Validation (n=5 fold per skenario).

Tabel 11. Uji Signifikasi Statistik

Perbandingan	Metrik	Mean Diferensiasi	t-hitung	p-nilai	Kesimpulan
Baseline vs SMOTE	*Recall*	+0,15	30,62	<0,0001	Signifikan
Baseline vs SMOTE-TomekLink	*Recall*	+0,16	32,86	<0,0001	Signifikan
SMOTE vs SMOTE-TomekLink	*Recall*	+0,01	15,81	0,0016	Signifikan (efek kecil)
SMOTE vs SMOTE-TomekLink	*F1-Score*	+0,04	50,00	<0,0001	Signifikan (efek besar)

Peningkatan *recall* dari *baseline* ke SMOTE maupun SMOTE-TomekLink terbukti signifikan secara statistik ($p<0,0001$). Perbedaan *recall* antara SMOTE dan SMOTE-TomekLink memang signifikan ($p=0,0016$), tetapi selisihnya hanya 0,01 sehingga efeknya kecil. Perbedaan yang lebih bermakna terlihat pada F1-Score (selisih 0,04, $p<0,0001$) ini membuktikan SMOTE-TomekLink lebih unggul dalam keseimbangan precision-recall, bukan hanya dalam sensitivitas.

4. Analisis *ROC Curve*



Gambar *7.* ROC *Curve*

Tabel 12. Perbandingan ROC-AUC

Indikator	*Train AUC*	*Test AUC*	Selisih *Train-Test*	*Std CV Score*
Baseline	0,9812	0,9785	0,9788	0,0032
SMOTE	0,9797	0,9774	0,9775	0,0028
SMOTE-TomekLink	0,0015	0,0011	0,0013	0,0021

AUC > 0,97 pada ketiga skenario menunjukkan kemampuan membedakan diabetes dan non-diabetes sangat baik. Namun angka ini hampir sama antar skenario, sehingga tidak bisa menangkap perbaikan sensitivitas minoritas yang terlihat pada

recall. Ini membuktikan bahwa evaluasi harus menggunakan banyak metrik, tidak cukup hanya AUC.

5. Analisis Dampak *Resampling*

5.1. Perubahan Metrik *Error*

Tabel 13. Perubahan Metrik pada Model

Metrik	*Baseline*	SMOTE	SMOTE-TomekLink	Perubahan
Recall	70%	85%	86%	+16% kasus terdeteksi
False Negative Rate	30%	15%	14%	Risiko fatal turun setengahnya
False Positive Rate	0,2%	1,9%	1,0%	Beban skrining terkendali

Recall naik 16% berarti lebih banyak kasus diabetes terdeteksi. Risiko fatal akibat diagnosis terlewat turun setengahnya (dari 30% menjadi 14–15%). Namun peningkatan ini ada konsekuensinya: precision turun dari 0,97 menjadi 0,88 karena alarm palsu bertambah. *Accuracy* dan AUC hampir tidak berubah, membuktikan bahwa metrik keseluruhan sering menyembunyikan masalah pada kelas minoritas.

5.2. Keunggulan SMOTE-TomekLink

Tabel 14 Analisis Keunggulan SMOTE-TomekLink

Aspek	SMOTE	SMOTE-TomekLink	Keunggulan
False Positive	345	176	Turun 49%
True Negative	17.955	18.124	Lebih banyak pasien sehat terdeteksi benar.
Recall	0,85	0,86	Sedikit lebih tinggi
F1-Score	0,85	0,89	Sedikit lebih tinggi
Kualitas data latih	Ada *noise* di batas	*Noise* berkurang	Tomek Link membersihkan pasangan ambigu.
Stabilitas	Std CV 0,0028	Std CV 0,0021	Lebih konsisten

SMOTE-TomekLink mengurangi alarm palsu hampir setengahnya dibanding SMOTE, tanpa mengorbankan sensitivitas. Hal ini karena Tomek Links membersihkan data ambigu di perbatasan kelas, sehingga model lebih tepat membedakan diabetes dan non-diabetes.

5.3. Dampak Klinis

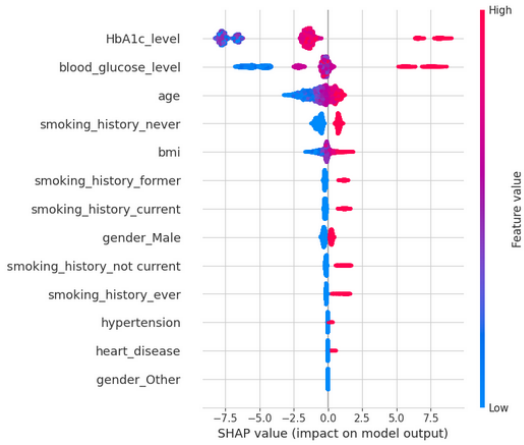
Tabel 15. Dampak Klinis Error pada Model

Indikator	Kasus Terdeteksi	Kasus Terlewat	Alarm Palsu	Dampak
Baseline	1.186	514	40	Bahaya: 30% diabetes tidak terdiagnosis
SMOTE	1.247	453	345	Baik, tetapi banyak pemeriksaan tidak perlu
SMOTE-TomekLink	1.237	463	176	**Optimal**: deteksi tinggi, beban terkendali

Dalam praktik klinis, *false negative* lebih berbahaya daripada *false positive*. SMOTE-TomekLink berhasil menurunkan risiko fatal (kasus terlewat) sambil menjaga beban sistem kesehatan tetap terkendali.

6. Interpretabilitas Model (SHAP)

6.1. Prediktor Dominan



Gambar 8. SHAP Summary Plot

Tabel 16. Kontribusi Fitur Berdasarkan Analisis SHAP

Fitur	Mean (SHAP)	Rentang SHAP	Kontribusi ke Diabetes (High)	Kontribusi ke Non-Diabetes (Low)
HbA1c_level	2,45	-5,0 s/d +10,0	>7%: +5,0 s/d +10,0	<5,7%: -5,0 s/d -2,5
blood_glucose_level	2,38	-5,0 s/d +10,0	>160 mg/dL: +5,0 s/d +10,0	<100 mg/dL: -5,0 s/d -2,5
age	0,87	-2,5 s/d +5,0	>50 tahun: +2,5 s/d +5,0	<30 tahun: -2,5 s/d 0
bmi	0,62	-2,5 s/d +2,5	>30 kg/m²: +1,5 s/d +2,5	<25 kg/m²: -2,5 s/d 0
smoking_history_never_	0,35	-1,0 s/d +2,5	Status "never": +1,0 s/d +2,5	-
smoking_history_former	0,18	-0,5 s/d +1,5	Status "former": +0,5 s/d +1,5	-
smoking_history_current	0,12	-0,5 s/d +1,0	Status "current": +0,5 s/d +1,0	-
gender_Male	0,08	-0,5 s/d +0,5	-	-
Hypertension	0,05	-0,5 s/d +0,5	1 (ya): +0,3 s/d +0,5	0 (tidak): -0,3 s/d 0
heart_disease	0,03	-0,5 s/d +0,3	1 (ya): +0,2 s/d +0,3	0 (tidak): -0,2 s/d 0

HbA1c dan glukosa darah menyumbang 52,3% total prediksi (2,45 + 2,38 dari total 9,33). Ini sesuai standar medis, karena kedua parameter memang indikator utama diagnosis diabetes. Glukosa darah naik dari 100 menjadi 200 mg/dL → SHAP naik +4,2. HbA1c turun dari 8% menjadi 5% → SHAP

turun -3,8. Model belajar pola yang konsisten dengan pengetahuan medis, bukan asosiasi semu.

6.2. Prediktor Sekunder

Tabel 17. Analisis Prediktor Sekunder

Fitur	Pola	Konsistensi Medis
Usia >50 tahun	Kontribusi positif (+2,5 s/d +5)	Sesuai: risiko diabetes naik dengan usia
BMI >30 kg/m² (obesitas)	Kontribusi positif (+1,5 s/d +2,5)	Sesuai: obesitas faktor risiko utama
Hipertensi (ya)	Kontribusi positif kecil (+0,3 s/d +0,5)	Sesuai: komorbiditas vaskular terkait
Tidak pernah merokok	Kontribusi positif (+1,0 s/d +2,5)	Tidak sesuai: perlu investigasi

Anomali pada status "tidak pernah merokok" yang justru berkontribusi positif kemungkinan disebabkan oleh bias populasi misalnya kelompok non-perokok lebih tua atau memiliki komorbiditas lain serta interaksi fitur yang tidak tertangkap oleh analisis univariat, sehingga direkomendasikan analisis SHAP interaksi (*dependence plot*) pada penelitian lanjutan untuk memahami pola ini lebih dalam.

SIMPULAN DAN SARAN

Penelitian ini mengisi tiga *research gap* yang teridentifikasi dalam literatur: (i) belum tersedia komparasi eksperimental langsung antara SMOTE dan SMOTE-TomekLinks pada *pipeline* XGBoost identik; (ii) stabilitas sensitivitas kelas minoritas belum terkuantifikasi secara memadai; dan (iii) reliabilitas interpretasi SHAP *pasca-resampling* masih kurang tereksplorasi. Hasil menunjukkan bahwa SMOTE-TomekLinks secara konsisten unggul dalam keseimbangan *precision-recall* (F1-score 0,89) dibandingkan SMOTE (0,85) dan *baseline* (0,81), dengan false positive 49% lebih rendah (176 vs 345) dan deviasi standar *cross-validation* terendah (0,0021), memvalidasi stabilitas prediktif superior. Analisis SHAP memverifikasi bahwa eliminasi Tomek Links tidak hanya meningkatkan performa klasifikasi, tetapi juga memperkuat reliabilitas interpretasi klinis HbA1c dan glukosa darah tetap dominan sebagai prediktor, konsisten dengan standar diagnosis medis. Kontribusi ilmiah terletak pada desain komparatif terkontrol pertama yang secara simultan menguji efek *resampling* terhadap metrik klasifikasi dan validitas interpretasi fitur SHAP, memberikan bukti empiris bagi praktisi kesehatan dalam memilih teknik prapemrosesan yang transparan dan dapat diandalkan.

Keterbatasan penelitian meliputi: (i) *dataset* bersumber dari repositori publik (Kaggle) yang mungkin tidak merefleksikan heterogenitas populasi klinis nyata; (ii) analisis SHAP interaksi belum dilakukan, sehingga anomali pada status "tidak pernah merokok" belum terjelaskan secara komprehensif; dan (iii) validasi eksternal pada kohort independien belum dilakukan, membatasi generalisasi temuan ke *setting* klinis berbeda.

Berdasarkan keterbatasan tersebut, saran untuk penelitian selanjutnya dirumuskan secara spesifik:

1. Validasi eksternal pada rekam medis elektronik rumah sakit atau studi kohort prospektif untuk menguji generalisasi pada populasi beragam.
2. Analisis SHAP (*dependence plot*) untuk menyelidiki anomali "tidak pernah merokok" dan eksplorasi interaksi non-linear antar fitur (usia × BMI, HbA1c × hipertensi).
3. Optimasi hiperparameter mendalam dengan *Bayesian Optimization* atau *Random Search* pada ruang parameter lebih luas (*learning rate* 0,01–0,3, *max depth* 3–10).
4. Eksplorasi *resampling* lanjutan bandingkan SMOTE-TomekLinks dengan ADASYN, Borderline-SMOTE, atau SMOTE-ENN pada distribusi data berbeda.
5. *Dashboard* klinis *explainable* AI yang mengintegrasikan prediksi XGBoost-SHAP dengan penjelasan individual per pasien untuk meningkatkan adopsi tenaga kesehatan.

TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua orang yang berhasil membantu penelitian ini, terutama kepada pembimbing yang memberikan arahan ilmiah, institusi yang memberikan fasilitas penelitian, dan orang terkait yang mendukung proses penelitian dan mengumpulan data.

DAFTAR PUSTAKA

Abousaber, I., Abdallah, H. F., & El-Ghaish, H. (2024). Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1499530

Adeleke, O., Adebowale, O. J., Ayoade, I., Olawuyi, I., & Jen, T.-C. (2026). Residential energy consumption dynamics: A SHAP-based interpretation, k-means clustering, and predictive modelling. *Green Energy and Resources*, 100169. https://doi.org/10.1016/j.gerr.2026.100169

Arabani, M., Safari, M., & Rahimabadi, M. M. (2026). SHAP-Based Ensemble Machine Learning for Predicting Hveem-Derived Shear Strength from Marshall Tests. *Results in Engineering*, 109853. https://doi.org/10.1016/j.rineng.2026.109853

Ashisha, G. R., Mary, X. A., Kanaga, E. G. M., Andrew, J., & Eunice, R. J. (2024). Random

Oversampling-Based Diabetes Classification via Machine Learning Algorithms. *International Journal of Computational Intelligence Systems*, *17*(1). https://doi.org/10.1007/s44196-024-00678-3

Ayoade, O. B., Shahrestani, S., & Ruan, C. (2025). Machine Learning and Deep Learning Approaches for Predicting Diabetes Progression: A Comparative Analysis. *Electronics (Switzerland)*, *14*(13). https://doi.org/10.3390/electronics14132583

Elhadad, A., Rayan, A., & Taloba, A. I. (2026). A novel APSGWO-XGBoost framework for heart disease detection: Enhancing diagnostic accuracy and efficiency in healthcare. *Journal of Radiation Research and Applied Sciences*, *19*(1), 102258. https://doi.org/10.1016/j.jrras.2026.102258

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, *113*(7), 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Fachmi, A. N., Dwi Patria, R., Ramadhan, F. R., Pudjiati, S., & Sari, A. P. (2026). *Perbandingan Kinerja Klasifikasi Data Mining untuk Diagnosis Diabetes*. https://ejournal.sisfokomtek.org/index.php/jumin/article/view/8024/4681

Ghufron Syifa, M., Anissa Maori, N., Sucipto, A., & Studi Tekni, P. (2026). *Penerapan SMOTE dan XGBoost untuk Klasifikasi Penyakit Ginjal Kronis pada Data yang Tidak Seimbang*.

Hairani, H., Anggrawan, A., & Priyanto, D. (2023). *Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link*. www.joiv.org/index.php/joiv

Islam, Md. N., Rimon, Md. M. H., Shamim, S. S.-E.-A., Fahad, Z. M., Mony, Md. J. I., & Chowdhury, Md. J. U. (2025). *An Improved Ensemble-Based Machine Learning Model with Feature Optimization for Early Diabetes Prediction*. http://arxiv.org/abs/2512.02023

Jang, Y. (2025). Feature-based ensemble modeling for addressing diabetes data imbalance using the SMOTE, RUS, and random forest methods: a prediction study. *Ewha Medical Journal*, *48*(2), e32. https://doi.org/10.12771/emj.2025.00353

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences (Switzerland)*, *13*(6). https://doi.org/10.3390/app13064006

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, *14*(7), 8459–8486. https://doi.org/10.1007/s12652-021-03612-z

Muhammad, M., Samodro, J., Sakti, M., Kumudasmoro, B. M., Fahmi, M., & Nahdli, M. (2026). *Analisis Pengaruh Ketidakseimbangan Data terhadap Kinerja Model Klasifikasi Penyakit Jantung* (Vol. 6, Number 1). https://ejurnal.umri.ac.id/index.php/SEIS/article/view/11050/4100

Netayawijit, P., Chansanam, W., & Sorn-In, K. (2025). Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support. *Healthcare (Switzerland)*, *13*(20). https://doi.org/10.3390/healthcare13202588

Optarina, Y., Suarna, N., Bahtiar, A., Rahaningsih, N., & Prihartono, W. (2026). OPTIMASI MODEL XGBOOST UNTUK PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN OPTUNA. *Jurnal Software Engineering and Information System (SEIS)*, *6*(1), 50–55. https://ejurnal.umri.ac.id/index.php/SEIS/article/view/10527/4096

Pinem, J., Astuti, W., & Adiwijaya. (2025). Explainable Ensemble Learning Framework with SMOTE, SHAP and LIME for Predicting 30-Day Readmission in Diabetic Patients. *Jurnal RESTI*, *9*(5), 983–990. https://doi.org/10.29207/resti.v9i5.6977

Rafie, Z., Talab, M. S., Koor, B. E. Z., Garavand, A., Salehnasab, C., & Ghaderzadeh, M. (2025). Leveraging XGBoost and explainable AI for accurate prediction of type 2 diabetes. *BMC Public Health*, *25*(1). https://doi.org/10.1186/s12889-025-24953-w

Sadeghi, S., Khalili, D., Ramezankhani, A., Mansournia, M. A., & Parsaeian, M. (2022). Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. *BMC Medical Informatics and Decision Making*, *22*(1). https://doi.org/10.1186/s12911-022-01775-z

Septiana Rizky, P., Haiban Hirzi, R., Hidayaturrohman, U., Hamzanwadi Selong Jl TGKH Muhammad Zainuddin Abdul Madjid Pancor, U., & Timur, L. (2022). Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang. In *J Statistika* (Vol. 15, Number 2). www.unipasby.ac.id

Sir, Y. A., & Soepranoto, A. H. H. (2022). Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas. *Jurnal Komputer Dan Informatika*,

10(1), 31–38. https://doi.org/10.35508/jicon.v10i1.6554

Sirag, A., & Nor, N. M. (2021). Out-of-pocket health expenditure and poverty: Evidence from a dynamic panel threshold analysis. *Healthcare (Switzerland)*, *9*(5). https://doi.org/10.3390/healthcare9050536

Sukamto, T. F., Prameswary, C. L., Royadi, D., & Sofia, D. (2025). Diabetes Disease Prediction on Unbalanced Data Using SMOTE-Tomek Links and Random Forest Algorithm. *G-Tech: Jurnal Teknologi Terapan*, *9*(3), 1194–1203. https://doi.org/10.70609/g-tech.v9i3.7164

Zhang, Y., Deng, L., & Wei, B. (2024). Imbalanced Data Classification Based on Improved Random-SMOTE and Feature Standard Deviation. *Mathematics*, *12*(11). https://doi.org/10.3390/math12111709

Zuhairi, A. H., Yakub, F., Omar, M., Sharifuddin, M., Razak, K. A., & Faruq, A. (2024). Imbalanced Flood Forecast Dataset Resampling Using SMOTE-Tomek Link. *International Exchange and Innovation Conference on Engineering and Sciences*, *10*, 845–850. https://doi.org/10.5109/7323359