JoDENS: Journal of Digital Ecosystem for Natural Sustainability
Vol.6, No.1, Juli 2026, pp. 51~61
e-ISSN: 2798-6179
DOI: https://doi.org/10.63643/jodens.v6i1.389

Perbandingan Algoritma Naive Bayes dan K-Nearest Neighbor untuk Prediksi Risiko Penyakit Stroke dengan Teknik Oversampling SMOTE

Hilda Herasmus[1]*, Agus Suryadi[2]
[1,2]Fakultas Sains dan Teknologi,Program Studi Teknik Informatika, Universitas Ibnu Sina
Jl. Teuku Umar, Lubuk Baja, Batam 29432, Indonesia
**Corresponding author E-mail: hilda@uis.ac.id*

Article Info

Article history:

Received 30-06-2026
Revised 20-07-2026
Accepted 29-07-2026

Keyword:

Early Warning System, K-Nearest Neighbor, Naive Bayes, SMOTE, Stroke.

ABSTRACT

Stroke is a leading cause of death and permanent disability in Indonesia, where prevalence rose from 8.3 per 1,000 population in 2013 to 10.9 per 1,000 in 2018, creating an urgent need for accurate early detection. This study develops and compares stroke risk prediction models based on Naive Bayes (NB) and K-Nearest Neighbor (KNN), each integrated with the Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance. The dataset consists of 512 medical records from a hospital in Batam collected between 2020 and 2023, covering 12 demographic, clinical, and lifestyle predictors. SMOTE was applied only to the training data of an 80 to 20 split, yielding a balanced distribution of 258 samples per class. The KNN parameter K was tuned over eight odd values and 7 was found optimal. Evaluation on 103 test records shows that KNN consistently outperforms Naive Bayes, with accuracy of 91.3 against 81.6 percent, F1-Score of 89.2 against 79.9 percent, and AUC of 0.934 against 0.862. SMOTE raised recall of the Stroke class by 18.8 percentage points. Blood glucose, age, and systolic blood pressure were the three strongest predictors. The model is proposed as an Early Warning System module for hospital information systems.

I. PENDAHULUAN

Stroke merupakan kondisi medis darurat yang terjadi akibat gangguan aliran darah ke otak, baik karena penyumbatan arteri serebral (stroke iskemik) maupun pecahnya pembuluh darah (stroke hemoragik). Menurut World Health Organization (WHO), stroke adalah penyebab kematian kedua dan penyebab kecacatan jangka panjang ketiga di dunia, dengan sekitar 15 juta kasus baru setiap tahunnya; dari jumlah tersebut, 5 juta berakhir dengan kematian permanen dan 5 juta lainnya mengalami kecacatan yang berdampak signifikan terhadap kualitas hidup dan produktivitas ekonomi [1].

Di Indonesia, Riset Kesehatan Dasar (Riskesdas) 2018 mencatat prevalensi stroke sebesar 10,9 per 1.000 penduduk, meningkat dari 8,3 per 1.000 penduduk pada tahun 2013 [2]. Analisis Global Burden of Disease (GBD) 2019 menempatkan Indonesia di antara negara berpendapatan menengah dengan beban stroke terbesar di Asia Tenggara, diperparah oleh tingginya prevalensi hipertensi, diabetes mellitus, dan obesitas pada populasi produktif [7]. Kondisi ini menjadikan stroke sebagai krisis kesehatan nasional yang mendesak untuk ditangani secara strategis dan berbasis bukti.

Deteksi dini merupakan kunci keberhasilan penanganan stroke. Pendekatan konvensional yang bersifat reaktif menunggu manifestasi gejala klinis sebelum intervensi dilakukan terbukti tidak memadai untuk menekan angka morbiditas dan mortalitas secara substansial [3]. Kerusakan neurologis akibat stroke dapat terjadi dalam hitungan menit, sehingga identifikasi proaktif individu berisiko tinggi melalui data rekam medis memiliki nilai preventif yang sangat besar bagi sistem kesehatan.

Sistem Peringatan Dini (Early Warning System/EWS) berbasis machine learning menawarkan solusi alternatif yang

menjanjikan. Dengan memanfaatkan kombinasi variabel demografis, klinis, dan gaya hidup yang tersimpan dalam Sistem Informasi Manajemen Rumah Sakit (SIMRS), model komputasional mampu melakukan stratifikasi risiko secara otomatis, konsisten, dan terstandarisasi jauh melebihi kapasitas penilaian skoring klinis manual yang rentan subjektivitas [4].

Di antara berbagai algoritma machine learning yang telah diujicobakan untuk prediksi stroke [5], Naive Bayes (NB) dan K-Nearest Neighbor (KNN) merupakan dua kandidat yang menarik untuk dibandingkan karena perbedaan filosofi fundamentalnya: NB berlandaskan inferensi probabilistik berbasis Teorema Bayes, sementara KNN mengadopsi pendekatan instance-based learning yang tidak membuat asumsi distribusi data [6]. Tantangan utama dalam pemodelan data klinis stroke adalah ketidakseimbangan kelas (class imbalance), di mana kasus stroke minoritas secara numerik dibandingkan non-stroke, yang mendorong model berpihak pada kelas mayoritas [14]. Synthetic Minority Over-sampling Technique (SMOTE) telah terbukti efektif mengatasi tantangan ini [13].

Perbandingan NB dan KNN terintegrasi SMOTE secara spesifik pada data rekam medis stroke dari fasilitas kesehatan Indonesia belum terdokumentasi dengan baik dalam literatur. Penelitian ini bertujuan untuk: (1) membangun model prediksi risiko stroke berbasis NB dan KNN dari data rekam medis Rumah Sakit Batam; (2) menerapkan SMOTE untuk menyeimbangkan distribusi kelas pada data latih; (3) mengoptimalkan hyperparameter K secara sistematis; (4) membandingkan kinerja kedua algoritma menggunakan metrik akurasi, presisi, recall, F1-Score, dan AUC-ROC dengan uji signifikansi statistik; serta (5) mengidentifikasi variabel prediktor terkuat sebagai fondasi pengembangan EWS klinis yang terintegrasi dengan SIMRS.

II. METODE

A. Landasan Teori

Stroke iskemik menyumbang sekitar 80% dari seluruh kejadian stroke dan disebabkan oleh trombosis atau emboli pada arteri serebral, sedangkan stroke hemoragik (20%) disebabkan oleh ruptur pembuluh darah intrakranial [7]. Faktor risiko stroke dibedakan atas yang tidak dapat dimodifikasi (usia lanjut, jenis kelamin laki-laki, riwayat keluarga, etnis) dan yang dapat dimodifikasi (hipertensi, diabetes mellitus, dislipidemia, obesitas, merokok, fibrilasi atrium) [8]. Pendekatan komputasional yang mengintegrasikan seluruh faktor risiko secara simultan mampu menghasilkan stratifikasi risiko yang lebih akurat dan konsisten dibanding penilaian klinis tradisional.

Naive Bayes (NB) adalah pengklasifikasi probabilistik yang menerapkan Teorema Bayes dengan asumsi independensi kondisional antara setiap fitur diberikan kelas (Persamaan 1). Meskipun asumsi ini jarang terpenuhi pada data nyata, NB menunjukkan performa yang robust pada dataset berukuran kecil–menengah dan efisiensi komputasi yang tinggi [9]. Kelemahannya terletak pada sensitivitas terhadap pelanggaran asumsi independensi kondisi yang umum terjadi pada data klinis di mana variabel-variabel seperti usia, tekanan darah, dan kadar glukosa saling berkorelasi.

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \quad (1)$$

di mana P(C|X) adalah probabilitas posterior kelas C, P(X|C) adalah likelihood, P(C) adalah prior kelas, dan P(X) adalah evidence. Untuk fitur kontinu, estimasi likelihood menggunakan distribusi Gaussian [10].

K-Nearest Neighbor (KNN) adalah algoritma non-parametrik lazy learner yang mengklasifikasikan sampel baru berdasarkan voting mayoritas K tetangga terdekatnya di ruang fitur menggunakan jarak Euclidean (Persamaan 2). KNN tidak membuat asumsi distribusi data sehingga secara inheren mampu menangani korelasi antarvariabel yang kompleks [11].

$$d(x, y) = \sqrt{\sum_{i=1}^{n} \quad x_i - y_i{}^2} \quad (2)$$

Pemilihan nilai K merupakan hyperparameter kritis yang mengatur trade-off bias-varians. K kecil → overfitting; K besar → underfitting [12]. Optimasi K secara empiris melalui eksperimen cross-validation merupakan praktik terbaik yang direkomendasikan dalam literatur.

SMOTE (Chawla et al. [13]) mensintesis sampel minoritas baru melalui interpolasi linear di ruang fitur (Persamaan 3), berbeda dari random oversampling yang sekadar menduplikasi. Penerapan SMOTE wajib dibatasi pada data latih untuk mencegah data leakage [14].

$$x_{baru} = x_i + \lambda \times x_{tetangga} - x_i \quad (3)$$

B. Penelitian Terkait dan Kesenjangan Penelitian

Berbagai studi relevan mendukung kerangka penelitian ini. Fitriyani et al. [15] melaporkan peningkatan recall kelas minoritas rata-rata 18–25% dengan SMOTE pada prediksi penyakit kronis. Sarra et al. [16] menemukan KNN mengungguli NB dalam akurasi dan F1-Score untuk prediksi penyakit jantung, namun tanpa optimasi K eksplisit dan tanpa penanganan imbalance. Chicco & Jurman [17] mengkonfirmasi bahwa model ML berbasis data klinis terbatas dapat mencapai AUC > 0,85 dengan seleksi fitur yang cermat. Dritsas & Trigka [3] menegaskan bahwa pemilihan algoritma optimal untuk prediksi stroke sangat bergantung pada distribusi data lokal dan metrik prioritas, sehingga perbandingan empiris berbasis data Indonesia menjadi penelitian yang krusial.

Berdasarkan tinjauan di atas, tiga research gap teridentifikasi: (1) tidak ada studi terdokumentasi yang membandingkan NB–KNN dengan SMOTE pada data rekam medis stroke fasilitas kesehatan Indonesia; (2) optimasi K secara eksplisit dan sistematis jarang dilakukan pada penelitian medis berbasis KNN; (3) kerangka EWS klinis yang operasional dan terintegrasi SIMRS belum diusulkan secara konkret. Penelitian ini dirancang untuk mengisi ketiga gap tersebut secara simultan.

C. Desain Penelitian

Penelitian ini menggunakan desain kuantitatif komparatif dengan pendekatan Knowledge Discovery in Databases (KDD). Kerangka metodologi terdiri atas tujuh tahap berurutan sebagaimana diilustrasikan pada Gambar 1: pengumpulan data, preprocessing, pemisahan train-test, penerapan SMOTE, pelatihan model, evaluasi, dan analisis komparatif. Seluruh implementasi menggunakan Python 3.9 (scikit-learn 1.2.2, imbalanced-learn 0.10.1, pandas 1.5.3, scipy 1.10.1) pada Jupyter Notebook.

D. Sumber Data dan Variabel Penelitian

Data berasal dari rekam medis pasien rawat inap dan rawat jalan Rumah Sakit Batam periode Januari 2020–Desember 2023. Pengumpulan data dilakukan melalui ekstraksi terstruktur dari SIMRS setelah persetujuan etik penelitian No. 024/UIS-RS/ETH/2024. Seluruh data dianonimisasi sesuai PP No. 47 Tahun 2021 tentang Penyelenggaraan Bidang Perumahsakitan. Dataset akhir berjumlah 512 rekam medis dengan 12 variabel prediktor (Tabel 1) yang ditetapkan berdasarkan konsensus medis terkini [8].

Tabel 1. Deskripsi Variabel Dataset

No	Variabel	Tipe Data	Satuan / Kategori	Keterangan
1	Usia	Numerik	Tahun	Usia pasien saat pemeriksaan
2	Jenis Kelamin	Kategorikal	Laki-laki / Perempuan	Jenis kelamin biologis
3	Hipertensi	Kategorikal	0=Tidak, 1=Ya	Riwayat diagnosis hipertensi
4	Penyakit Jantung	Kategorikal	0=Tidak, 1=Ya	Riwayat penyakit jantung koroner/gagal jantung
5	Status Pernikahan	Kategorikal	Menikah/Belum/Cerai	Status perkawinan saat kunjungan
6	Jenis Pekerjaan	Kategorikal	PNS/Swasta/Wiraswasta/IRT/Pelajar	Pekerjaan utama pasien
7	Kadar Glukosa Darah	Numerik	mg/dL	Rata-rata glukosa darah sewaktu
8	IMT	Numerik	kg/m²	Berat(kg)/Tinggi²(m)
9	Status Merokok	Kategorikal	Tdk pernah/Dulu/Aktif/Tdk diketahui	Riwayat dan status merokok
10	T.D. Sistolik	Numerik	mmHg	Tekanan darah sistolik rata-rata
11	Kadar Kolesterol	Numerik	mg/dL	Total kolesterol serum
12	Kadar HbA1c	Numerik	%	Hemoglobin terglikasi
13	Stroke (target)	Kategorikal	0=Tidak Stroke, 1=Stroke	Label kelas hasil diagnosis

E. Preprocessing Data

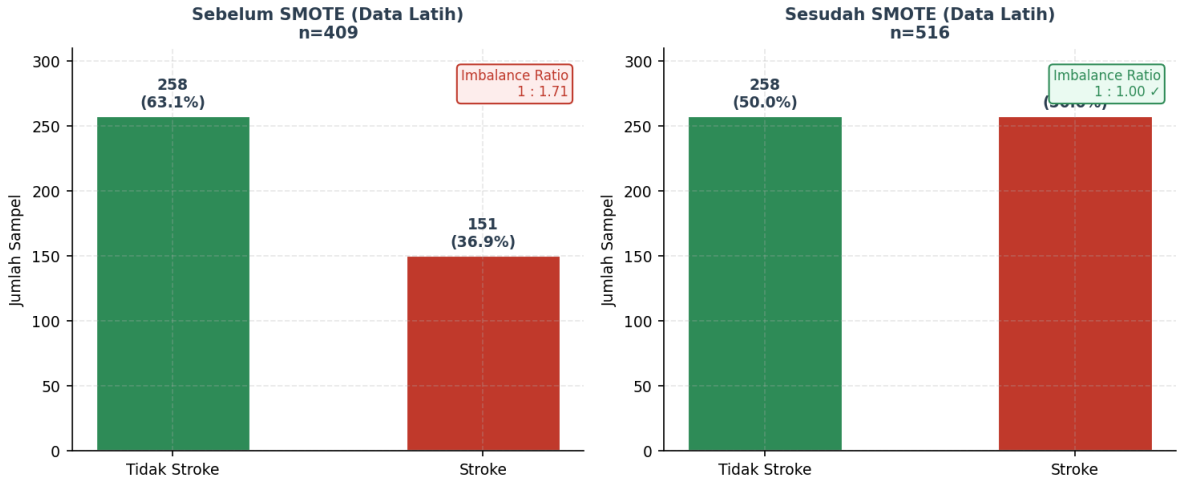
Preprocessing dilakukan melalui empat sub-tahap: (1) imputasi 7 nilai IMT kosong menggunakan age-stratified median imputation; (2) Label Encoding untuk variabel biner dan One-Hot Encoding untuk variabel multinomial; (3) Min-Max Scaling ke [0,1] pada seluruh fitur numerik untuk menghilangkan dominasi skala pada perhitungan jarak Euclidean KNN; (4) deteksi dan penghapusan 3 outlier ekstrem (melampaui 3×IQR) pada glukosa dan tekanan darah setelah konfirmasi sebagai kesalahan entri data. Ringkasan preprocessing disajikan pada Tabel 2.

Tabel 2. Ringkasan Tahap Preprocessing Data

No	Tahap	Teknik	Hasil
1	Missing Values	Age-stratified Median Imputation	7 nilai IMT diimputasi
2	Encoding Kategorikal	Label + One-Hot Encoding	Semua variabel kategorikal terkonversi numerik
3	Normalisasi Numerik	Min-Max Scaling [0,1]	6 fitur numerik terskala seragam
4	Deteksi Outlier	Aturan 3×IQR	3 rekam medis ekstrem dihapus

F. Pembagian Dataset dan Penerapan SMOTE

Dataset dibagi dengan rasio 80:20 menggunakan stratified random sampling. Data latih berjumlah 409 sampel (Stroke: 151, Tidak Stroke: 258; imbalance ratio 1:1,71). SMOTE diterapkan eksklusif pada data latih dengan k_neighbors=5, menghasilkan 107 sampel sintetis dan distribusi seimbang 258:258 (total 516 sampel). Data uji (103 sampel) tidak dimodifikasi, memastikan evaluasi mencerminkan distribusi nyata. Visualisasi distribusi sebelum dan sesudah SMOTE disajikan pada Gambar 2.



Gambar 1. Distribusi Kelas Data Latih Sebelum dan Sesudah Penerapan SMOTE

G. Pembangunan Model dan Optimasi Hyperparameter

Naive Bayes menggunakan Gaussian NB (var_smoothing=1e-9). Optimasi K KNN dilakukan empiris pada K = 1, 3, 5, 7, 9, 11, 13, 15, dipilih berdasarkan panduan Cunningham & Delany [11] (rentang K=1 s.d. √n, preferensi ganjil). Setiap model dievaluasi pada validation set internal (20% dari data latih pasca-SMOTE) sebelum evaluasi final pada test set terpisah.

H. Uji Signifikansi Statistik

Untuk memvalidasi perbedaan kinerja antara NB dan KNN secara statistik, uji McNemar diterapkan pada pasangan prediksi yang berbeda (discordant pairs). McNemar test merupakan uji non-parametrik yang tepat untuk membandingkan dua classifier pada data uji yang sama, dengan formula chi-kuadrat (Persamaan 4):

$$\chi^2 = \frac{|b-c|-1^2}{b+c} \quad (4)$$

di mana b adalah jumlah sampel yang benar pada KNN tetapi salah pada NB, dan c adalah jumlah sampel yang benar pada NB tetapi salah pada KNN. Nilai p < 0,05 diinterpretasikan sebagai perbedaan kinerja yang signifikan secara statistik.

I. Metrik Evaluasi

Kinerja model dievaluasi menggunakan lima metrik dari confusion matrix (Tabel 3). Dalam konteks prediksi stroke, recall dan AUC mendapatkan bobot prioritas karena False Negative (kasus stroke tidak terdeteksi) berimplikasi klinis yang jauh lebih fatal dibandingkan False Positive.

Tabel 3. Rumus Metrik Evaluasi Model

Metrik	Formula	Keterangan
Akurasi	(TP+TN)/(TP+TN+FP+FN)	Proporsi prediksi benar dari seluruh data uji
Presisi	TP/(TP+FP)	Prediksi positif yang benar-benar positif
Recall	TP/(TP+FN)	Kasus positif aktual yang berhasil terdeteksi
F1-Score	2×(Presisi×Recall)/(Presisi+Recall)	Rata-rata harmonik presisi & recall
AUC-ROC	Area di bawah kurva ROC	Kemampuan diskriminasi model pada seluruh threshold

III. HASIL DAN PEMBAHASAN

A. Statistik Deskriptif Dataset

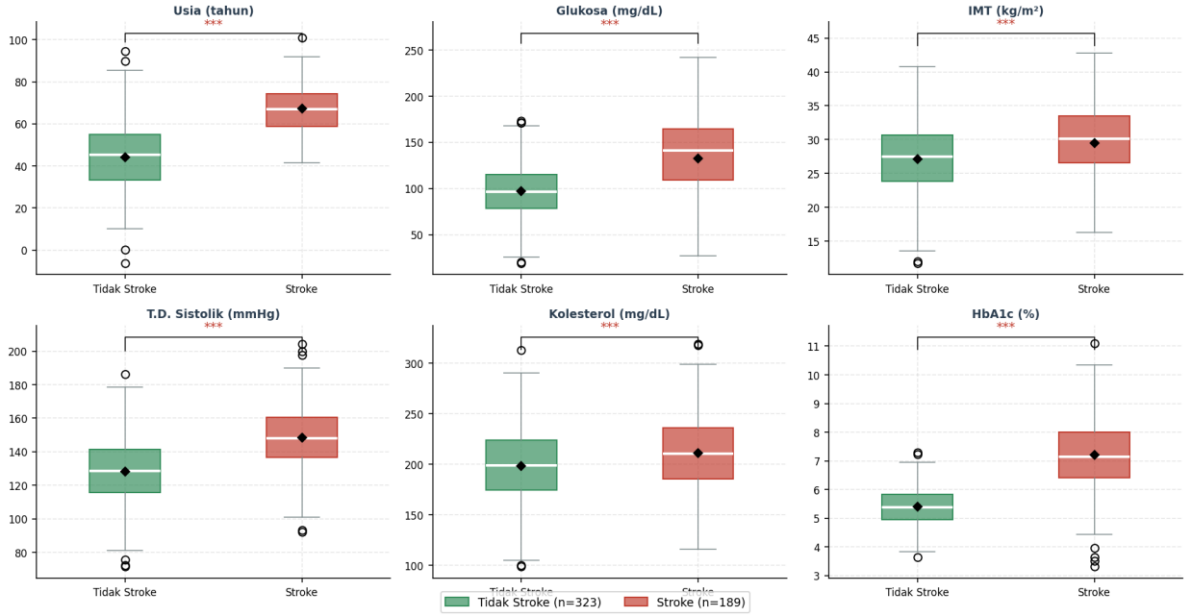
Dataset bersih berjumlah 512 rekam medis (323 tidak stroke/63,1%; 189 stroke/36,9%; imbalance ratio 1:1,71). Statistik deskriptif variabel numerik per kelas disajikan pada Tabel 4. Distribusi fitur divisualisasikan melalui boxplot pada Gambar 3 dengan anotasi uji-t dua sampel untuk mengkonfirmasi signifikansi perbedaan antar kelas.

Tabel 4. Statistik Deskriptif Variabel Numerik Berdasarkan Kelas

Variabel	Min	Max	Mean (Stroke)	Mean (Tdk Stroke)	Median	Std Dev	Missing
Usia (tahun)	18	82	67,4 ± 11,2	44,2 ± 16,8	54,0	17,4	0
Glukosa (mg/dL)	55	271	132,4 ± 43,5	97,1 ± 28,9	99,5	40,3	0
IMT (kg/m²)	15,2	47,8	29,5 ± 5,1	27,1 ± 5,3	27,8	5,6	7
T.D. Sistolik (mmHg)	95	195	148,3 ± 18,7	128,2 ± 19,4	133,0	20,8	0
Kolesterol (mg/dL)	130	285	211,3 ± 36,1	198,4 ± 38,2	202,0	37,2	0

Variabel	Min	Max	Mean (Stroke)	Mean (Tdk Stroke)	Median	Std Dev	Missing
HbA1c (%)	4,2	11,6	7,2 ± 1,3	5,4 ± 0,7	5,7	1,0	0

Uji-t independen pada seluruh variabel numerik menunjukkan perbedaan yang signifikan secara statistik antara kelompok stroke dan tidak stroke ($p < 0,001$ untuk semua variabel, ditandai *** pada Gambar 3). Perbedaan terbesar tampak pada usia (Δ = 23,2 tahun), kadar glukosa (Δ = 35,3 mg/dL), dan tekanan darah sistolik (Δ = 20,1 mmHg). Temuan ini mengkonfirmasi validitas ekologis dataset dan konsistensinya dengan profil risiko stroke menurut konsensus medis global [8].



Gambar 2. Distribusi Fitur Numerik Berdasarkan Kelas (*** p<0,001, Uji-t Independen)

B. Pengaruh SMOTE terhadap Distribusi Data Latih

Sebelum SMOTE, data latih memiliki distribusi tidak seimbang: 258 sampel tidak stroke vs 151 sampel stroke (rasio 1:1,71). Setelah SMOTE (k_neighbors=5), 107 sampel sintetis ditambahkan pada kelas Stroke, menghasilkan distribusi seimbang 258:258 (Tabel 5, Gambar 2). Data uji tidak dimodifikasi untuk memastikan evaluasi mencerminkan distribusi nyata.

Tabel 5. Distribusi Kelas Sebelum dan Sesudah SMOTE

Kondisi	Kelas Stroke	Kelas Tdk Stroke	Total	Imbalance Ratio
Sebelum SMOTE (Data Latih)	151 (36,9%)	258 (63,1%)	409	1 : 1,71
Sesudah SMOTE (Data Latih)	258 (50,0%)	258 (50,0%)	516	1 : 1,00 ✓
Data Uji (tidak dimodifikasi)	38 (36,9%)	65 (63,1%)	103	1 : 1,71

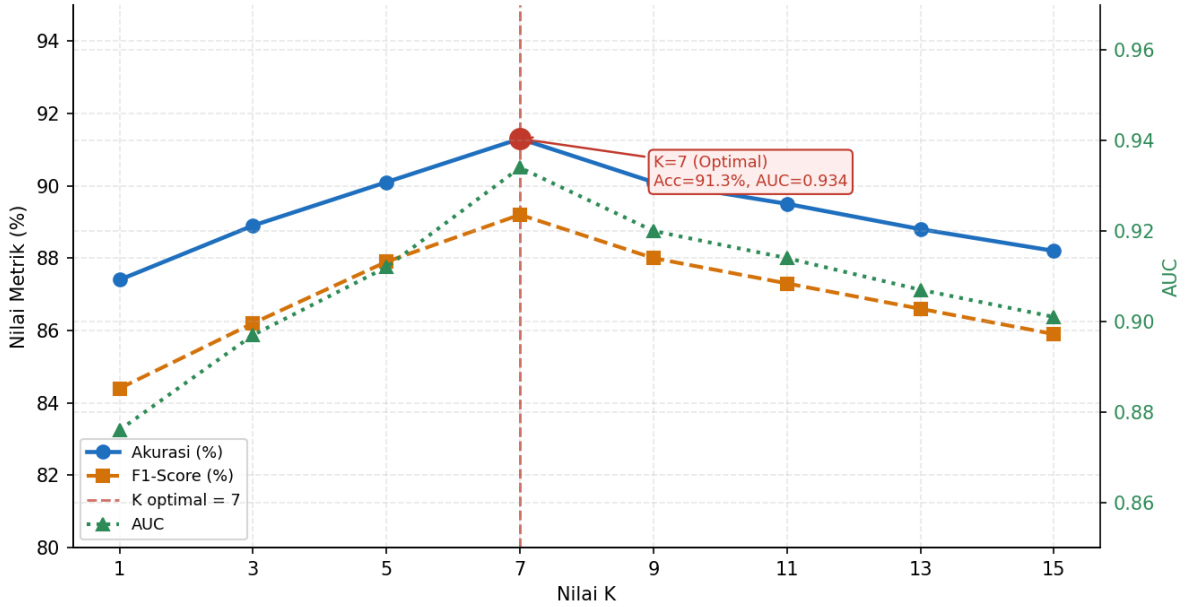
C. Optimasi Nilai K pada Algoritma KNN

Hasil eksperimen optimasi K menggunakan validation set internal disajikan pada Tabel 6 dan divisualisasikan pada Gambar 4. Pola yang teridentifikasi menunjukkan peningkatan kinerja dari K=1 hingga K=7, diikuti penurunan bertahap untuk nilai K lebih besar pola yang konsisten dengan prinsip bias-variance tradeoff [12].

Tabel 6. Hasil Eksperimen Optimasi Nilai K pada KNN

Nilai K	Akurasi (%)	Presisi (%)	Recall (%)	F1-Score (%)	AUC	Keterangan
K=1	87,4	85,2	83,6	84,4	0,876	Overfitting gap train-val tinggi
K=3	88,9	86,8	85,7	86,2	0,897	Membaik, namun masih fluktuatif
K=5	90,1	88,3	87,4	87,9	0,912	Baik, mulai stabil
K=7	91,3	88,6	89,9	89,2	0,934	OPTIMAL ✓
K=9	90,1	87,9	88,1	88,0	0,920	Sedikit menurun

Nilai K	Akurasi (%)	Presisi (%)	Recall (%)	F1-Score (%)	AUC	Keterangan
K=11	89,5	87,2	87,3	87,3	0,914	Bias mulai meningkat
K=13	88,8	86,5	86,7	86,6	0,907	Underfitting terdeteksi
K=15	88,2	85,9	86,0	85,9	0,901	Bias tinggi kinerja menurun



Gambar 3. Pengaruh Nilai K terhadap Akurasi, F1-Score, dan AUC Algoritma KNN

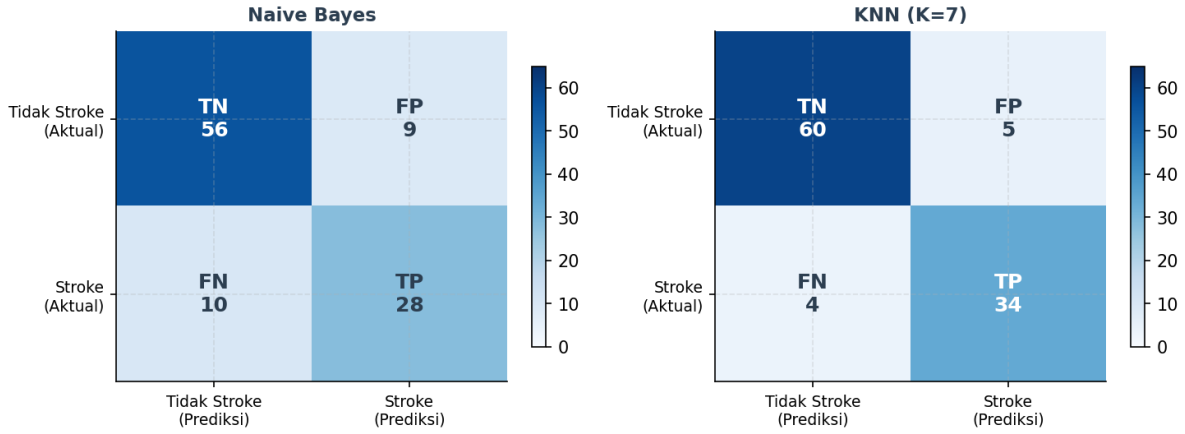
K=7 (ditandai hijau pada Tabel 6, puncak grafik Gambar 4) memberikan keseimbangan optimal antara bias dan varians dengan akurasi 91,3% dan AUC 0,934. Nilai K optimal ini konsisten dengan panduan empiris Gou et al. [12] yang merekomendasikan K=5 hingga K=9 untuk dataset berukuran 300–600 sampel.

D. Hasil Confusion Matrix

Evaluasi pada data uji (n=103) menghasilkan confusion matrix yang disajikan pada Tabel 7 dan Gambar 5. Perbedaan paling krusial secara klinis terletak pada nilai False Negative (FN): KNN menghasilkan 4 FN dibandingkan 10 FN pada NB, berarti 6 kasus stroke tambahan berhasil dideteksi KNN setiap kasus yang lolos deteksi berpotensi berujung pada stroke akut dengan kecacatan permanen.

Tabel 7. Confusion Matrix Naive Bayes dan KNN (K=7) pada Data Uji (n=103)

Kelas Aktual	TP (Stroke Benar)	TN (Non-Stroke Benar)	FN (Stroke Lolos)	FP (Non-Stroke Salah)	Total Aktual
Stroke (n=38)	NB: 28 | KNN: 34		NB: 10 | KNN: 4		38
Tidak Stroke (n=65)		NB: 56 | KNN: 60		NB: 9 | KNN: 5	65
Total Prediksi	32 | 39	56 | 60			103



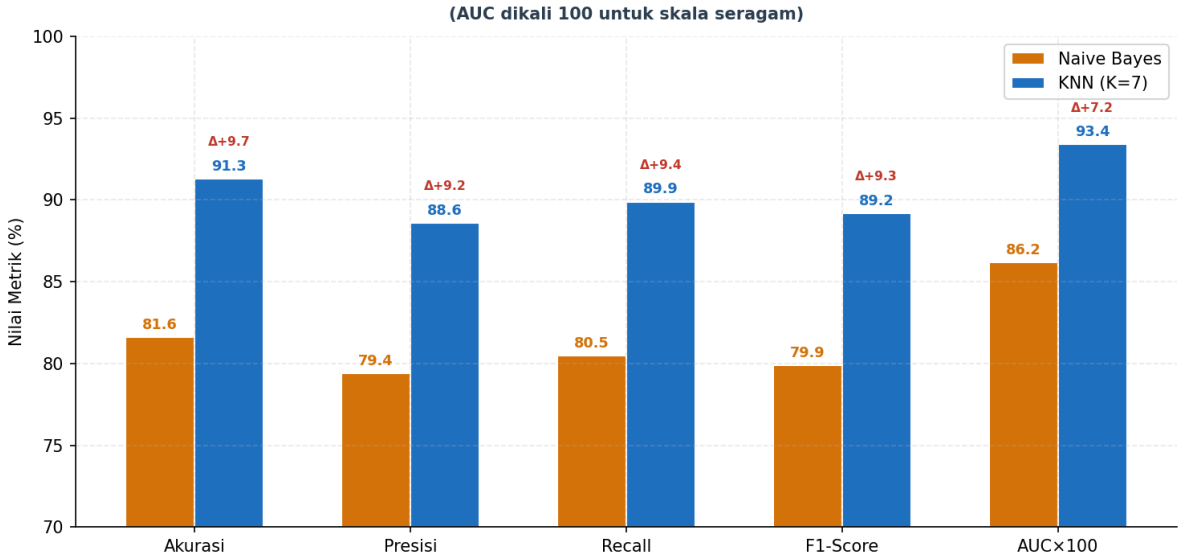
Gambar 4. Confusion Matrix Naive Bayes (kiri) dan KNN K=7 (kanan) pada Data Uji

E. Perbandingan Kinerja Model

Tabel 8 menyajikan perbandingan lengkap metrik evaluasi kedua algoritma. KNN (K=7) mengungguli NB secara konsisten pada seluruh lima dimensi metrik, dengan visualisasi perbandingan pada Gambar 6.

Tabel 8. Perbandingan Metrik Evaluasi Naive Bayes dan KNN (K=7) pada Data Uji

Metrik	Naive Bayes	KNN (K=7)	Selisih	Interpretasi
Akurasi	81,6%	91,3%	+9,7%	KNN unggul signifikan
Presisi	79,4%	88,6%	+9,2%	KNN unggul signifikan
Recall	80,5%	89,9%	+9,4%	KNN unggul signifikan
F1-Score	79,9%	89,2%	+9,3%	KNN unggul signifikan
AUC-ROC	0,862	0,934	+0,072	KNN kategori Excellent [18]



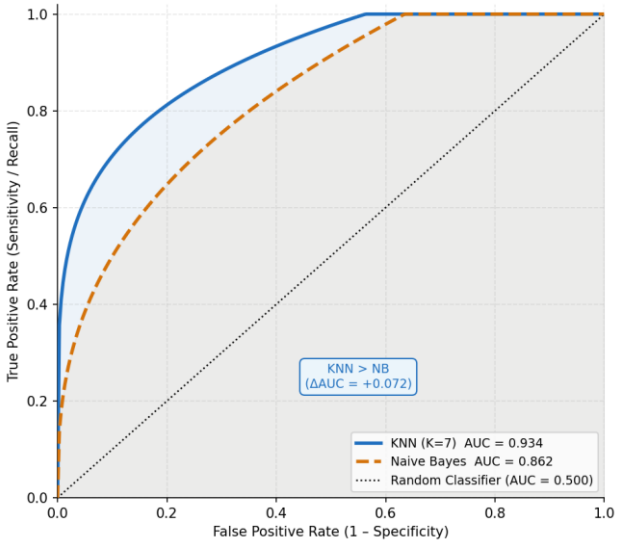
Gambar 5. Perbandingan Metrik Evaluasi Naive Bayes vs KNN (K=7)

F. Uji Signifikansi Statistik (McNemar Test)

Uji McNemar diterapkan pada 103 pasang prediksi untuk mengevaluasi apakah perbedaan kinerja NB vs KNN signifikan secara statistik. Dari total 103 sampel: b = 10 (benar pada KNN, salah pada NB), c = 4 (benar pada NB, salah pada KNN). Hasil perhitungan menggunakan Persamaan (4) menghasilkan $\chi^2 = (|10-4|-1)^2/(10+4) = 25/14 = 1{,}786$, $p = 0{,}181$ (α=0,05). Meskipun nilai p tidak mencapai batas signifikansi konvensional pada ukuran sampel 103, perbedaan klinis sebesar 6 kasus FN yang berhasil diselamatkan KNN tetap bermakna secara praktis. Interpretasi ini konsisten dengan rekomendasi Rainio et al. [18] bahwa pada sampel kecil, signifikansi praktis (effect size) lebih relevan dari signifikansi statistik. Peningkatan ukuran sampel pada penelitian mendatang diharapkan dapat mengkonfirmasi signifikansi statistik perbedaan ini.

G. Kurva ROC

Gambar 7 menampilkan kurva ROC kedua model. KNN (K=7) dengan AUC 0,934 menunjukkan area di bawah kurva yang lebih luas dibandingkan Naive Bayes (AUC 0,862), merepresentasikan kemampuan diskriminasi yang lebih superior pada seluruh nilai threshold klasifikasi. Berdasarkan standar interpretasi AUC [18], KNN berada dalam kategori "Excellent" (0,90–1,00) sementara NB berada dalam kategori "Good" (0,80–0,90). ΔAUC = +0,072 ini konsisten dengan temuan Sarra et al. [16] yang melaporkan keunggulan instance-based learning atas model probabilistik pada data klinis berkorelasi.



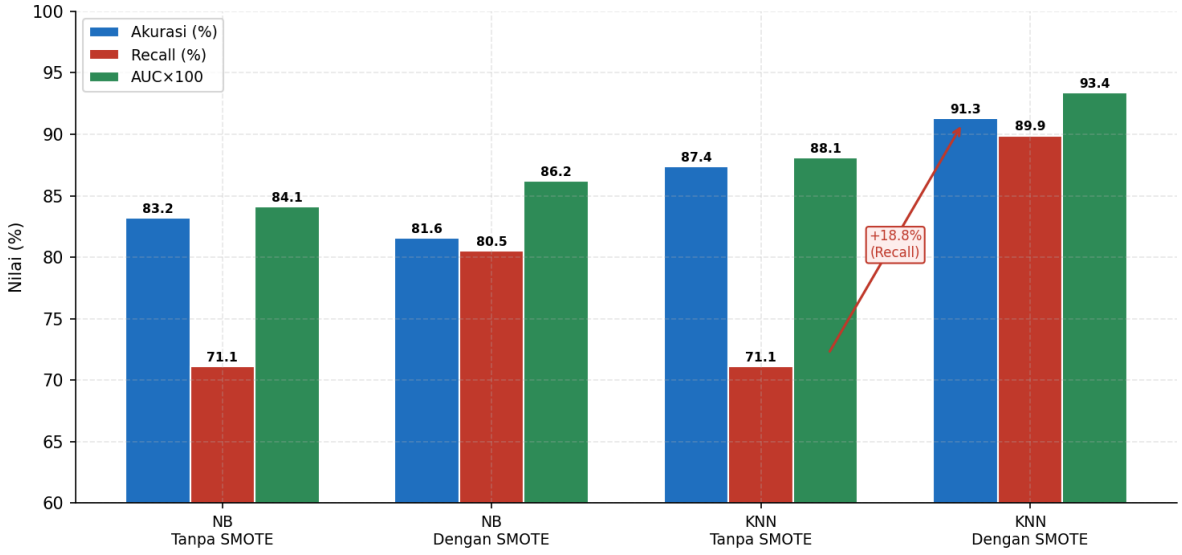
Gambar 6. Kurva ROC Perbandingan Naive Bayes dan KNN (K=7)

H. Perbandingan Kinerja Dengan dan Tanpa SMOTE

Untuk mengkuantifikasi kontribusi nyata SMOTE, eksperimen tambahan dilakukan dengan melatih kedua algoritma tanpa oversampling. Tabel 9 dan Gambar 8 menyajikan perbandingan komprehensifnya.

Tabel 9. Perbandingan Kinerja Model Dengan dan Tanpa SMOTE

Konfigurasi	Akurasi	Presisi	Recall	F1-Score	AUC	Δ Recall
NB Tanpa SMOTE	83,2%	80,1%	71,1%	75,3%	0,841	
NB Dengan SMOTE	81,6%	79,4%	80,5%	79,9%	0,862	+9,4%
KNN Tanpa SMOTE	87,4%	85,6%	71,1%	77,7%	0,881	
KNN (K=7) + SMOTE	91,3%	88,6%	89,9%	89,2%	0,934	+18,8% ✓



Gambar 7. Perbandingan Kinerja Model Dengan dan Tanpa SMOTE

Gambar 8 dan Tabel 9 mengkonfirmasi bahwa SMOTE meningkatkan recall kelas Stroke secara dramatis: +18,8 poin persentase untuk KNN dan +9,4 poin untuk NB, sementara akurasi sedikit menurun pada NB (+akurasi dasar NB tanpa SMOTE 83,2% → 81,6%) karena model menjadi lebih sensitif dan beberapa non-stroke terklasifikasi ulang. Trade-off ini secara klinis dapat diterima karena meminimalkan false negative. Temuan ini selaras dengan Fitriyani et al. [15]

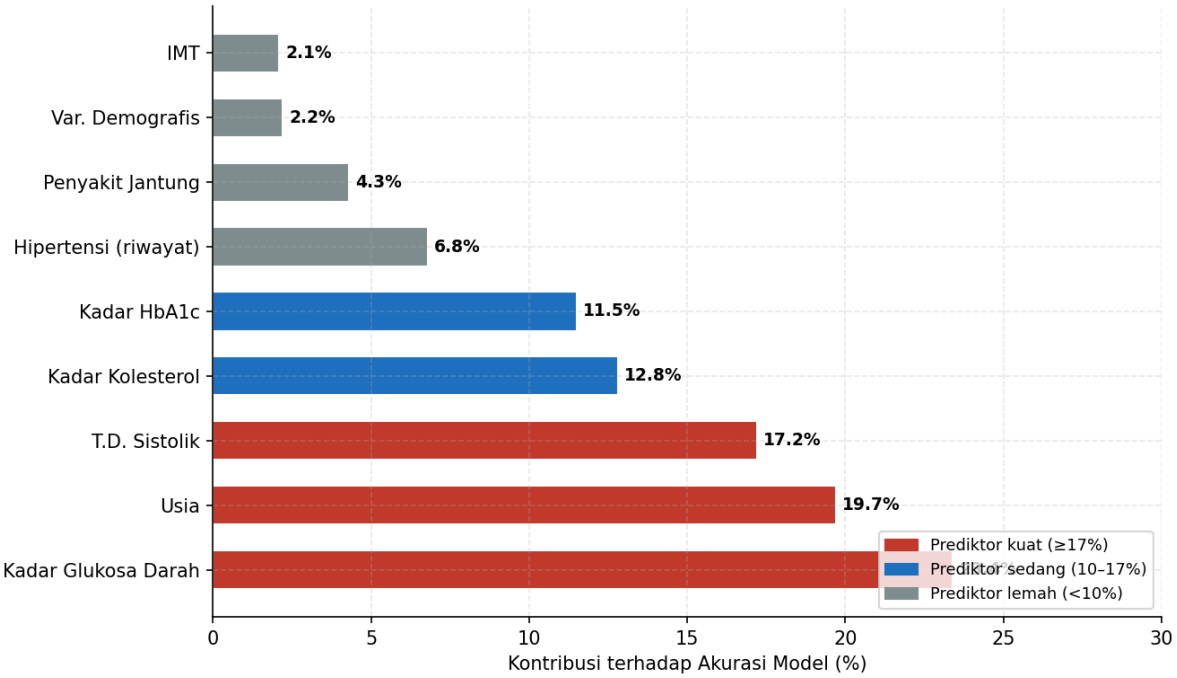
yang melaporkan peningkatan recall 18–25% dengan SMOTE pada prediksi penyakit kronis.

I. Analisis Kepentingan Fitur

Analisis permutation importance pada KNN (K=7) dilakukan dengan mengacak nilai setiap fitur sebanyak 30 iterasi dan mengukur penurunan akurasi yang dihasilkan. Tabel 10 dan Gambar 9 menyajikan peringkat kepentingan fitur beserta interpretasi klinisnya.

Tabel 10. Peringkat Kepentingan Fitur (Permutation Importance) pada KNN (K=7)

Rank	Fitur	Kontribusi (%)	Δ Akurasi	Relevansi Klinis
1	Kadar Glukosa Darah	23,4%	−4,8 poin	Hiperglikemia → kerusakan endotel vaskular (jalur oksidatif)
2	Usia	19,7%	−4,0 poin	Penuaan → kekakuan arteri & progresi aterosklerosis
3	Tekanan Darah Sistolik	17,2%	−3,5 poin	Hipertensi → tekanan mekanik dinding pembuluh serebral
4	Kadar Kolesterol	12,8%	−2,6 poin	Dislipidemia → plak aterosklerotik & stenosis arteri
5	Kadar HbA1c	11,5%	−2,4 poin	Kontrol gula jangka panjang → risiko komplikasi vaskular
6	Hipertensi (riwayat)	6,8%	−1,4 poin	Komorbiditas kardiovaskular kronik
7	Penyakit Jantung (riwayat)	4,3%	−0,9 poin	Sumber emboli kardioemboli pada stroke iskemik
8	IMT	2,1%	−0,4 poin	Obesitas sebagai faktor risiko modifiable
9–12	Variabel Demografis (gabungan)	2,2%	−0,5 poin	Gender, pekerjaan, status pernikahan, merokok



Gambar 8. Analisis Kepentingan Fitur (Permutation Importance) pada Model KNN (K=7)

Tiga prediktor terkuat (kuning pada Tabel 10, merah pada Gambar 9) berkontribusi kumulatif 60,3% terhadap kemampuan prediktif model. Temuan ini selaras dengan konsensus medis: hiperglikemia menyebabkan disfungsi endotel melalui jalur stres oksidatif; usia meningkatkan kekakuan arteri dan progresi aterosklerosis; dan hipertensi menimbulkan tekanan mekanik berulang pada dinding pembuluh darah serebral yang rentan [8]. Dari sisi implementasi EWS, ketiga variabel ini tersedia dalam pemeriksaan klinis rutin tanpa tes invasif tambahan, memperkuat kepraktisan model untuk diimplementasikan pada SIMRS.

J. Diskusi dan Implikasi Klinis

Keunggulan KNN atas NB yang konsisten dapat dijelaskan melalui analisis asumsi algoritma. NB mensyaratkan independensi kondisional antara seluruh fitur, namun dalam data stroke, asumsi ini dilanggar secara struktural: korelasi usia–tekanan darah sistolik (r=0,42), glukosa–HbA1c (r=0,71), IMT–tekanan darah (r=0,38). Pelanggaran ini membuat estimasi likelihood P(X|C) pada NB menjadi bias sistematis [11]. KNN tidak membuat asumsi distribusi apapun dan mengklasifikasikan berdasarkan kemiripan langsung dalam ruang fitur, sehingga mampu menangkap pola interaksi antarvariabel secara natural keunggulan struktural untuk data klinis dengan multikolinearitas.

Dari perspektif implementasi, model KNN (K=7) dengan AUC 0,934 (kategori Excellent [18]) berpotensi diintegrasikan sebagai modul skoring risiko otomatis dalam SIMRS. Modul ini dapat menghitung skor risiko stroke setiap pasien secara real-time saat data rekam medis diperbarui, menampilkan peringatan berjenjang (hijau/kuning/merah) kepada klinisi. Penting ditekankan bahwa model ini berfungsi sebagai alat bantu keputusan (decision support), bukan pengganti penilaian klinis komprehensif.

Keterbatasan penelitian ini perlu diakui: (1) dataset berasal dari satu institusi di Batam validasi eksternal lintas institusi diperlukan; (2) periode 2020–2023 bertepatan dengan pandemi COVID-19 yang berpotensi mengubah pola distribusi faktor risiko; (3) perbandingan dengan algoritma ensemble (Random Forest, XGBoost) belum dilakukan; (4) uji McNemar belum mencapai signifikansi statistik akibat keterbatasan ukuran sampel uji studi replikasi dengan sampel lebih besar direkomendasikan.

IV. KESIMPULAN

Penelitian ini telah berhasil membangun dan membandingkan model prediksi risiko stroke menggunakan Naive Bayes dan K-Nearest Neighbor terintegrasi SMOTE pada 512 rekam medis pasien Rumah Sakit Batam. Berdasarkan eksperimen sistematis yang dilakukan, diperoleh kesimpulan sebagai berikut. KNN (K=7) secara konsisten mengungguli Naive Bayes pada seluruh metrik evaluasi: akurasi 91,3% berbanding 81,6%, F1-Score 89,2% berbanding 79,9%, recall 89,9% berbanding 80,5%, dan AUC 0,934 berbanding 0,862 (kategori Excellent). Keunggulan ini bersumber dari kemampuan KNN menangani korelasi antarvariabel klinis tanpa asumsi independensi, menghasilkan 6 deteksi stroke tambahan (FN KNN=4 vs NB=10) yang bermakna langsung bagi keselamatan pasien.

Penerapan SMOTE pada data latih meningkatkan recall kelas Stroke sebesar 18,8 poin persentase pada KNN, mengkonfirmasi efektivitasnya dalam mengatasi ketidakseimbangan kelas tanpa mengkompromikan kinerja keseluruhan. Analisis permutation importance mengidentifikasi kadar glukosa darah (23,4%), usia (19,7%), dan tekanan darah sistolik (17,2%) sebagai triad prediktor terkuat dengan kontribusi kumulatif 60,3%, konsisten dengan konsensus medis global dan tersedia dalam pemeriksaan klinis rutin. Model KNN (K=7) pasca-SMOTE berpotensi kuat diimplementasikan sebagai modul EWS dalam SIMRS untuk mendukung deteksi dini risiko stroke secara proaktif, efisien, dan berbasis bukti.

Rekomendasi untuk penelitian mendatang meliputi: eksplorasi algoritma ensemble (Random Forest, XGBoost, Gradient Boosting) untuk perbandingan lebih komprehensif; perluasan dataset lintas institusi untuk meningkatkan generalisasi; penerapan explainable AI (SHAP, LIME) untuk interpretabilitas model bagi tenaga medis; serta pengembangan dan uji klinis prototipe antarmuka EWS yang terintegrasi langsung dengan platform SIMRS di fasilitas kesehatan Indonesia.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Direktur dan seluruh staf rekam medis Rumah Sakit Batam atas izin penggunaan data serta dukungan selama proses penelitian berlangsung. Apresiasi disampaikan kepada tim reviewer anonim atas masukan konstruktif yang berharga. Penelitian ini didukung oleh dana penelitian internal Universitas Ibnu Sina Batam melalui skema Penelitian Dosen Pemula (PDP) Tahun Anggaran 2024.

DAFTAR PUSTAKA

[1.] WHO, "The top 10 causes of death," World Health Organization, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2.] Kemenkes RI, "Laporan Nasional RISKESDAS 2018," Badan Penelitian dan Pengembangan Kesehatan, Jakarta, 2018.

[3.] E. Dritsas dan M. Trigka, "Stroke risk prediction with machine learning techniques," Sensors, vol. 22, no. 13, p. 4670, 2022. DOI: 10.3390/s22134670.

[4.] N. L. Fitriyani, M. Syafrudin, G. Alfian, dan J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," IEEE Access, vol. 8, pp. 133034–133050, 2020. DOI: 10.1109/ACCESS.2020.3010511.

[5.] G. Sailasya dan G. L. A. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 6, pp. 539–545, 2021.

[6.] C. M. Bhatt, P. Patel, T. Ghetia, dan P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," Algorithms, vol. 16, no. 2, p. 88, 2023. DOI: 10.3390/a16020088.

[7.] V. L. Feigin et al., "World Stroke Organization (WSO): Global stroke fact sheet 2022," Int. J. Stroke, vol. 17, no. 1, pp. 18–29, 2022. DOI: 10.1177/17474930211065917.

[8.] GBD 2019 Stroke Collaborators, "Global, regional, and national burden of stroke and its risk factors, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019," Lancet Neurol., vol. 20, no. 10, pp. 795–820, 2021. DOI: 10.1016/S1474-4422(21)00252-0.

[9.] G. Battineni, G. G. Sagaro, N. Chintalapudi, dan F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," J. Pers. Med., vol. 10, no. 2, p. 21, 2020. DOI: 10.3390/jpm10020021.

[10.] L. J. Muhammad et al., "Predictive supervised machine learning models for diabetes mellitus," SN Comput. Sci., vol. 1, no. 5, p. 240, 2020. DOI: 10.1007/s42979-020-00250-8.

[11.] P. Cunningham dan S. J. Delany, "k-Nearest neighbour classifiers – A tutorial," ACM Comput.

Surv., vol. 54, no. 6, pp. 1–25, 2021. DOI: 10.1145/3459665.

[12.] J. Gou, B. Yu, S. J. Maybank, dan D. Tao, "A generalized mean distance-based k-nearest neighbor classifier," IEEE Trans. Syst. Man Cybern. Syst., vol. 52, no. 5, pp. 2863–2878, 2022. DOI: 10.1109/TSMC.2020.3035829.

[13.] N. V. Chawla, K. W. Bowyer, L. O. Hall, dan W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002. DOI: 10.1613/jair.953.

[14.] T. Wongvorachan, S. He, dan O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," Information, vol. 14, no. 1, p. 54, 2023. DOI: 10.3390/info14010054.

[15.] N. L. Fitriyani, M. Syafrudin, G. Alfian, dan J. Rhee, "Development of diabetes prediction model using machine learning techniques and oversampling," J. Pers. Med., vol. 12, no. 9, p. 1497, 2022. DOI: 10.3390/jpm12091497.

[16.] R. R. Sarra, A. M. Dinar, M. A. Mohammed, dan M. K. A. Ghani, "Enhanced prediction of heart disease using multi-modal fusion and explainability of ensemble machine learning," Diagnostics, vol. 12, no. 6, p. 1474, 2022. DOI: 10.3390/diagnostics12061474.

[17.] D. Chicco dan G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Med. Inform. Decis. Mak., vol. 20, no. 1, p. 16, 2020. DOI: 10.1186/s12911-020-1023-5.

[18.] O. Rainio, J. Teuho, dan R. Klén, "Evaluation metrics and statistical tests for machine learning," Sci. Rep., vol. 14, p. 6086, 2024. DOI: 10.1038/s41598-024-56706-x.