Jurnal Informatika: Jurnal pengembangan IT
Vol. 10, No. 4, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i4.8809

Analisis Pengaruh SMOTE terhadap Kinerja Model KNN untuk Prediksi Risiko Stroke

Cinantya Paramita[1], Calvin Samuel Simbolon[2], Azriel Sebastian Pamungkas[3], Justin Matthew Triono[4], Emanuel Pinesthi Widi Utomo[5], Egia Rosi Subhiyakto[6]
[1,2,3,4,5,6] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jl. Imam Bonjol, Pendrikan Kidul, Semarang, 50131, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-05-22
Revised 2025-09-08
Accepted 2025-09-16

Corresponding Author:
Cinantya Paramita
Email:
cinantya.paramita@dsn.dinus.ac.id

Abstract *– This study investigates the challenge of data imbalance in stroke risk classification, where the proportion of non-stroke cases is considerably smaller than that of stroke cases. Such imbalance often causes bias toward the majority class, thereby reducing the reliability of classification models. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized to generate synthetic data for the minority class, while the K-Nearest Neighbor (KNN) algorithm was applied as the classification method. The research workflow involved data preprocessing, application of SMOTE, model training, and performance evaluation using widely recognized metrics including accuracy, precision, recall, and F1-score. Experimental results show that integrating SMOTE with KNN improved classification performance, achieving 91.87% accuracy, 94.27% precision, 89.20% recall, and a 91.66% F1-score. These findings demonstrate that the proposed approach is effective in handling class imbalance and provides reliable detection of stroke risk. The contribution of this research lies in presenting a comparative perspective on the role of SMOTE in medical datasets, while also emphasizing its potential to support the development of more robust early detection systems and contribute to better healthcare services in the future..*

Keywords: *KNN, Machine Learning, SMOTE, Stroke Prediction*

Abstrak *– Penelitian ini membahas masalah ketidakseimbangan data dalam klasifikasi risiko stroke, di mana kasus non-stroke secara signifikan lebih rendah daripada kasus stroke. Ketidakseimbangan kelas cenderung menimbulkan bias terhadap kelas mayoritas, yang menyebabkan berkurangnya efektivitas klasifikasi. Untuk mengatasi hal ini, SMOTE (Synthetic Minority Over-sampling Technique) digunakan untuk mengatasi ketidakseimbangan kelas dalam dataset dan algoritma K-Nearest Neighbor (KNN) digunakan untuk klasifikasi. Dataset mengalami preprocessing, aplikasi SMOTE, dan algoritma KNN dilatih dan dievaluasi menggunakan metrik standar termasuk akurasi, presisi, recall, dan F1-score. Penerapan SMOTE bersama dengan KNN menghasilkan peningkatan yang signifikan dalam hasil klasifikasi, mencapai akurasi 91,87%, presisi 94,27%, recall 89,20%, dan F1-score 91,66%. Temuan ini menegaskan bahwa pendekatan yang diimplementasikan berkinerja dengan baik dalam mendeteksi risiko stroke, meskipun set data yang digunakan memiliki ketidaksimbangan data di dalamnya. Tujuan dari penelitian ini adalah untuk menginformasikan kemajuan teknologi deteksi dini stroke yang lebih kuat dan mendukung peningkatan dalam penyediaan layanan kesehatan.*

Kata Kunci*: KNN, Pembelajaran Mesin, Prediksi Stroke, SMOTE*

I. PENDAHULUAN

Penyakit stroke dapat terjadi yang disebabkan oleh masalah dalam peredarah darah di otak yang mengakibatkan masalah pada jaringan otak. Sehingga penderita penyakit Stroke dapat mengalami kelumpuhan dan *defisit neurologis* yang menyerang secara mendadak [3]. Stroke merupakan penyakit berbahaya yang bisa menjangkit kapan saja dan mengakibatkan kerusakan pada otak dengan cepat dan progresif. Data kontrol atau pemeriksaan ulang penderita stroke di Indonesia, ke fasilitas pelayanan yaitu rutin sebanyak 39,4%, tidak rutin 38,7%, tidak memeriksa ulang 21,9% [20].

Setiap tahunnya, stroke telah menyebabkan kecacatan bagi lebih dari 12 juta orang di dunia [1]. Stroke terjadi ketika pembuluh darah tersumbat, mengganggu pasokan darah ke otak. Penyakit saraf ini termasuk yang paling banyak ditemui dan dapat berakibat fatal, menyebabkan kecacatan permanen atau bahkan kematian [6]. Stroke telah menyebabkan 5,5 juta orang meninggal dari 13,7 juta kasus setiap tahunnya, Hal tersebut menjadikan stroke termasuk dalam kategori penyakit mematikan di dunia. [5]. Berbagai aspek risiko seperti hipertensi, diabetes, obesitas, dan gaya hidup tidak sehat turut berkontribusi terhadap peningkatan angka kejadian stroke di berbagai negara [4].

Di Indonesia, stroke telah masuk dalam kategori penyakit yang mematikan. [27]. Menurut data pada tahun 2023 oleh Survei Kesehatan Indonesia (SKI) menunjukkan bahwa dari 1.000 penduduk Indonesia, kejadian

stroke telah menggapai angka 8,3. Angka ini mengindikasikan tingginya tingkat kejadian stroke di kalangan masyarakat, dengan tingkat kematian akibat stroke sebesar 18,5% dari total kematian [29]. Stroke adalah penyakit *catastrophic* yang membebani sistem kesehatan dengan biaya tinggi. Pada tahun 2023, biaya penanganan stroke diperkirakan mencapai Rp 5,2 triliun, menjadikannya penyakit dengan pembiayaan tertinggi ketiga setelah kanker dan jantung. [30]. Pemerintah melalui Kementerian Kesehatan telah berupaya meningkatkan deteksi dini penyakit ini, termasuk dengan meningkatkan skrining faktor risiko seperti hipertensi dan diabetes. Namun, upaya ini masih memiliki berbagai keterbatasan seperti akses dan sumber daya terhadap teknologi kesehatan yang memadai.

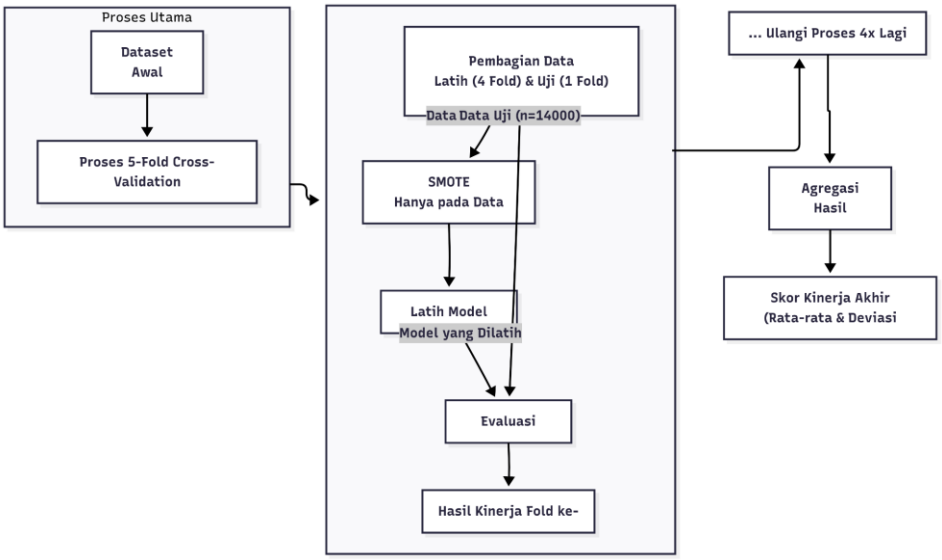
Stroke dapat menyerang ketika suplai darah ke otak mengalami gangguan secara tiba - tiba. Untuk mencegah kecacatan dan kompilasi lebih lanjut, stroke memerlukan penanganan yang cepat dan tepat [20]. Untuk mendukung hal tersebut, kemajuan teknologi kecerdasan buatan (*AI*) telah berhasil membuka peluang untuk meningkatkan deteksi dini stroke [22]. Memanfaatkan sistem cerdas yang mempertimbangkan baik gejala maupun tes diagnostik dapat sangat membantu dalam diagnosis dini dan pencegahan kondisi terkait jantung, berpotensi menyelamatkan nyawa serta mengurangi beban pada sistem pelayanan kesehatan [2].

Dalam penelitian ini, masalah ketidakseimbangan kelas dalam *dataset* stroke diatasi dengan teknik *Synthetic Minority Over-sampling Technique (SMOTE)*. *SMOTE* merupakan teknik *oversampling* yang digunakan untuk menyeimbangkan distribusi data pada kelas minoritas (kasus stroke) melalui proses interpolasi linier antar titik data terdekat dalam ruang fitur, sehingga jumlah sampel kelas minoritas meningkat [10]. Dengan cara ini, distribusi kelas menjadi lebih seimbang dan model klasifikasi dapat mengenali pola kelas minoritas secara lebih efektif, meningkatkan kemampuan prediksi terhadap kasus stroke.

Selanjutnya, klasifikasi dilakukan menggunakan algoritma *K-Nearest Neighbors (KNN)*, yang bekerja dengan cara mengklasifikasikan data baru berdasarkan kesamaan fitur dengan sejumlah k-tetangga terdekat dalam ruang fitur. Metode *KNN* dipilih karena kemampuannya dalam menangkap pola lokal dalam data melalui perhitungan jarak *Euclidean*, yang efektif dalam mendeteksi pola penyakit. Penelitian sebelumnya menunjukkan bahwa algoritma ini mencapai akurasi hingga 90% dalam deteksi penyakit jantung, dengan nilai k yang optimal [1].

II. METODE

Alur dari penelitian ini meliputi beberapa hal seperti pengumpulan data, pengolahan data dan proses evaluasi pada model.



Gambar 1. Alur Penelitian

A. Pengumpulan Data

Dalam proses koleksi data, berbagai metode dapat diterapkan, seperti studi literatur dan ekstraksi informasi dari berbagai referensi. Penggunaan dataset terhadap penelitian ini diperoleh dari *dataset* yang diunduh melalui platform *Kaggle*. Dataset yang kami gunakan memiliki judul *"Stroke Risk Dataset"* dalam format csv [31]. *Dataset* tersebut terdiri dari delapan belas atribut, sebagai berikut *Chest Pain, Shortness of Breath, Irregular Heartbeat, Fatigue and Weakness, Dizziness, Swelling (Edema), Pain in Neck/Jaw/Shoulder/Back, Nausea/Vomiting, Chest Discomfort (Activity), Excessive Sweating, Persistent Cough, High Blood Pressure,*

Cold Hands/Feet, Snoring/Sleep Apnea, Anxiety/Feeling of Doom, Age, Stroke Risk (%), At Risk (Binary). Data tersebut akan diolah untuk mendukung tujuan penelitian.

TABEL 1
ATRIBUT

No	Atribut	Penggunaan
1	Chest pain	Riwayat nyeri dada
2	Shortness of breath	Riwayat sesak napas
3	Irregular heartbeat	Riwayat detak jantung tidak teratur
4	Fatigue and weakness	Riwayat kelelahan dan kelemahan
5	Dizziness	Riwayat pusing
6	Swelling (Edema)	Riwayat pembengkakan (Edema)
7	Pain in Neck / Jaw / Shoulder / Back	Riwayat nyeri pada leher, rahang, bahu, punggung
8	Excessive Sweating	Riwayat keringat berlebih
9	Persistent Cough	Riwayat batuk terus-menerus
10	Nausea/Vomiting	Riwayat mual / muntah
11	Chest Discomfort (Activity)	Riwayat rasa tidak nyaman pada dada saat beraktivitas
12	Snoring/Sleep Apnea	Riwayat mendengkur saat tidur
13	Anxiety/Feeling of Doom	Riwayat kecemasan
14	Cold Hands/Feet	Riwayat tangan / kaki dingin
15	High Blood Pressure	Riwayat tekanan darah tinggi
16	Age	Umur subjek
17	Stroke Risk (%)	Persentase resiko stroke pada subjek
18	At Risk (Binary)	Tingkat resiko (ya / tidak)

B. Pengolahan Data

Pada tahap pengolahan data, akan dilaksanakan kegiatan *Preprocessing*, dimana di dalam tahap *Preprocessing* tersebut terdapat beberapa tahap di dalamnya, diantaranya Pembersihan Data *(Data Cleaning)*, Transformasi Data, Pengelompokan Data *(Aggregation)*, Penyaringan Data *(Filtering)*, dan Integrasi Data.

C. Implementasi Teknik SMOTE

SMOTE merupakan salah satu langkah *oversampling* yang digunakan untuk menyeimbangkan distribusi data pada kelas minoritas [21]. Penting untuk dicatat bahwa dalam penelitian ini, teknik *SMOTE* diterapkan di dalam setiap iterasi dari proses *K-Fold Cross-Validation*. Secara spesifik, *SMOTE* hanya diaplikasikan pada bagian data yang berfungsi sebagai data latih (training set) pada fold tersebut. Langkah ini krusial untuk mencegah kebocoran data (*data leakage*) dan memastikan bahwa data uji (*validation set*) pada setiap *fold* tetap menjadi representasi data asli yang belum pernah dilihat oleh model. *SMOTE* memperluas dataset latih dengan menganalisis dan mensintesis sampel baru untuk kelas minoritas. Operasi dari teknik *SMOTE* meliputi langkah-langkah berikut ini:

1. Menentukan Tetangga Terdekat: Untuk setiap sampel pada kelas minoritas, dihitung jaraknya dengan sampel minoritas lainnya untuk menemukan himpunan tetangga terdekatnya. Perhitungan jarak ini menggunakan metrik jarak *Euclidean*, yang formulanya disajikan pada Persamaan (1).

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

Keterangan:
- $d(x,y)$: Jarak *Euclidean* antara titik data p dan q.
- n : Jumlah fitur (dimensi) pada dataset.
- $x_i \; dan \; y_i$: Nilai fitur ke-i dari titik data p dan q.

2. Menetapkan Tingkat Sampling: Menetapkan tingkat sampling berdasarkan rasio ketidakseimbangan dan menentukan rasio sampling.

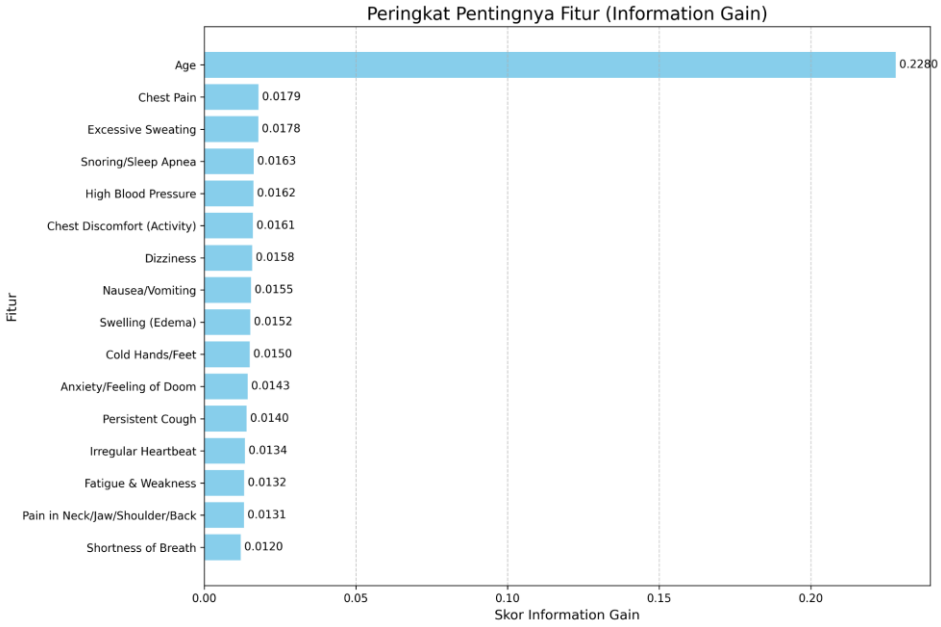
3. Membangkitkan Sampel Baru: Untuk setiap sampel kelas minoritas, dipilih sampel secara acak dari tetangga terdekatnya dan dibangkitkan sampel baru (x_baru) sesuai dengan Persamaan (2) [12].

$$x_{baru} = x + rand.(\underline{x} - x) \quad (2)$$

Keterangan:
- x : Sampel dari kelas minoritas.
- $\underline{x}$: Salah satu tetangga terdekat dari x yang dipilih acak.
- $rand$: Angka acak antara 0 dan 1.

D. Permodelan K-Nearest Neighbor

Tahap pemodelan diawali dengan analisis pentingnya fitur menggunakan metode *Information Gain*. Analisis ini bertujuan untuk memahami kontribusi informasi dari setiap fitur terhadap penentuan risiko stroke dan mengidentifikasi fitur dengan daya prediksi tertinggi. Hasil analisis, yang disajikan pada Gambar 2, menunjukkan bahwa fitur 'Usia' (*Age*) memiliki nilai *Information Gain* paling signifikan. Meskipun demikian, untuk membangun model yang komprehensif, seluruh 16 fitur tetap diikutsertakan dalam proses pemodelan.



Gambar 2. Peringkat Pentingnya Fitur (*Information Gain*)

Selanjutnya diterapkan algoritma *K-Nearest Neighbor (KNN)* untuk melakukan klasifikasi. Dalam implementasi penelitian ini, model *KNN* dikonfigurasi dengan beberapa parameter spesifik untuk memastikan transparansi dan reprodusibilitas. Nilai K atau jumlah tetangga terdekat ditetapkan sebesar 5, dan metrik jarak yang digunakan adalah Jarak *Euclidean*, yang sesuai dengan parameter default *'minkowski'* dengan p=2 pada pustaka *Scikit-learn.* Rumus untuk metrik ini dinyatakan pada Persamaan (1). Selain itu, metode pembobotan yang diterapkan adalah *'uniform'*, yang berarti setiap dari K tetangga terdekat memiliki pengaruh yang sama dalam menentukan kelas data baru.

Parameter-parameter ini dipilih sebagai konfigurasi dasar yang umum digunakan. Dengan konfigurasi tersebut, proses kerja algoritma dimulai dengan menghitung jarak antara data baru dan semua data dalam set pelatihan untuk menemukan 5 tetangga terdekat, yang kemudian diikuti dengan penentuan kelas data baru berdasarkan suara mayoritas dari kelima tetangga tersebut.

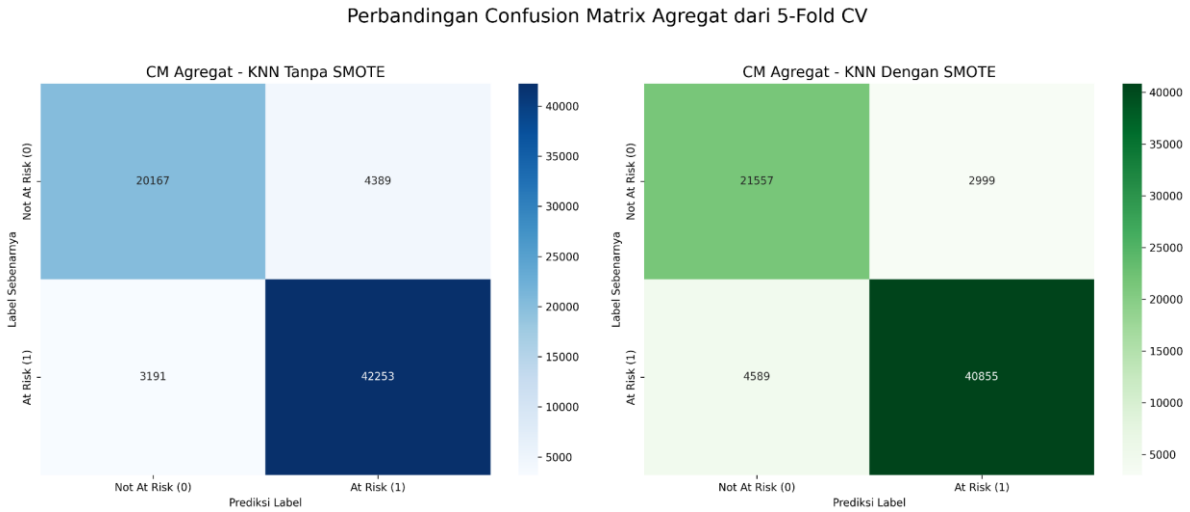
Jarak *Euclidean* adalah metrik jarak standar dan paling umum digunakan untuk data numerik kontinu. Validasi penggunaannya dalam model *KNN* ini bersifat implisit: keberhasilan model *KNN* dalam menghasilkan akurasi dan metrik evaluasi yang baik menunjukkan bahwa Jarak *Euclidean* merupakan pilihan metrik yang tepat untuk mengukur "kedekatan" atau "kemiripan" antar data dalam dataset ini. Dalam rumus ini, jarak dihitung sebagai akar kuadrat dari jumlah selisih nilai pada setiap atribut antara data uji dan data latih.

E. Evaluasi Model

Evaluasi kinerja model dilakukan dengan menggunakan serangkaian metrik standar yang dihitung dari hasil *Stratified K-Fold Cross-Validation*. Karena proses validasi dilakukan sebanyak 5 kali (5-*fold*), metrik performa akhir disajikan sebagai rata-rata (*mean*) dan standar deviasi (*standard deviation*) untuk memberikan gambaran yang komprehensif dan andal mengenai kinerja model.

Metrik utama yang digunakan untuk evaluasi adalah Akurasi, Presisi, *Recall*, dan *F1-Score*. Penggunaan rata-rata dari hasil *5-fold CV* memberikan estimasi performa yang stabil, sementara standar deviasi menunjukkan seberapa konsisten performa model pada subset data yang berbeda.

Selain metrik kuantitatif tersebut, *confusion matrix* agregat juga disajikan untuk memberikan gambaran visual mengenai distribusi kesalahan prediksi secara keseluruhan. Matriks pada Gambar 3 merupakan hasil penjumlahan nilai *True Positive*, *True Negative*, *False Positive*, dan *False Negative* dari kelima iterasi pengujian dalam proses cross-validation.



Gambar 3. Perbandingan *confusion matrix*

Dari matriks tersebut, dapat dianalisis secara lebih mendalam tipe kesalahan yang cenderung dibuat oleh model, apakah itu *False Positive* (memprediksi pasien berisiko padahal sehat) atau *False Negative* (memprediksi pasien sehat padahal berisiko).

F. Uji Signifikansi Statistik

Untuk memvalidasi apakah perbedaan kinerja antara model *KNN* dengan dan tanpa *SMOTE* bermakna secara statistik, digunakan uji *Wilcoxon signed-rank test*. Uji non-parametrik ini dipilih karena sesuai untuk membandingkan dua set skor berpasangan yang dihasilkan dari proses *5-fold cross-validation*. Tingkat signifikansi (*alpha*) ditetapkan sebesar 0.05. Sebuah p-value di bawah 0.05 akan dianggap menunjukkan adanya perbedaan yang signifikan secara statistik antara kedua model.

III. HASIL DAN PEMBAHASAN

Tujuan bagian ini adalah menganalisa efektivitas *SMOTE* dalam meningkatkan kinerja algoritma *KNN* dalam mendeteksi stroke. *SMOTE* bekerja dengan melakukan penyeimbangan terhadap ketidakseimbangan *dataset*. Pembahasan difokuskan pada perbandingan kinerja model sebelum dan setelah penerapan *SMOTE*.

A. Pengumpulan Data

Tahap ini menjelaskan atribut-atribut dalam *dataset* yang digunakan penelitian. Proses pengumpulan data ini diperlukan dalam memahami semua variabel beserta serta hubungan masing-masing variabel dengan kondisi yang ingin diprediksi, yaitu risiko yang dihadapi oleh individu, apakah termasuk dalam kategori *"At Risk"* atau *"Not At Risk"*. Pemahaman terhadap atribut yang tersedia akan membantu dalam menentukan variabel mana yang memiliki peranan paling signifikan dalam membangun model klasifikasi serta bagaimana atribut tersebut mempengaruhi hasil klasifikasi yang dilakukan menggunakan algoritma *KNN*, termasuk penyeimbangan data dengan metode *SMOTE*.

Berikut ini menyajikan sebagian dari *dataset* yang digunakan dalam penelitian ini. Setiap baris mewakili satu individu dengan atribut-atribut kesehatan yang relevan, seperti *"Chest Pain", "Shortness of Breath", "Age", dan "At Risk (Binary)"*. Data ini digunakan untuk membangun model klasifikasi yang memprediksi risiko stroke berdasarkan berbagai faktor medis dan kondisi kesehatan individu.

TABEL 2.
SAMPEL DATA

No	Chest Pain	Shortness of Breath	…	Age	At Risk (Binary)
0	0	1	…	54	1
1	0	0	…	49	0
2	1	0	…	62	1
3	1	0	…	48	1
4	0	0	…	61	1
5	1	1	…	34	0
6	0	1	…	74	1

…	…	…	…	…	…
69997	1	1	…	49	0
69998	0	1	…	45	0
69999	0	1	…	74	1

B. Pembersihan Data

Pada tahap ini, dilakukan pemeriksaan menyeluruh terhadap *dataset* untuk memastikan integritas data sebelum proses analisis lebih lanjut. Pemeriksaan meliputi identifikasi nilai yang hilang (*missing values*) serta deteksi data duplikat pada seluruh fitur dalam *dataset*. Hasil pemeriksaan menunjukkan bahwa *dataset* yang digunakan sudah lengkap, sehingga tidak ditemukan nilai yang hilang maupun data duplikat. Oleh karena itu, proses imputasi data tidak diperlukan dalam penelitian ini.

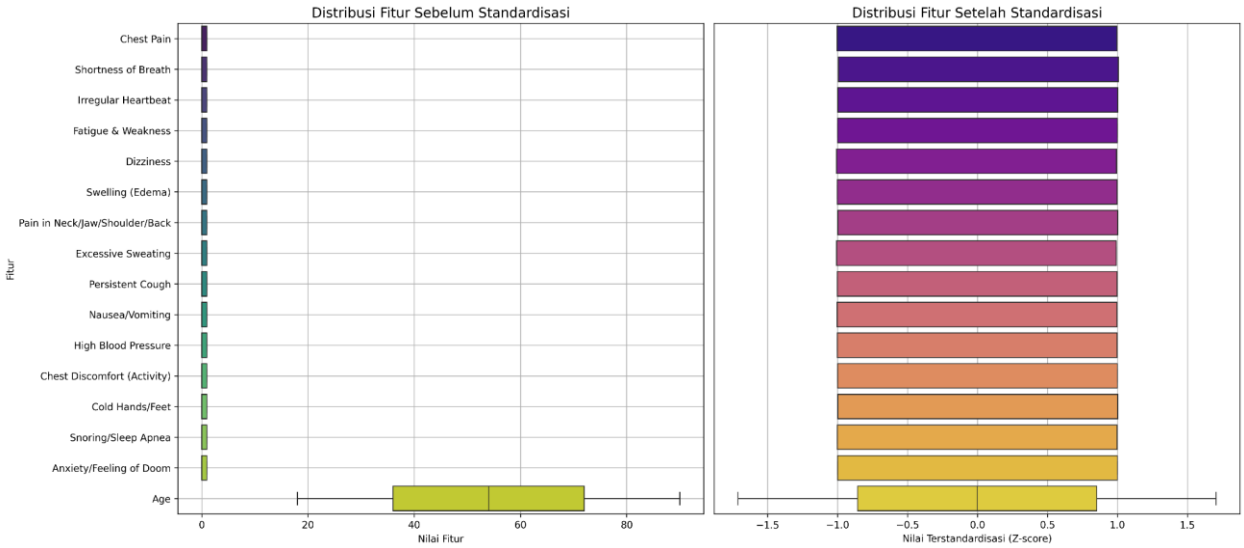
Selain itu, dilakukan juga pemisahan fitur dan target dengan menghapus kolom yang tidak digunakan sebagai prediktor, yaitu *Stroke Risk* (%) dan *At Risk* (*Binary*). Kolom *Stroke Risk* (%) merupakan variabel numerik yang menjadi indikator resiko namun tidak relevan untuk input klasifikasi biner, sedangkan *At Risk* (*Binary*) merupakan label target yang dipisahkan untuk keperluan pelatihan model.

Selanjutnya, untuk memastikan kualitas dan konsistensi analisis, dilakukan proses standardisasi pada seluruh fitur. Proses standardisasi bertujuan untuk mengubah skala setiap fitur sehingga memiliki distribusi dengan rata-rata (mean) 0 dan standar deviasi 1. Langkah ini sangat penting, terutama dalam penerapan algoritma berbasis jarak seperti *KNN*, karena perbedaan skala antar fitur dapat menyebabkan bias dalam perhitungan jarak antar data. Dengan demikian, proses standardisasi ini penting untuk meningkatkan performa model dalam mengklasifikasikan data secara lebih akurat dan adil.

Sebagai tambahan, dilakukan visualisasi distribusi data sebelum dan sesudah standardisasi untuk memastikan tidak ada outlier ekstrim yang dapat mempengaruhi hasil analisis. Seluruh proses pembersihan data ini bertujuan untuk meminimalkan potensi bias dan memastikan bahwa data yang digunakan telah siap untuk tahap pemodelan selanjutnya.

C. Transformasi Data

Dalam tahapan ini, transformasi data dilakukan untuk mempersiapkan dataset agar siap digunakan dalam model klasifikasi. Langkah utamanya adalah standardisasi fitur menggunakan *StandardScaler* dari pustaka *Scikit-learn*. Proses ini memastikan bahwa setiap fitur numerik disesuaikan skalanya, sehingga tidak ada satu fitur pun yang mendominasi fitur lainnya, khususnya pada algoritma *KNN* yang kinerjanya sangat sensitif terhadap variasi skala data.



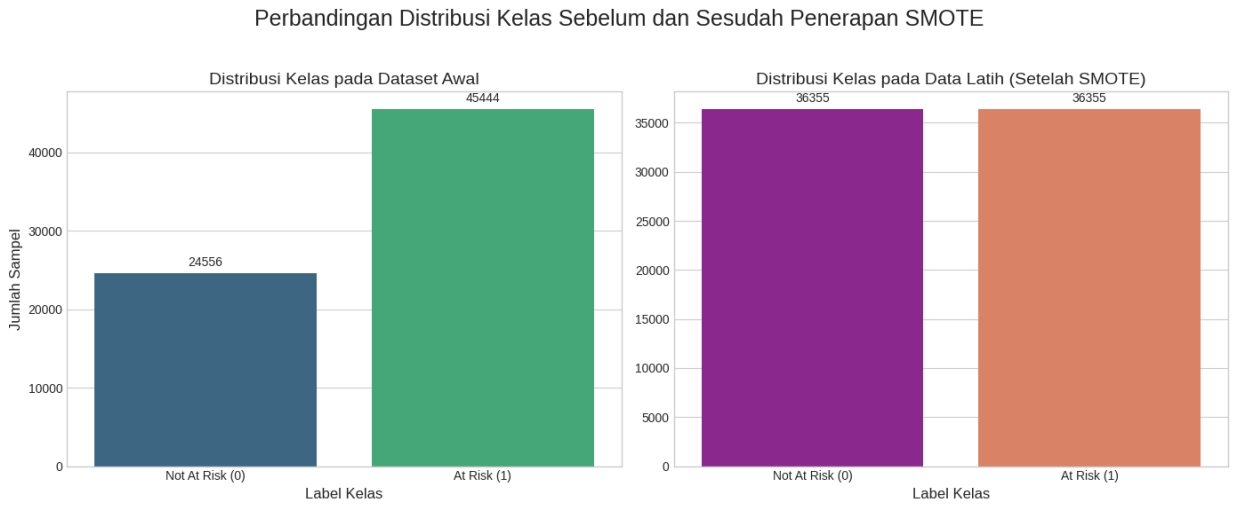
Gambar 4. Perbandingan distribusi fitur

D. Penyeimbangan Data dengan SMOTE

Analisis awal pada dataset menunjukkan adanya ketidakseimbangan kelas yang signifikan, di mana jumlah data untuk kelas mayoritas ('*At Risk*') jauh lebih banyak daripada kelas minoritas ('*Not At Risk*'). Kondisi ini dapat menyebabkan model menjadi bias, terutama pada algoritma yang sensitif terhadap distribusi data seperti *KNN*.

Untuk mengatasi masalah ini, diterapkan teknik *oversampling* yaitu *SMOTE* (*Synthetic Minority Over-sampling Technique*). Penting untuk dicatat bahwa *SMOTE* tidak diterapkan pada keseluruhan dataset di awal, melainkan diintegrasikan ke dalam alur kerja K-Fold Cross-Validation menggunakan Pipeline.

Pada setiap dari 5 iterasi (*fold*) validasi silang, proses *SMOTE* diaplikasikan secara eksklusif hanya pada data latih temporer untuk *fold* tersebut. Teknik ini menyeimbangkan kelas dengan membuat sampel sintetis baru untuk kelas minoritas. Sementara itu, data uji untuk *fold* tersebut dibiarkan dalam kondisi aslinya (tidak seimbang) untuk memastikan evaluasi model yang adil dan tidak bias. Pendekatan ini merupakan praktik terbaik untuk mencegah kebocoran data (*data leakage*) dan memastikan validitas hasil evaluasi.



Gambar 5. Grafik perbandingan data

Gambar 5 mengilustrasikan efek dari *SMOTE*. Grafik kiri menunjukkan distribusi kelas yang tidak seimbang pada keseluruhan *dataset* awal. Grafik kanan menunjukkan distribusi kelas yang telah seimbang pada sebuah set data latih yang representatif setelah *SMOTE* diterapkan, di mana jumlah sampel untuk kedua kelas menjadi setara.

E. Evaluasi Model KNN dengan SMOTE

Evaluasi model *KNN* dengan penerapan *SMOTE* dilakukan pada tahap ini setelah pelatihan menggunakan data seimbang. Pengukuran akurasi menjadi tahap pertama dalam proses evaluasi. Beberapa hal yang diuji antara lain adalah *precision*, *recall*, dan *F1-score*. Evaluasi ini bertujuan menilai performa model pada proses pengklasifikasian data yang telah diseimbangkan melalui metode *SMOTE*. serta untuk memastikan bahwa model tidak terpengaruh oleh ketidakseimbangan kelas dalam *dataset*.

1. Akurasi: Mengukur prosentasi prediksi data benar dibandingkan dengan semua data yang diambil.
2. *Precision* dan *Recall*: Mengukur kemampuan model dalam mengidentifikasi kelas minoritas (kelas yang lebih jarang), yang penting dalam masalah klasifikasi biner.
3. *F1-Score*: Menunjukkan grafik keseimbangan diantara *precision* dengan *recall*.

TABEL 3
HASIL EVALUASI MODEL *KNN* DENGAN *SMOTE*

Metode	*Mean ± SD*			
	Accuracy	*Precision*	*Recall*	*F1-Score*
SMOTE + KNN	0.8916 ± 0.0020	0.9316 ± 0.0027	0.8990 ± 0.0025	0.9150 ± 0.0016

F. Evaluasi Model *KNN tanpa SMOTE*

Selanjutnya, dilakukan evaluasi terhadap model *KNN* tanpa *SMOTE* yang menggunakan *dataset* asli tanpa penyeimbangan kelas. Evaluasi serupa dilakukan untuk mengukur kinerja model pada *dataset* yang tidak diseimbangkan. Dengan membandingkan hasil ini dengan model *KNN* yang menggunakan *SMOTE*, kita dapat mengukur seberapa besar pengaruh *SMOTE* terhadap kinerja model.

TABEL 4
HASIL EVALUASI MODEL *KNN* TANPA *SMOTE*

Metode	*Mean ± SD*			
	Accuracy	*Precision*	*Recall*	*F1-Score*
KNN	0.8917 ± 0.0024	0.9059 ± 0.0037	0.9298 ± 0.0009	0.9177 ± 0.0017

G. Perbandingan Kinerja KNN dengan dan Tanpa SMOTE

Pada tahap ini, perbandingan dilakukan antara model *KNN* dengan *SMOTE* dan *KNN* tanpa *SMOTE* berdasarkan metrik-metrik yang telah disebutkan sebelumnya. Pembandingan ini memberikan gambaran tentang seberapa besar pengaruh teknik penyeimbangan kelas menggunakan *SMOTE* terhadap kinerja model *KNN*.

TABEL 5
HASIL PERBANDINGAN EVALUASI *MODEL*

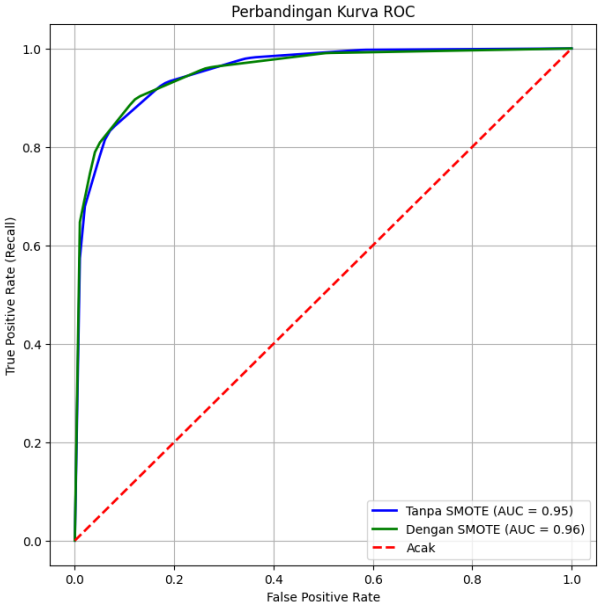
Metode	*Mean ± SD*			
	Accuracy	*Precision*	*Recall*	*F1-Score*
SMOTE + KNN	0.8916 ± 0.0020	0.9316 ± 0.0027	0.8990 ± 0.0025	0.9150 ± 0.0016
KNN	0.8917 ± 0.0024	0.9059 ± 0.0037	0.9298 ± 0.0009	0.9177 ± 0.0017

H. Analisis Hasil

Analisis perbandingan kinerja antara model *KNN* dengan dan tanpa *SMOTE* dievaluasi menggunakan *5-fold cross-validation*, dengan hasil disajikan pada Tabel 5. Poin utama dari analisis ini adalah, meskipun terdapat sedikit perbedaan pada beberapa metrik, uji signifikansi statistik (*Wilcoxon test*) menunjukkan tidak ada perbedaan kinerja yang signifikan secara statistik antara kedua model (semua p-value > 0.05).

Selain metrik evaluasi pada Tabel 5, analisis visual menggunakan kurva *Receiver Operating Characteristic* (*ROC*) dan *Precision-Recall (PR)* juga dilakukan untuk mendapatkan pemahaman yang lebih mendalam mengenai kinerja model di berbagai ambang batas klasifikasi.

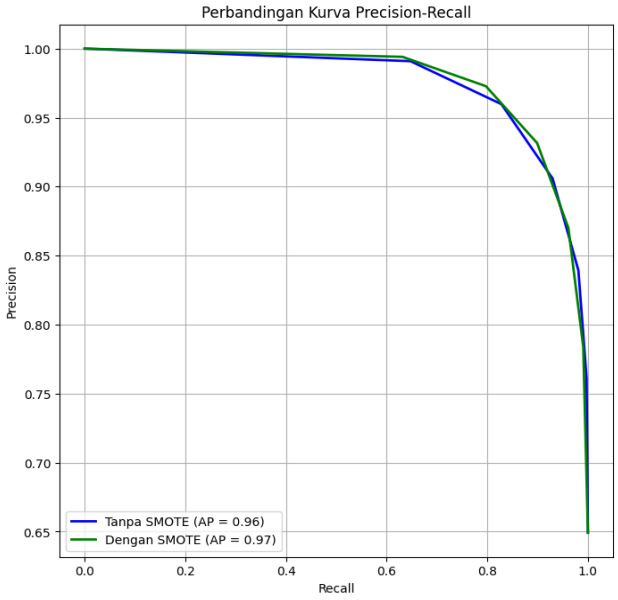
Kurva *ROC* digunakan untuk mengevaluasi kemampuan diskriminatif sebuah model, yaitu sejauh mana model dapat membedakan antara kelas positif (*'At Risk'*) dan kelas negatif (*'Not At Risk').* Hasil perbandingan kurva *ROC* antara model dengan dan tanpa *SMOTE* disajikan pada Gambar 6.



Gambar 6.Perbandingan Kurva *ROC*

Dari gambar tersebut, terlihat bahwa kedua model menunjukkan performa yang sangat baik. Model tanpa *SMOTE* menghasilkan nilai *Area Under the Curve* (*AUC*) sebesar 0.95, sedangkan model dengan *SMOTE* mencapai *AUC* yang sedikit lebih tinggi, yaitu 0.96. Nilai *AUC* yang mendekati 1.0 mengindikasikan bahwa kedua model memiliki kemampuan yang sangat kuat dalam memisahkan kelas. Peningkatan tipis pada model dengan *SMOTE* menunjukkan bahwa penyeimbangan data memberikan sedikit keunggulan dalam hal kemampuan diskriminasi secara keseluruhan.

Kurva *Precision-Recall (PR*) sangat informatif untuk *dataset* yang tidak seimbang karena berfokus pada kinerja model terhadap kelas minoritas. Kurva ini memvisualisasikan *trade-off* antara presisi dan recall.



Gambar 7. Perbandingan Kurva Precision-Recall

Berdasarkan Gambar 7, kedua model kembali menunjukkan kinerja yang sangat tinggi. Model tanpa *SMOTE* mencatatkan nilai *Average Precision (AP)* sebesar 0.96, sementara model dengan *SMOTE* sedikit unggul dengan *AP* sebesar 0.97. Nilai *AP* yang lebih tinggi pada model dengan *SMOTE* mengindikasikan bahwa model tersebut mampu mempertahankan presisi yang tinggi bahkan saat berusaha mengidentifikasi sebagian besar sampel kelas positif (*recall* tinggi). Hal ini memperkuat gagasan bahwa *SMOTE* membantu model menjadi lebih andal dalam prediksi kelas positif, yang sangat penting dalam konteks medis. Analisis visual ini melengkapi evaluasi numerik dan menunjukkan bahwa

1. Akurasi: Kinerja kedua model pada metrik akurasi hampir identik. Model tanpa *SMOTE* mencatatkan rata-rata akurasi 0.8917 ± 0.0024, sementara model dengan *SMOTE* sedikit lebih rendah di 0.8916 ± 0.0020. Klaim peningkatan akurasi oleh SMOTE tidak terbukti dalam hasil ini.
2. Presisi: Model dengan *SMOTE* menunjukkan rata-rata presisi yang sedikit lebih tinggi (0.9316) dibandingkan model tanpa *SMOTE* (0.9059). Ini mengindikasikan bahwa model dengan *SMOTE* cenderung lebih akurat ketika memberikan prediksi positif, meskipun perbedaan ini tidak signifikan secara statistik (p=0.0625).
3. *Recall*: Sebaliknya, model tanpa *SMOTE* unggul secara jelas dalam hal recall dengan skor 0.9298, melampaui model dengan *SMOTE* yang hanya mencapai 0.8990. Ini menunjukkan bahwa model asli lebih sensitif dan mampu mengidentifikasi lebih banyak kasus positif yang sebenarnya.
4. *F1-Score*: Dalam hal *F1-score*, model tanpa *SMOTE* mencatatkan skor yang sedikit lebih unggul (0.9177) dibandingkan model dengan *SMOTE* (0.9150), menandakan keseimbangan harmonis antara presisi dan *recall* yang sedikit lebih baik.

Secara keseluruhan, meskipun *SMOTE* berhasil sedikit meningkatkan presisi, hal ini tidak cukup untuk memberikan keunggulan yang signifikan secara statistik dan justru mengorbankan sedikit kemampuan recall. Dalam konteks deteksi medis di mana sensitivitas (*recall*) untuk menemukan semua kasus positif sangat krusial, model tanpa *SMOTE* menunjukkan profil kinerja yang sedikit lebih unggul dan lebih disarankan.

IV. SIMPULAN

Berdasarkan hasil analisis dan uji signifikansi, disimpulkan bahwa penerapan *SMOTE* tidak efektif dalam meningkatkan performa model *KNN* secara signifikan pada dataset ini. Meskipun *SMOTE* sedikit meningkatkan presisi, hal ini diiringi dengan penurunan recall dan tidak terbukti signifikan secara statistik ($p > 0.05$).

Temuan ini menyoroti bahwa *SMOTE*, meskipun merupakan teknik yang populer, tidak selalu menjamin peningkatan kinerja dan efektivitasnya sangat bergantung pada karakteristik dataset. Mengingat tidak adanya bukti peningkatan yang signifikan dan adanya penurunan pada metrik recall yang krusial untuk deteksi medis, model *KNN* tanpa *SMOTE* dapat dianggap sebagai model yang lebih efisien dan memadai untuk kasus ini. Penelitian ini menggarisbawahi pentingnya validasi statistik dalam menentukan dampak sebenarnya dari teknik penyeimbangan data.

DAFTAR PUSTAKA

[1] F. N. Syahreza, Puspita Nurul Sabrina, and Edvin Ramadhan, "PREDIKSI PENYAKIT STROKE MENGGUNAKAN METODE K-NEAREST NEIGHBOUR DAN INFORMATION GAIN," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 8, no. 6, pp. 11354–11359, Nov. 2024, doi: https://doi.org/10.36040/jati.v8i6.11427.

[2] P. N. Srinivasu, U. Sirisha, K. Sandeep, S. P. Praveen, L. P. Maguluri, and T. Bikku, "An Interpretable Approach with Explainable AI for Heart Stroke Prediction," Diagnostics, vol. 14, no. 2, p. 128, Jan. 2024, doi: https://doi.org/10.3390/diagnostics14020128.

[3] D. U. maula Rachmad, H. Oktavianto, and M. Rahman, "Perbandingan Metode K-Nearest Neighbors dan Gaussian Naive Bayes untuk Klasifikasi Penyakit Stroke," Jurna Smart Teknologi, vol. 3, no. 4, pp. 405–412, May 2022, Accessed: Apr. 30, 2025. [Online]. Available: https://jurnal.unmuhjember.ac.id/index.php/JST/article/view/7601

[4] A. Byna and M. Basit, "PENERAPAN METODE ADABOOST UNTUK MENGOPTIMASI PREDIKSI PENYAKIT STROKE DENGAN ALGORITMA NAÏVE BAYES," Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 9, no. 3, Aug. 2020, doi: https://doi.org/10.32736/sisfokom.v9i3.1023.

[5] M. N. Maskuri, K. Sukerti, and R. M. H. Bhakti, "Penerapan Algoritma K-Nearest Neighbor untuk Memprediksi Penyakit Stroke," Jurnal Ilmiah Intech : Information Technology Journal of UMUS, vol. 4, no. 1, pp. 130–140, May 2022, Accessed: Apr. 26, 2025. [Online]. Available: https://jurnal.umus.ac.id/index.php/intech/article/view/751

[6] P. W. S. Aji, S. Supriyanto, and R. Dijaya, "Prediksi Penyakit Stroke Menggunakan Metode Random Forest," KESATRIA Jurnal Penerapan Sistem Informasi (Komputer & Manajemen), vol. 4, no. 4, pp. 916–924, Oct. 2023, Accessed: Apr. 20, 2025. [Online]. Available: https://tunasbangsa.ac.id/pkm/index.php/kesatria/article/view/242

[7] Novianti Puspitasari, Anindita Septiarini, and Abdul Razak Aliudin, "METODE K-NEAREST NEIGHBOR DAN FITUR WARNA UNTUK KLASIFIKASI DAUN SIRIH BERDASARKAN CITRA DIGITAL," Prosisko/Prosisko: jurnal pengembangan riset dan observasi sistem komputer, vol. 10, no. 2, pp. 165–172, Aug. 2023, doi: https://doi.org/10.30656/prosisko.v10i2.6924.

[8] Z. Umar, None Dityo Kreshna Argeshwara, None Aji Prasetya Wibawa, A. Nur, and S. Hadi, "Pemodelan Sistem Deteksi Kadar Unsur Hara Tanah Berdasarkan Nilai NPK Menggunakan Metode Fuzzy Mamdani," Jurnal sains dan informatika, pp. 77–88, Aug. 2023, doi: https://doi.org/10.34128/jsi.v9i1.523.

[9] V. Saini, L. Guada, and D. R. Yavagal, "Global Epidemiology of Stroke and Access to Acute Ischemic Stroke Interventions," Neurology, vol. 97, no. 20 Supplement 2, pp. S6–S16, Nov. 2021, doi: https://doi.org/10.1212/WNL.0000000000012781.

[10] Tanapol Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," vol. 16, no. 1, Apr. 2023, doi: https://doi.org/10.1186/s13040-023-00330-4.

[11] Amir Reza Salehi and Majid Khedmati, "A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data," Scientific reports (Nature Publishing Group), vol. 14, no. 1, Mar. 2024, doi: https://doi.org/10.1038/s41598-024-55598-1.

[12] Z. Zhou, C. Xu, Y. Qiao, J. Xiong, and J. Yu, "Enhancing Equipment Health Prediction with Enchanced SMOTE-KNN," JIEAS Journal of Industrial Engineering and Applied Science, vol. 2, no. 2, Apr. 2024, Accessed: Mar. 25, 2025. [Online]. Available: https://www.suaspress.org/ojs/index.php/JIEAS/article/view/v2n2a03

[13] M. Khushi et al., "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," IEEE Access, vol. 9, pp. 109960–109975, 2021, doi: https://doi.org/10.1109/access.2021.3102399.

[14] Oladunjoye John Abiodun and A. I. Wreford, "Stroke Prediction Using Smote for Data Balancing, XGBoost and KNN Ensemble Algorithms," Deleted Journal, pp. 42–53, Aug. 2023, doi: https://doi.org/10.56557/japsi/2023/v15i18349.

[15] H. M. Merdas, "Elastic Net – MLP – SMOTE (EMS)-Based Model for Enhancing Stroke Prediction," Medinformatics, pp. 73–78, Apr. 2024, doi: https://doi.org/10.47852/bonviewmedin42022470.

[16] F. Yagin, I. Cicek, and Z. Kucukakcali, "Classification of stroke with gradient boosting tree using smote-based oversampling method," Medicine Science | International Medical Journal, vol. 10, no. 4, p. 1510, 2021, doi: https://doi.org/10.5455/medscience.2021.09.322.

[17] M. A. Aish, F. Nasim, K. I. Ali, S. Akhter, and S. Azeem, "Improving Stroke Prediction Accuracy through Machine Learning and Synthetic Minority Over-sampling," Journal of Computing & Biomedical Informatics, vol. 7, no. 02, Sep. 2024, Accessed: Apr. 26, 2025. [Online]. Available: https://jcbi.org/index.php/Main/article/view/566/469

[18] K. Swain et al., "Enhancing Stroke Prediction Using LightGBM With SMOTE-ENN and Fine-Tuning: A Comprehensive Analysis," Cureus Journals, Dec. 2024, doi: https://doi.org/10.7759/s44389-024-02268-y.

[19] Fitri Handayani and Reny Medikawati Taufiq, "Komparasi Algoritma Menggunakan Teknik Smote Dalam Melakukan Klasifikasi Penyakit Stroke Otak," Jurnal CoSciTech (Computer Science and Information Technology), vol. 5, no. 2, pp. 367–372, Aug. 2024, doi: https://doi.org/10.37859/coscitech.v5i2.7439.

[20] F. Fadmadika, H. H. Handayani, T. A. Mudzakir, and J. Indra, "PENGARUH SMOTE TERHADAP PERFORMA ALGORITMA RANDOM FOREST DAN ALGORITMA GRADIENT BOOSTING DALAM MEMPREDIKSI PENYAKIT STROKE," Jurnal Teknik Informasi dan Komputer (Tekinkom), vol. 7, no. 2, p. 837, Dec. 2024, doi: https://doi.org/10.37600/tekinkom.v7i2.1575.

[21] L. Pasiolo, I. Afrianty, E. Budianita, and R. Abdillah, "PENERAPAN TEKNIK SMOTE PADA KLASIFIKASI PENYAKIT STROKE DENGAN ALGORITMA SUPPORT VECTOR MACHINE," ZONAsi: Jurnal Sistem Informasi, vol. 7, no. 1, Jan. 2025, Accessed: Apr. 23, 2025. [Online]. Available: https://journal.unilak.ac.id/index.php/zn/article/view/24731

[22] Desti Mualfah, Wahyu Fadila, and Rahmad Firdaus, "Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest," Jurnal CoSciTech (Computer Science and Information Technology), vol. 3, no. 2, pp. 107–113, Aug. 2022, doi: https://doi.org/10.37859/coscitech.v3i2.3912.

[23] Oladunjoye John Abiodun and A. I. Wreford, "Stroke Prediction Using Smote for Data Balancing, XGBoost and KNN Ensemble Algorithms," Deleted Journal, pp. 42–53, Aug. 2023, doi: https://doi.org/10.56557/japsi/2023/v15i18349.

[24] K. Akmal, A. Faqih, and Fatihanursari Dikananda, "PERBANDINGAN METODE ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBORS UNTUK KLASIFIKASI PENYAKIT STROKE," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 7, no. 1, pp. 470–477, Mar. 2023, doi: https://doi.org/10.36040/jati.v7i1.6367.

[25] M. Hafidz Ariansyah, Sri Winarno, Esmi Nur Fitri, and Retha, "Multi-Layer Perceptron For Diagnosing Stroke With The SMOTE Method In Overcoming Data Imbalances," Innovation in Research of Informatics (INNOVATICS), vol. 5, no. 1, Mar. 2023, doi: https://doi.org/10.37058/innovatics.v5i1.6565.

[26] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," Jurnal Teknologi dan Sistem Komputer, vol. 8, no. 2, pp. 89–93, Feb. 2020, doi: https://doi.org/10.14710/jtsiskom.8.2.2020.89-93.

[27] N. Y. Paramitha, A. Nuryaman, A. Faisol, E. Setiawan, and D. E. Nurvazly, "Klasifikasi Penyakit Stroke Menggunakan Metode Naive Bayes," Jurnal Siger Matematika, vol. 04, no. 01, Mar. 2023, Accessed: Apr. 19, 2025. [Online]. Available: https://jsm.fmipa.unila.ac.id/index.php/jsm/article/view/33

[28] H. Siregar, A. Tumanggor, and None Akhwan Rahmadani, "Penerapan K-Nearest Neighbors (KNN) dalam Memprediksi dan Menghitung Akurasi Data Penyakit Stroke," Jurnal Penelitian Rumpun Ilmu Teknik, vol. 2, no. 4, pp. 146–154, Nov. 2023, doi: https://doi.org/10.55606/juprit.v2i4.3040.

[29] "Stroke pada Lansia di Indonesia: Gambaran Faktor Risiko Berdasarkan Gender (SKI 2023)," Jurnal Biostatistik, Kependudukan, dan Informatika Kesehatan, vol. 5, no. 1, Dec. 2024, doi: https://doi.org/10.7454/bikfokes.v5i1.1092.

[30] Siti Retno Wulandari, "Pembiayaan Penyakit Stroke Masih Tinggi Hingga Rp5,2 Triliun," Mediaindonesia.com, Oct. 25, 2024. https://mediaindonesia.com/humaniora/712147/pembiayaan-penyakit-stroke-masih-tinggi-hingga-rp52-triliun (accessed Apr. 30, 2025).J. Padhye, V. Firoiu, &D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 199

[31] M. A. Tusher, "Stroke Risk Prediction Dataset Based on Symptoms," Kaggle, 2025. [Online]. Tersedia: https://doi.org/10.34740/KAGGLE/DSV/10754870.

[32] C. Supriyanto, A. Salam, J. Zeniarja, D. W. Utomo, I. N. Dewi, C. Paramita, A. Wijaya, and N. Z. M. Safar, "A Bibliometric Review of Deep Learning Approaches in Skin Cancer Research," Computation, vol. 13, no. 3, p. 78, Mar. 2025.

[33] A. A. Dzaky, J. Zeniarja, C. Supriyanto, G. F. Shidik, C. Paramita, E. R. Subhiyakto, and S. Rakasiwi, "Optimization Chatbot Services Based on DNN-Bert for Mental Health of University Students," Journal of Applied Informatics and Computing (JAIC), vol. 8, no. 1, pp. 13–21, Jul. 2024.

[34] C. Paramita, F. A. Rafrastara, and L. I. Kencana, "Pengembangan Sistem Klasifikasi Karakteristik Siswa Berbasis Website dengan menggunakan Algoritma C4.5," Jurnal Informatika: Jurnal pengembangan IT (JPIT), vol. 8, no. 1, pp. 17-21, Jan. 2023.

[35] E. R. Subhiyakto, S. Rakasiwi, J. Zeniarja, C. Paramita, G. F. Shidik, Z. A. Hasibuan, and M. G. Kesić, "Evaluation of Resampling Techniques in CNN-Based Heartbeat Classification," Ingénierie des Systèmes d'Information, vol. 29, no. 4, pp. 1323–1332, Aug. 2024