

Prediksi Risiko Stroke berbasis *Algoritma SVM* dan *XGBoost* dengan *SMOTE-ENN*

Stroke Risk Prediction based on SVM and XGBoost Algorithms with SMOTE-ENN

¹Rianindya Chandra Hardika*, ²Dedy Mirwansyah, ³Abadi Nugroho

¹Program Studi Sistem Informasi, Sekolah Tinggi Teknologi Bontang, Indonesia

^{2,3}Program Studi Teknik Informatika, Sekolah Tinggi Teknologi Bontang, Indonesia

^{1,2,3}Jl. Letjen S Parman, Bontang Barat, Bontang, Kalimantan Timur, Indonesia

*e-mail: hardikachandra@stitek.ac.id

(received: 25 August 2026, revised: 11 September 2026, accepted: 12 September 2026)

Abstrak

Stroke merupakan salah satu gangguan neurologis yang memerlukan identifikasi risiko secara dini. Salah satu permasalahan dalam prediksi stroke menggunakan *machine learning* adalah ketidakseimbangan kelas yang dapat menyebabkan model lebih dominan mengenali kelas mayoritas. Penelitian ini bertujuan membandingkan kinerja *Support Vector Machine* (SVM) dan *Extreme Gradient Boosting* (XGBoost) dalam prediksi stroke dengan menerapkan *Synthetic Minority Over-sampling Technique–Edited Nearest Neighbours* (SMOTE-ENN). Dataset yang digunakan merupakan *Stroke Prediction Dataset* yang terdiri atas 43.400 data dan 12 atribut. Tahapan penelitian terdiri dari *preprocessing*, pembagian data dengan rasio 80:20, penerapan SMOTE-ENN pada data latih, pelatihan SVM dan XGBoost, kemudian evaluasi menggunakan *accuracy*, *precision*, *recall*, *F1-score*, *confusion matrix*, ROC-AUC, dan waktu komputasi. Hasil penelitian menunjukkan bahwa SMOTE-ENN berhasil meningkatkan representasi kelas stroke dari 1,8% menjadi distribusi seimbang pada data latih. Model SMOTE-ENN dan XGBoost memperoleh *accuracy* 96,96%, *precision* 96,83%, *recall* 97,11%, dan *F1-score* 96,97% dengan waktu pelatihan 4,68 detik. Model SMOTE-ENN-SVM memperoleh *accuracy* 90,13%, *precision* 88,07%, *recall* 92,83%, dan *F1-score* 90,39% dengan waktu pelatihan 21,2 menit. Berdasarkan hasil tersebut SMOTE-ENN-XGBoost memberikan kinerja klasifikasi dan efisiensi komputasi yang lebih baik dibandingkan SMOTE-ENN-SVM dalam prediksi stroke. Hasil ini menunjukkan bahwa penerapan SMOTE-ENN pada XGBoost mampu mendukung deteksi risiko stroke dengan kinerja klasifikasi tinggi dan waktu komputasi yang lebih efisien.

Kata kunci : machine learning, SMOTE-ENN, stroke, SVM, XGBoost

Abstract

Stroke is a neurological disorder that requires early risk identification. One of the challenges in stroke prediction using machine learning is class imbalance, which can cause models to favor the majority class. This study aims to compare the performance of *Support Vector Machine* (SVM) and *Extreme Gradient Boosting* (XGBoost) for stroke prediction using the *Synthetic Minority Over-sampling Technique–Edited Nearest Neighbours* (SMOTE-ENN). The dataset used was the *Stroke Prediction Dataset*, consisting of 43,400 records and 12 attributes. The research procedure included data preprocessing, an 80:20 train-test split, application of SMOTE-ENN to the training data, SVM and XGBoost model training, and evaluation using *accuracy*, *precision*, *recall*, *F1-score*, *confusion matrix*, ROC-AUC, and computational time. The results show that SMOTE-ENN successfully increased the representation of the stroke class from 1.8% to a balanced distribution in the training data. The SMOTE-ENN-XGBoost model achieved an *accuracy* of 96.96%, *precision* of 96.83%, *recall* of 97.11%, and *F1-score* of 96.97%, with a training time of 4.68 seconds. In comparison, the SMOTE-ENN-SVM model achieved an *accuracy* of 90.13%, *precision* of 88.07%, *recall* of 92.83%, and *F1-score* of 90.39%, with a training time of 21.2 minutes. These results indicate that SMOTE-ENN-XGBoost provides better classification performance and computational efficiency than SMOTE-ENN-SVM for stroke prediction. The findings suggest that applying SMOTE-ENN with XGBoost can

support stroke risk detection with high classification performance and more efficient computational time.

Keywords: machine learning, SMOTE-ENN, stroke, SVM, XGBoost.

1 Pendahuluan

Stroke merupakan salah satu gangguan neurologis akut yang memiliki dampak signifikan terhadap kondisi kesehatan. Gangguan ini terjadi akibat terhambatnya aliran darah menuju otak, baik karena penyumbatan maupun perdarahan pada pembuluh darah, yang dapat mengganggu suplai oksigen dan nutrisi ke jaringan otak [1]. Kondisi tersebut dapat menyebabkan kerusakan jaringan dan gangguan fungsi neurologis dalam waktu relatif singkat. Organisasi Kesehatan Dunia (WHO) menempatkan stroke sebagai salah satu penyebab utama kematian di dunia, dengan kontribusi sekitar 11% terhadap tingkat kematian global [1]. Tingginya risiko kematian dan potensi disabilitas yang ditimbulkan oleh stroke menegaskan pentingnya identifikasi risiko secara dini sebagai langkah strategis dalam mendukung pencegahan dan penanganan yang lebih baik.

Perkembangan teknologi digital dan kesadaran masyarakat terhadap kesehatan membuka peluang pemanfaatan data untuk mengidentifikasi pola yang berkaitan dengan kejadian penyakit stroke. Kemampuan *machine learning* dalam mempelajari hubungan antarvariabel memungkinkan data klinis diolah menjadi informasi prediktif yang dapat mendukung proses identifikasi risiko [2]. Berbagai karakteristik klinis, seperti kadar glukosa, *Body Mass Index* (BMI), dan tekanan darah, dapat digunakan sebagai atribut dalam proses klasifikasi. Pendekatan berbasis *machine learning* dengan demikian memiliki potensi untuk melengkapi proses penilaian risiko dengan memanfaatkan pola yang terdapat pada data historis. Namun, keberhasilan model prediksi tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga oleh karakteristik dan distribusi data yang menjadi dasar proses pembelajaran.

Analisis prediksi merupakan pendekatan berbasis data yang digunakan untuk memperkirakan kemungkinan terjadinya suatu kejadian berdasarkan pola dan karakteristik data historis. Pendekatan ini telah diterapkan pada berbagai sektor, seperti keuangan [3], industri [4], perdagangan, transportasi, pertanian [5], dan kesehatan [1][4][5]. Perkembangan *machine learning* semakin memperluas penerapan analisis prediktif karena mampu mengidentifikasi pola kompleks dari berbagai variabel. Dalam bidang kesehatan, pendekatan ini telah dimanfaatkan untuk memperkirakan risiko penyakit berdasarkan karakteristik dan faktor klinis pasien, salah satunya dalam prediksi risiko stroke. Permasalahan metodologis dalam prediksi stroke adalah ketidakseimbangan kelas (*class imbalance*). Permasalahan tersebut terjadi saat jumlah observasi pada kelas stroke jauh lebih sedikit dibandingkan dengan kelas tidak stroke. Distribusi kelas yang timpang dapat menyebabkan model lebih dominan mempelajari karakteristik kelas mayoritas. Akibatnya kemampuan dalam mengenali kelas stroke sebagai kelas minoritas menjadi kurang optimal. Kondisi tersebut juga dapat menghasilkan *accuracy* yang relatif tinggi tanpa diikuti kemampuan deteksi kelas stroke yang memadai [6].

Pendekatan yang dapat digunakan untuk menangani permasalahan tersebut adalah *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE meningkatkan representasi kelas minoritas dengan membentuk sampel sintetis berdasarkan karakteristik observasi minoritas yang berdekatan [8]. Berbeda dengan pendekatan yang hanya menggandakan observasi yang telah tersedia, mekanisme tersebut menghasilkan variasi observasi baru yang dapat memberikan informasi tambahan bagi proses pembelajaran model. Namun, penambahan sampel sintetis juga dapat menghasilkan observasi yang kurang konsisten atau berpotensi menjadi *noise*. Perlu menggunakan SMOTE-ENN sebagai pendekatan hibrida yang mengombinasikan SMOTE dengan *Edited Nearest Neighbours* (ENN). SMOTE berperan dalam meningkatkan representasi kelas minoritas, sedangkan ENN digunakan untuk menyaring observasi yang tidak konsisten berdasarkan hubungan terdekat.

Pemanfaatan *machine learning* dalam bidang kesehatan telah berkembang luas, ditandai dengan penerapan berbagai algoritma untuk klasifikasi dan prediksi penyakit berdasarkan karakteristik data klinis yang beragam. Berbagai pendekatan menunjukkan bahwa pemilihan algoritma menjadi salah satu faktor penting dalam memperoleh kinerja prediksi yang sesuai dengan karakteristik *dataset*. SVM telah diterapkan pada sejumlah penelitian prediksi penyakit diabetes [6] [7], tumor otak [9], dan penyakit jantung [1][10][11][12], sedangkan KNN, Random Forest, dan XGBoost juga digunakan pada berbagai permasalahan klasifikasi medis [2][9][10][13]. Berdasarkan penelitian yang telah

<http://sistemasi.ftik.unisi.ac.id>

dilakukan SVM menunjukkan kemampuan yang baik pada data dengan karakteristik berdimensi tinggi sedangkan XGBoost menunjukkan performa kompetitif pada pengolahan dataset berukuran besar. Namun performa algoritma dapat dipengaruhi oleh karakteristik *dataset*, distribusi kelas, atribut yang digunakan, serta strategi *preprocessing* yang diterapkan.

Berdasarkan kondisi tersebut, terdapat ruang penelitian untuk menguji secara komparatif kinerja SVM dan XGBoost setelah penerapan teknik penyeimbangan kelas yang sama, yaitu SMOTE-ENN, pada dataset stroke dengan tingkat ketidakseimbangan kelas yang ekstrem. Penggunaan dua algoritma dengan mekanisme pembelajaran yang berbeda memungkinkan analisis terhadap efektivitas SMOTE-ENN pada model berbasis *kernel* dan model berbasis *ensemble boosting*. Efektivitas teknik penyeimbangan kelas tidak hanya dipengaruhi oleh distribusi data, tetapi juga dapat berbeda sesuai dengan karakteristik dan mekanisme pembelajaran algoritma yang digunakan. Oleh karena itu, penelitian ini tidak hanya membandingkan kinerja klasifikasi SVM dan XGBoost, tetapi juga mengevaluasi kemampuan keduanya dalam mengenali kelas stroke serta efisiensi waktu pelatihan.

2 Tinjauan Literatur

Pemanfaatan *machine learning* dalam bidang kesehatan telah berkembang pesat melalui penerapan berbagai algoritma untuk menganalisis data medis dan membangun model klasifikasi. Setiap algoritma memiliki karakteristik dan mekanisme pembelajaran yang berbeda dalam mengenali pola data, termasuk *Support Vector Machine* (SVM) dan *Extreme Gradient Boosting* (XGBoost) yang memiliki pendekatan berbeda dalam melakukan klasifikasi. Berdasarkan penelitian terdahulu terhadap 17 sumber literatur, *Support Vector Machine* (SVM) merupakan salah satu algoritma yang paling banyak digunakan, dengan penerapan 11 penelitian. Sejumlah penelitian menunjukkan bahwa SVM memiliki kinerja yang baik dalam menangani data berdimensi tinggi, antara lain pada prediksi diabetes dengan akurasi 89,45% [6], deteksi tumor otak sebesar 92,47% [9], serta klasifikasi risiko kekambuhan kanker tiroid yang mencapai 91% [8]. Sementara itu, *K-Nearest Neighbors* (KNN) ditemukan dalam 8 penelitian dan beberapa penelitian menunjukkan kinerja yang lebih rendah dibandingkan SVM, seperti pada klasifikasi diabetes dengan akurasi 78,57% [7] dan tumor otak sebesar 85,63% [14]. Namun demikian, kinerja KNN tidak selalu lebih rendah, sebagaimana ditunjukkan pada klasifikasi anemia aplastik yang menghasilkan akurasi 97,56% setelah dikombinasikan dengan teknik SMOTE [15]. Temuan tersebut menunjukkan bahwa kinerja suatu algoritma tidak hanya ditentukan oleh karakteristik algoritmanya, tetapi juga dipengaruhi oleh karakteristik *dataset* dan teknik pengolahan data yang diterapkan.

Algoritma berbasis *ensemble dan boosting* juga menunjukkan dominasi yang signifikan. Penelitian yang menggunakan algoritma XGBoost terbukti sangat efisien untuk dataset besar dengan pola *non-linear*, mengungguli SVM pada prediksi penyakit Parkinson 92% [16] dan memberikan hasil presisi pada prediksi penipuan rantai pasok dengan *F1-score* 99,31% [4] dan harga rumah dengan nilai *F1-Score* 91,77% [13]. Sementara itu, penelitian yang menggunakan *Random Forest* (RF) sangat efektif dalam meminimalkan *overfitting* pada prediksi stroke dengan akurasi 82,07% [1] dan pada analisis sentimen transportasi publik yang dianggap sebagai model paling seimbang. Penggunaan algoritma klasik lainnya *Logistic Regression* interpretabilitas akurasi mencapai 95,02% [2] pada prediksi stroke dan *Naive Bayes* menunjukkan hasil yang berfluktuasi antara 77,40% hingga 82% tergantung pada karakteristik data [4].

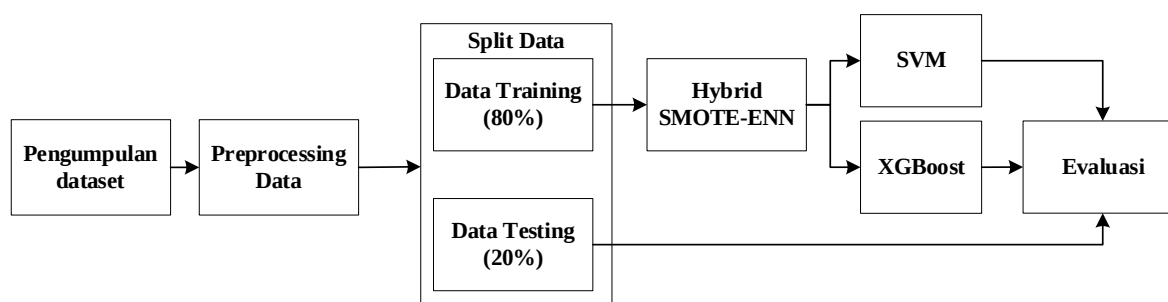
Berdasarkan tinjauan literatur, sejumlah penelitian telah menerapkan berbagai teknik penyeimbangan kelas untuk meningkatkan kemampuan *machine learning* dalam menangani *dataset* dengan distribusi kelas yang tidak seimbang. Salah satu pendekatan yang memiliki karakteristik berbeda adalah SMOTE-ENN, yang menggabungkan SMOTE untuk membentuk sampel sintetis pada kelas minoritas dengan ENN untuk menyaring data yang memiliki karakteristik tidak konsisten atau berpotensi menjadi *noise*. Kombinasi tersebut memberikan keunggulan karena proses penyeimbangan tidak hanya meningkatkan representasi kelas minoritas, tetapi juga disertai proses pembersihan data sebelum model melakukan pembelajaran. Perkembangan penelitian menunjukkan bahwa penggunaan teknik *balancing* seperti SMOTE telah banyak dikombinasikan dengan berbagai algoritma klasifikasi, termasuk SVM, *Random Forest*, dan XGBoost. Sementara itu, SMOTE-ENN mulai digunakan sebagai alternatif yang mengintegrasikan proses *oversampling* dan pembersihan data, tetapi

efektivitasnya dapat berbeda ketika diterapkan pada algoritma dengan mekanisme pembelajaran yang berbeda.

Penelitian ini dilakukan untuk menguji secara komparatif penerapan SMOTE-ENN pada SVM dan XGBoost untuk *dataset* stroke dengan ketidakseimbangan kelas yang ekstrem. Kedua algoritma dipilih karena memiliki mekanisme pembelajaran yang berbeda, yaitu SVM berbasis *kernel* dan XGBoost berbasis *ensemble boosting*. Perbandingan ini digunakan untuk memperoleh gambaran kinerja kedua algoritma setelah memperoleh perlakuan penyeimbangan kelas yang sama, tidak hanya berdasarkan metrik klasifikasi tetapi juga efisiensi waktu pelatihan. Dengan demikian, penelitian ini melengkapi penelitian terdahulu melalui analisis komparatif efektivitas SMOTE-ENN pada dua pendekatan klasifikasi yang berbeda dalam kondisi ketidakseimbangan kelas yang ekstrem.

3 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental komparatif untuk membandingkan performa algoritma *machine learning* dalam mengklasifikasikan risiko stroke. Penelitian ini menggunakan bahasa pemrograman *Python* dengan *library pandas*, *scikit-learn*, *matplotlib*, dan *imbalanced-learn*. Dalam penelitian ini menggunakan dataset rekam medis bersumber dari data publik kaggle.com yang mencakup faktor demografis dan kondisi kesehatan pasien. Penelitian yang dilakukan menerapkan tahapan yang terdiri dari pengumpulan data, preprocessing data, transformasi data, melakukan *oversampling* dan *undersampling* pada data *training*, pelatihan model *machine learning* dan evaluasi model yang tergambar pada Gambar 1 alur penelitian.



Gambar 1 Alur penelitian

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data sekunder bersifat publik yang diperoleh dari repositori *Kaggle.com* dengan judul *stroke prediction dataset*. Dataset terdiri dari 43.400 data dan 12 atribut. Pada Tabel 1 adalah definisi dan Deskripsi dengan 12 atribut yang digunakan pada penelitian.

Tabel 1 Definisi Variabel

No.	Nama	Deskripsi
1	<i>Id</i>	Indetifikasi unik untuk setiap pasien
2	<i>Gender</i>	Jenis kelamin pasien yang dikategori <i>male</i> , <i>female</i> , <i>other</i>
3	<i>Age</i>	Usia pasien dalam satuan tahun
4	<i>Hypertension</i>	Riwayat hipertensi, 1 memiliki hipertensi 0 adalah tidak
5	<i>Heart_disease</i>	Riwayat penyakit jantung
6	<i>Ever_married</i>	Status pernikahan
7	<i>work_type</i>	Jenis pekerjaan pasien
8	<i>residence_type</i>	Jenis tempat tinggal
9	<i>avg_glucose_level</i>	Kadar glukosa rata-rata dalam pasien
10	<i>bmi</i>	Indeks masa tubuh
11	<i>Smoking_status</i>	Status merokok pasien
12	<i>stroke</i>	Variabel target menunjukan mengalami stroke atau tidak

3.2 Preprocessing Data

Tahap *preprocessing* data dilakukan secara sistematis untuk menjamin kualitas, integritas, dan kesiapan dataset sebelum memasuki fase pemodelan. Tanpa pemrosesan yang tepat, model cenderung menghasilkan prediksi yang tidak akurat karena data mentah sering kali mengandung noise atau format yang tidak kompatibel dengan algoritma pembelajaran mesin [14]. Aktivitas pada *preprocessing* data terdiri dari pembersihan data (*data cleaning*) dengan menghapus atribut yang tidak relevan seperti identitas unik (*id*) untuk menyederhanakan kompleksitas data. Penanganan nilai hilang (*missing values*) untuk mempertahankan jumlah data. teknik yang digunakan adalah imputasi nilai mean atau median seperti pada atribut BMI dan pengisian kategori tertentu untuk data kategorikal yang kosong.

Aktivitas selanjutnya adalah transformasi data *feature scaling*. Variabel kategorikal dikonversi menjadi representasi numerik dengan teknik *One-Hot Encoding*. *Feature scaling* menggunakan instrumen *StandardScaler* untuk menormalisasi rentang nilai fitur, sehingga variabel dengan skala besar tidak mendominasi proses pelatihan model, terutama pada algoritma yang sensitif terhadap jarak seperti SVM [7][12][15]. *Preprocessing* sangat penting dilakukan untuk meminimalkan bias, meningkatkan efisiensi komputasi, dan memastikan model mampu melakukan generalisasi dengan baik pada data medis yang kompleks [16].

3.3 Split Data

Proses pembagian data dilakukan untuk memisahkan dataset menjadi dua bagian data latih (*training set*) yang berfungsi sebagai basis pembangunan model dan data uji (*testing set*) yang digunakan untuk mengevaluasi performa model pada data baru untuk menguji kemampuan generalisasinya [7]. Rasio pada penelitian yang dilakukan adalah 80% untuk data latih dan 20% untuk data uji. Pembagian data diimplementasikan menggunakan fungsi *train_test_split* dari library *Scikit-learn*. Dalam dataset medis yang memiliki masalah ketidakseimbangan kelas (*class imbalance*) seperti pada data prre [1], diabetes [6], dan kanker [8] penggunaan teknik *Stratified Sampling* menjadi langkah penting. Teknik ini memastikan bahwa proporsi kelas target pada himpunan data latih dan data uji tetap konsisten dengan distribusi dataset asli, sehingga mencegah terjadinya bias dan memastikan hasil evaluasi model tetap representatif.

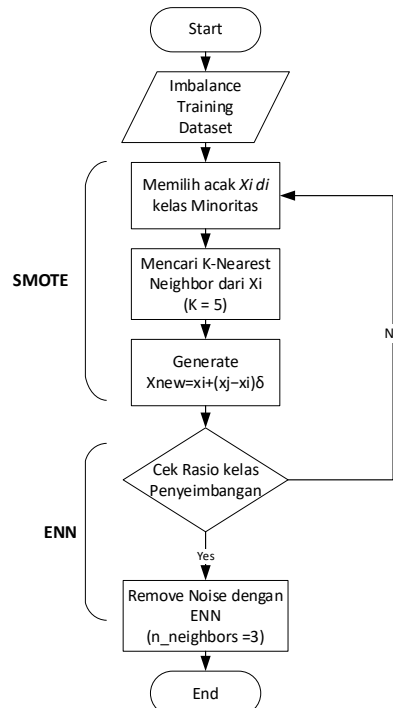
3.4 Imbalancing data

Penanganan ketidakseimbangan kelas dalam penelitian ini menggunakan *Synthetic Minority Over-sampling Technique-Edited Nearest Neighbours* (SMOTE-ENN) untuk mengatasi distribusi kelas yang sangat tidak seimbang pada dataset stroke. Berdasarkan karakteristik dataset, kelas stroke hanya terdiri atas 783 data (1,8%), sedangkan kelas tidak stroke mencapai 42.617 data (98,2%). Kondisi tersebut dapat menyebabkan model lebih dominan mempelajari karakteristik kelas mayoritas dan kurang optimal dalam mengenali kelas minoritas. SMOTE-ENN dipilih karena menggabungkan proses pembentukan sampel sintetis melalui SMOTE dengan proses penyaringan data menggunakan ENN. SMOTE berperan meningkatkan representasi kelas minoritas, sedangkan ENN digunakan untuk mengurangi observasi yang tidak konsisten atau berpotensi menjadi noise berdasarkan hubungan ketetanggaan antarobservasi.

Pada tahap SMOTE, parameter *k_neighbors* = 5 digunakan untuk menentukan lima tetangga terdekat dari observasi kelas minoritas dalam proses pembentukan sampel sintetis. Selanjutnya, ENN diterapkan dengan parameter *n_neighbors* = 3 untuk mengidentifikasi dan menghapus observasi yang tidak konsisten berdasarkan tiga tetangga terdekatnya. Parameter *k_neighbors* pada SMOTE dan *n_neighbors* pada ENN dipilih melalui pengujian beberapa konfigurasi pada data latih. Konfigurasi *k_neighbors* = 5 dan *n_neighbors* = 3 dipilih karena menghasilkan kinerja validasi terbaik berdasarkan *F1-score*. Pemilihan konfigurasi tersebut dilakukan pada data latih agar proses penentuan parameter tidak melibatkan data uji.

Setelah proses SMOTE-ENN, dilakukan *Random Under-Sampling* (RUS) untuk memperoleh distribusi kelas yang seimbang. RUS menggunakan *sampling strategy 1.0*, yang dimana jumlah kelas mayoritas disesuaikan dengan jumlah kelas minoritas setelah proses SMOTE-ENN. Hasil penerapan SMOTE-ENN pada data latih menghasilkan 33.618 data kelas stroke dan 34.094 data kelas tidak

stroke. Selanjutnya, RUS mengurangi jumlah kelas tidak stroke menjadi 33.618 data, sehingga diperoleh distribusi akhir yang seimbang sebesar 50:50, dengan total 67.236 data latih. Seluruh proses penyeimbangan kelas, termasuk SMOTE-ENN dan RUS, hanya diterapkan pada data latih, sedangkan data uji dipertahankan pada distribusi aslinya. Pendekatan tersebut diterapkan untuk mencegah data leakage dan memastikan evaluasi model mencerminkan kemampuan generalisasi terhadap data yang belum digunakan dalam proses penyeimbangan maupun pelatihan.



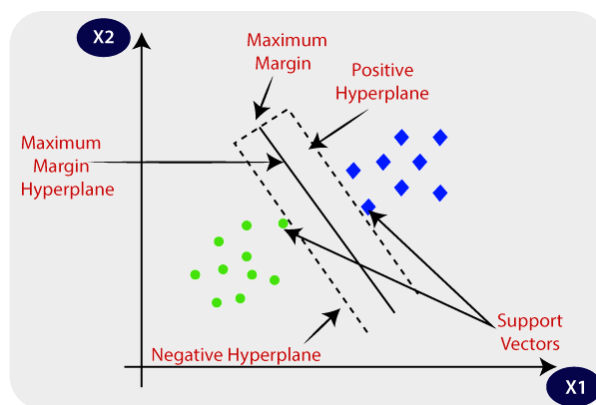
Gambar 2 Flowchart SMOTE-ENN

3.5 Model Klasifikasi

Penelitian ini menggunakan SVM dan XGBoost karena keduanya memiliki mekanisme pembelajaran yang berbeda, sehingga dapat digunakan untuk memperoleh perbandingan kinerja model setelah penerapan SMOTE-ENN.

3.5.1 Support Vector Machine (SVM)

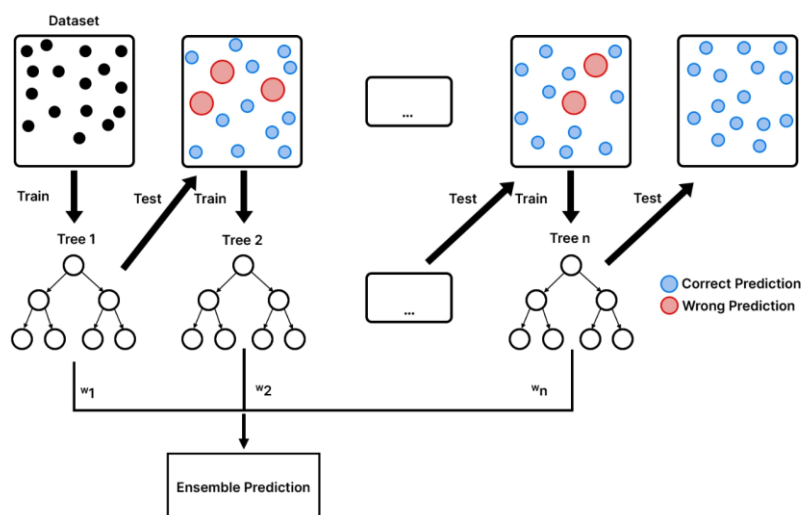
Support Vector Machine (SVM) merupakan algoritma klasifikasi yang mencari hyperplane optimal untuk memisahkan data dari dua kelas. *Hyperplane* dipilih dengan memaksimalkan *margin*, yaitu jarak antara *hyperplane* dan observasi terdekat dari masing-masing kelas yang disebut *support vectors*. SVM sangat efektif dalam menangani data dengan hubungan *non-linear* melalui penggunaan fungsi kernel [17]. *Kernel RBF* digunakan dalam penelitian ini untuk menangani pola data klinis yang kompleks dan *non linear*. Mekanisme klasifikasi SVM dalam memisahkan data berdasarkan *hyperplane* dan *support vectors* ditunjukkan pada Gambar 3.



Gambar 3 Klasifikasi support vector machine [18]

3.5.2 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) merupakan algoritma ensemble berbasis *gradient boosting* yang membangun sejumlah pohon keputusan secara bertahap. Setiap pohon baru digunakan untuk memperbaiki kesalahan prediksi yang dihasilkan oleh model sebelumnya [2]. XGBoost bekerja secara iteratif dengan menggabungkan banyak model, umumnya berupa pohon keputusan atau *Classification and Regression Trees* (CART) untuk membentuk satu model prediksi [3]. Kelebihan utama dari XGBoost adalah kecepatan dan efisiensi, penanganan nilai hilang dan fleksibilitas yang dapat digunakan untuk berbagai tugas seperti klasifikasi regresi dan perankingan. Arsitektur XGBoost dalam membangun model secara bertahap melalui kombinasi sejumlah pohon keputusan ditunjukkan pada Gambar 4.



Gambar 4 Arsitektur XGBoost [9]

3.6 Evaluasi Model Klasifikasi

Evaluasi model dilakukan untuk mengetahui kemampuan SMOTE-ENN dan SVM dan SMOTE-ENN dan XGBoost dalam membedakan kelas stroke dan tidak stroke. penelitian yang dilakukan berfokus pada data yang tidak seimbang, evaluasi tidak hanya menggunakan *accuracy*, tetapi juga mempertimbangkan *precision*, *recall*, *F1-score*, *specificity*, serta *confusion matrix*. Selain performa klasifikasi, waktu pelatihan digunakan untuk membandingkan efisiensi komputasi kedua model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

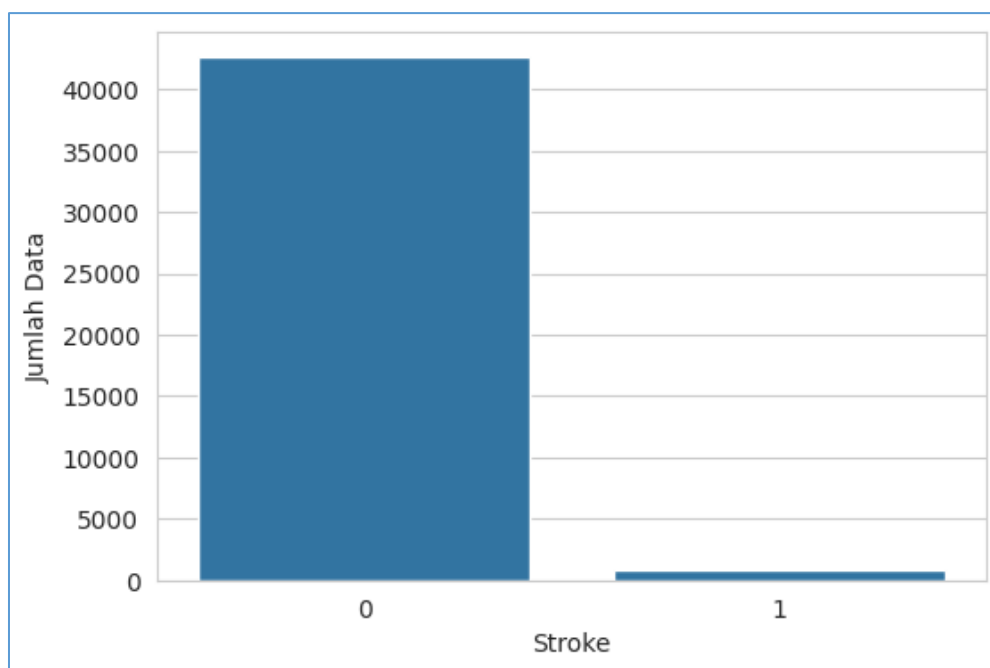
4 Hasil dan Pembahasan

Berikut adalah penyajian hasil penelitian dan pembahasan yang disusun secara sistematis sesuai dengan tahapan metode penelitian yang telah dilakukan untuk mengetahui kinerja model pada prediksi stroke. Selanjutnya, akan diinterpretasikan dengan matriks evaluasi.

4.1 Analisis Karakteristik Dataset

Sebelum melakukan preprocessing data dilakukan analisis terhadap *dataset* yang digunakan. Dalam penelitian ini dataset stroke memiliki data dengan total 43.400 data dan 12 atribut/variabel yang mencakup parameter demografis dan kondisi klinis pasien. Analisis dimulai dengan melihat kualitas data bahwa tidak ditemukan data duplikat, tetapi teridentifikasi adanya missing values pada atribut bmi sebanyak 1.462 data atau 3,37% pada atribut smoking_status sebanyak 13.292 atau 30,63%.

Pada Gambar 5 menunjukkan karakteristik paling terlihat dari dataset ini adalah distribusi dari label target stroke yang memiliki kategori 1 yang berarti stroke dan 0 adalah yang tidak stroke. Fenomena ketidakseimbangan kelas yang sangat ekstrem pada variabel target stroke, di mana kasus positif stroke hanya mencakup 1,8% atau 783 data dibandingkan dengan kasus tidak stroke yang mencapai 98,2% atau 42.617 data.



Gambar 5 Distribusi label target stroke

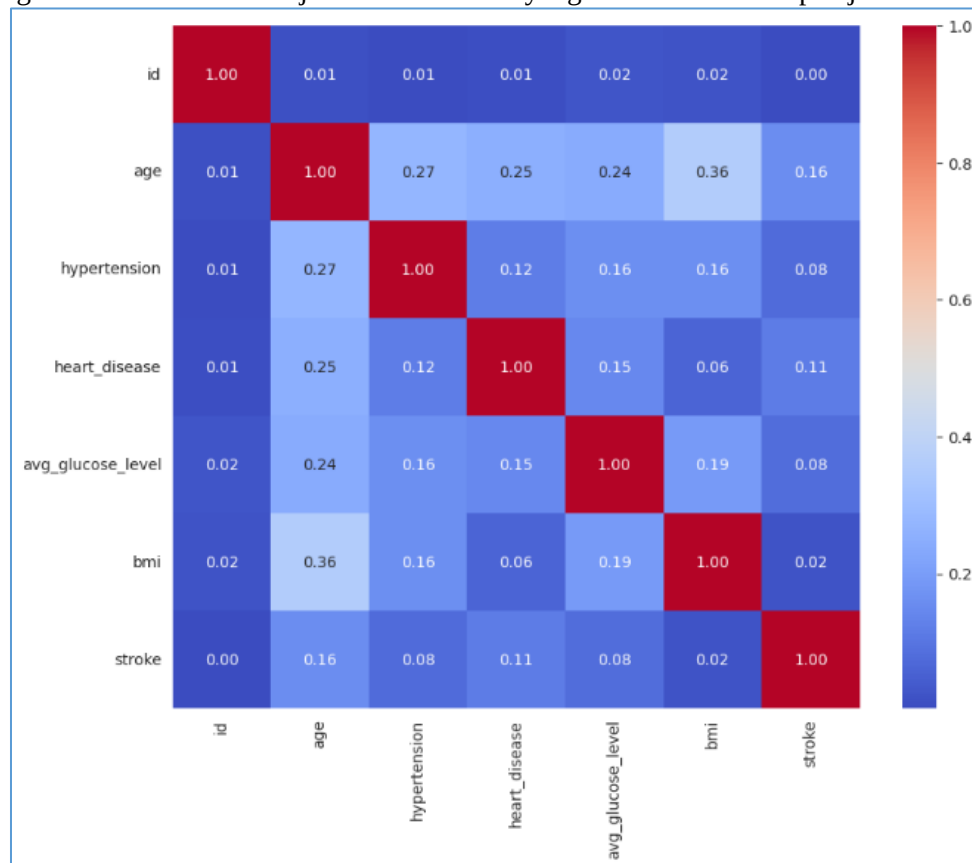
Selanjutnya, dilakukan analisis karakteristik atribut numerik berdasarkan statistik deskriptif yang disajikan pada Tabel 2. Hasil analisis menunjukkan variasi distribusi yang luas, seperti usia pasien yang mencakup rentang dari bayi hingga lansia dan adanya nilai ekstrem (*outliers*) pada variabel *avg_glucose_level* dan *bmi*.

Tabel 2 Deskripsi data

Atribut	Count	Mean	Min	25%	50%	75%	Max
---------	-------	------	-----	-----	-----	-----	-----

age	43400.0	42.217894	0.08	24.00	44.00	60.00	82.00
hypertension	43400.0	0.093571	0.00	0.00	0.00	0.00	1.00
heart_disease	43400.0	0.047512	0.00	0.00	0.00	0.00	1.00
avg_glucose_level	43400.0	104.482750	55.00	77.54	91.58	112.07	291.05
bmi	41938.0	28.605038	10.10	23.20	27.70	32.90	97.60
stroke	43400.0	0.018041	0.00	0.00	0.00	0.00	1.00
age	43400.0	42.217894	0.08	24.00	44.00	60.00	82.00

Pada Gambar 6 analisis korelasi melalui *heatmap* menunjukkan bahwa tidak terdapat multikolinearitas antar fitur independen dengan nilai korelasi di bawah 0,8. pada atribut hipertensi, dan kadar glukosa rata-rata menunjukkan keterkaitan yang lebih kuat terhadap kejadian stroke.



Gambar 6 Heatmap EDA

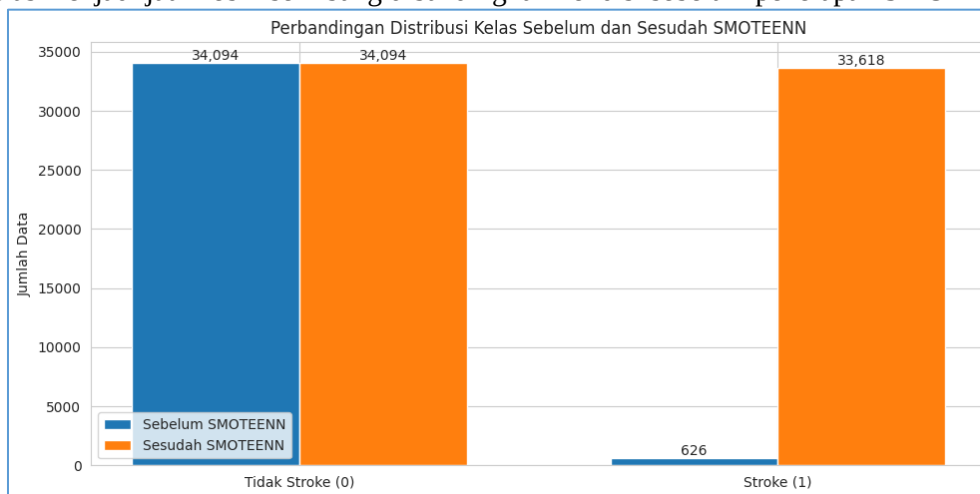
Secara keseluruhan, karakteristik dataset ini memerlukan tahap *preprocessing* yang intensif, terutama penyeimbangan kelas, guna memastikan model tidak bias terhadap kelas mayoritas

4.2 Analisis Deskriptif Preprocessing data

Tahap *preprocessing* data dalam penelitian ini dilakukan secara sistematis untuk meningkatkan kualitas dan kesiapan dataset stroke yang berjumlah 43.400 data. Langkah awal melakukan pembersihan data dengan menghapus atribut id yang tidak relevan secara klinis. Untuk menangani nilai hilang pada atribut bmi 3,37% menggunakan imputasi median dan atribut *smoking_status* 30,63% dengan pengisian kategori 'Unknown' menjaga integritas jumlah sampel. Selanjutnya, dilakukan transformasi data di mana variabel kategorikal dikonversi menjadi format numerik menggunakan teknik *One-Hot Encoding* agar dapat diproses oleh algoritma *machine learning*. Setelah dilakukan *preprocessing* membagi data pada penelitian prediksi stroke dengan total 43.400 data, pembagian data latih dan data uji menggunakan *rasio* 80:20. Pada kondisi *baseline* sebelum penyeimbangan kelas dilakukan, *rasio* ini menghasilkan 34.720 data latih dan 8.680 data uji.

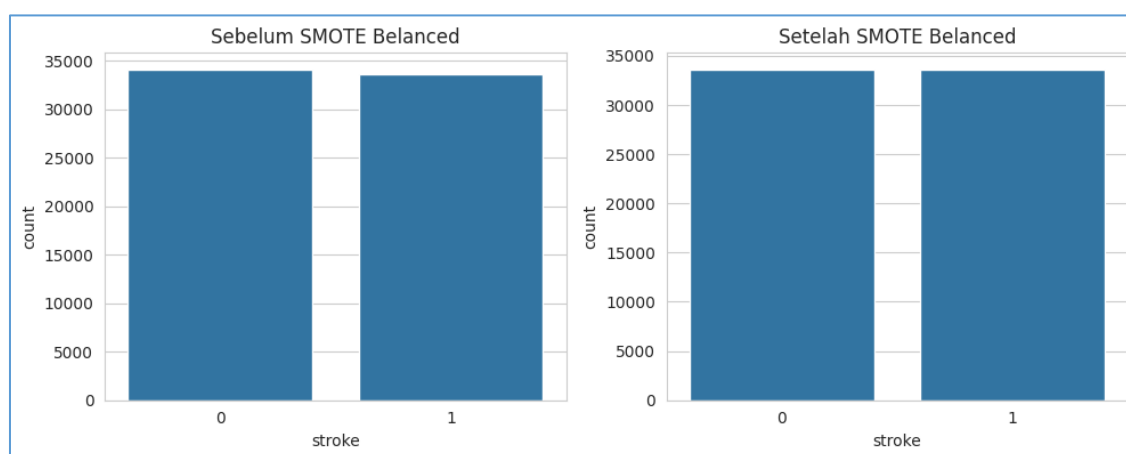
4.3 Hasil Oversampling SMOTE-ENN

Penerapan teknik SMOTE-ENN dalam penelitian ini untuk mengatasi masalah ketidakseimbangan kelas yang ekstrem pada dataset stroke. Proses SMOTE-ENN hanya diterapkan pada data latih untuk mencegah terjadinya *data leakage* terhadap data uji. Sebelum SMOTE-ENN data latih terdiri atas 34.094 data pada kelas mayoritas tidak stroke dan 626 data pada kelas minoritas stroke. Setelah SMOTE-ENN Jumlah kelas minoritas stroke meningkat secara signifikan menjadi 33.618 data, sementara kelas mayoritas tidak stroke menjadi 34.094 data. Dengan demikian, distribusi kedua kelas menjadi jauh lebih seimbang dibandingkan kondisi sebelum penerapan SMOTE-ENN.



Gambar 7 Perbandingan Distribusi kelas sebelum dan sesudah SMOTE-ENN

Pada Gambar 7 menunjukkan perubahan distribusi kelas setelah penerapan SMOTE-ENN, di mana kesenjangan jumlah antara kelas tidak stroke dan stroke berkurang secara signifikan. Kondisi tersebut menghasilkan distribusi kelas yang lebih seimbang sebagai data masukan untuk proses pelatihan model SVM dan XGBoost. Selanjutnya dilakukan *Random under sampling* Kelas mayoritas dikurangi secara acak sehingga kedua kelas memiliki jumlah seimbang, masing-masing sebanyak 33.618 data dengan distribusi 50:50.



Gambar 8 Perbandingan sebelum dan sesudah *random under sampling*

4.4 Implementasi dan Performa Model Klasifikasi

Pada skenario pertama, model klasifikasi dibangun menggunakan dataset tanpa menerapkan teknik balancing sebagai model *baseline*. Dataset terdiri atas 43.400 data, dengan distribusi kelas tidak stroke sebanyak 42.617 data atau 98,2% dan kelas stroke 1 sebanyak 783 data atau 1,8%. Data kemudian dibagi menggunakan stratified split dengan proporsi 80:20, sehingga diperoleh 34.720 data latih dan 8.680 data uji. Penggunaan *stratified split* bertujuan untuk mempertahankan proporsi masing-masing kelas pada data latih dan data uji. implementasi model yang digunakan terdiri atas

<http://sistemasi.ftik.unisi.ac.id>

XGBoost dan SVM. XGBoost diimplementasikan menggunakan XGBClassifier dengan *random_state=42* dan *eval_metric='logloss'*. Sementara itu, SVM menggunakan SVC dengan *kernel RBF* (*kernel='rbf'*) dan fitur numerik yang telah dinormalisasi menggunakan StandardScaler. Pada Tabel 3 menunjukkan hasil pengujian XGBoost memperoleh akurasi sebesar 98,13%, sedangkan SVM memperoleh akurasi sebesar 98,19%. Model baseline menghasilkan akurasi yang tinggi tetapi kedua model belum mampu mendeteksi kelas stroke dengan baik. Pada kelas minoritas, XGBoost maupun SVM menghasilkan nilai *Precision*, *Recall*, dan *F1-Score* sebesar 0,00%. Hasil tersebut menunjukkan bahwa akurasi yang tinggi pada kondisi ketidakseimbangan kelas belum merepresentasikan kemampuan model dalam mengenali kasus stroke. Oleh karena itu, skenario *baseline* ini digunakan sebagai pembandingan terhadap model yang menerapkan teknik penyeimbangan data pada skenario berikutnya.

Tabel 3 Hasil implementasi *baseline*

Model	Accuracy	Precision	Recall	F1-Score
SVM	98,18%	0 %	0 %	0 %
XGBoost	98.13%	0 %	0 %	0 %

Untuk mengatasi ketidakseimbangan kelas yang menyebabkan bias pada model *baseline*, teknik SMOTE-ENN diterapkan pada data latih untuk mencegah terjadinya data *leakage* dan meningkatkan representasi kelas minoritas tanpa memengaruhi data uji. Hasil proses penyeimbangan menghasilkan distribusi kelas yang seimbang, dengan masing-masing kelas berjumlah 33.618 data. Pada model SMOTE-ENN dan XGBoost, data latih hasil SMOTE-ENN digunakan untuk melatih XGBClassifier. Berdasarkan Tabel 4 model SMOTE-ENN-XGBoost menunjukkan kinerja yang lebih baik dibandingkan SMOTE-ENN-SVM pada sebagian besar metrik evaluasi. SMOTE-ENN dengan XGBoost memperoleh *accuracy* sebesar 96,96%, *precision* 96,83%, *recall* 97,11%, dan *F1-score* 96,97%, dengan waktu pelatihan 4,6781 detik. Hasil tersebut menunjukkan bahwa model SMOTE-ENN dengan XGBoost memberikan performa klasifikasi yang tinggi sekaligus memiliki waktu pelatihan yang relatif singkat dibandingkan model SMOTE-ENN dengan SVM.

Tabel 4 Hasil implementasi SMOTE-ENN dengan *machine learning*

Model	Accuracy	Precision	Recall	F1-Score	Waktu Komputasi
SMOTEENN-XGBoost	96,96%	96,82%	97,10%	96,96%	4,67 Detik
SMOTEENN-SVM	90,12%	88,07%	92,83%	90,38%	21,2 Menit

Sementara itu, model SMOTE-ENN dan SVM menggunakan SVC dengan kernel RBF dan parameter *C=10* serta *gamma='scale'*. Setelah proses standarisasi fitur, model menghasilkan akurasi 90,13%, *Precision* 88,07%, *Recall* 92,83%, dan *F1-Score* 90,39%. Namun, waktu pelatihan SVM mencapai 21,2 menit, sehingga jauh lebih lama dibandingkan XGBoost. Hasil tersebut menunjukkan bahwa penerapan SMOTE mampu meningkatkan kemampuan kedua algoritma dalam mengenali kelas minoritas dibandingkan model *baseline*, dengan XGBoost memberikan performa keseluruhan dan efisiensi komputasi yang lebih baik.

4.5 Evaluasi

Berdasarkan Gambar 9 hasil *classification report*, model SMOTE-ENN dengan XGBoost menunjukkan performa yang lebih unggul dibandingkan SMOTE-ENN dengan SVM dalam melakukan klasifikasi data stroke. Pada XGBoost, kelas tidak stroke memperoleh *Precision* 97,10%, *Recall* 96,82%, dan *F1-Score* 96,96%, sedangkan kelas stroke memperoleh *Precision* 96,83%, *Recall* 97,11%, dan *F1-Score* 96,97%, dengan akurasi keseluruhan sebesar 96,96%. Nilai *Recall* kelas stroke sebesar 97,11% menunjukkan bahwa model mampu mengenali sebagian besar data pasien stroke dengan tingkat sensitivitas yang tinggi. Sementara itu, model SMOTE-ENN dan SVM menghasilkan akurasi sebesar 90,13%, dengan *Precision*, *Recall*, dan *F1-Score* pada kelas stroke masing-masing sebesar 88,07%, 92,83%, dan 90,39%. Meskipun SVM memiliki *Recall* kelas stroke yang cukup

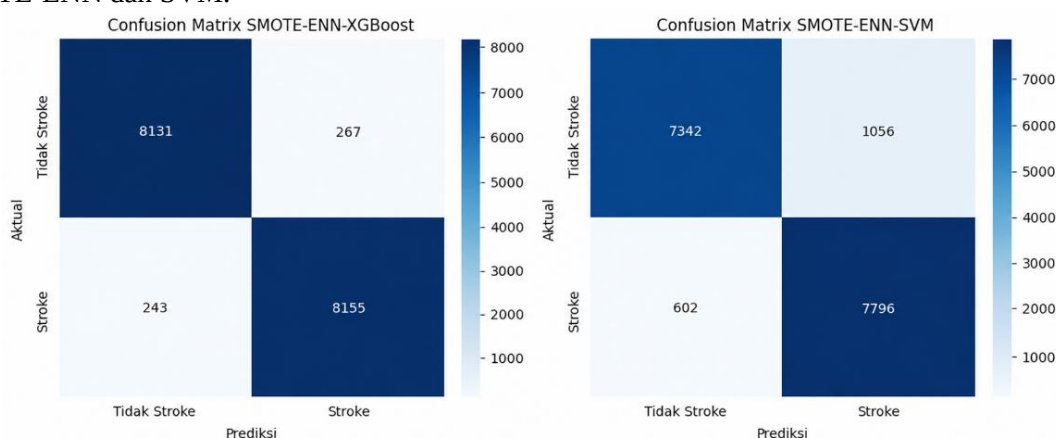
tinggi, performanya masih berada di bawah XGBoost pada seluruh metrik utama. Hal ini menunjukkan bahwa SMOTE-ENN dan XGBoost memberikan kinerja klasifikasi yang lebih konsisten dan seimbang dalam membedakan kelas stroke dan tidak stroke, sebagaimana terlihat dari nilai *Precision*, *Recall*, dan *F1-Score* yang relatif tinggi dan hampir seimbang pada kedua kelas.

=== SMOTE-XGBoost ===				
	precision	recall	f1-score	support
Tidak Stroke	0.9710	0.9682	0.9696	8398
Stroke	0.9683	0.9711	0.9697	8398
accuracy			0.9696	16796
macro avg	0.9696	0.9696	0.9696	16796
weighted avg	0.9696	0.9696	0.9696	16796

=== SMOTE-SVM ===				
	precision	recall	f1-score	support
Tidak Stroke	0.9242	0.8743	0.8985	8398
Stroke	0.8807	0.9283	0.9039	8398
accuracy			0.9013	16796
macro avg	0.9025	0.9013	0.9012	16796
weighted avg	0.9025	0.9013	0.9012	16796

Gambar 9 Klasifikasi report oversampling dan machine learning

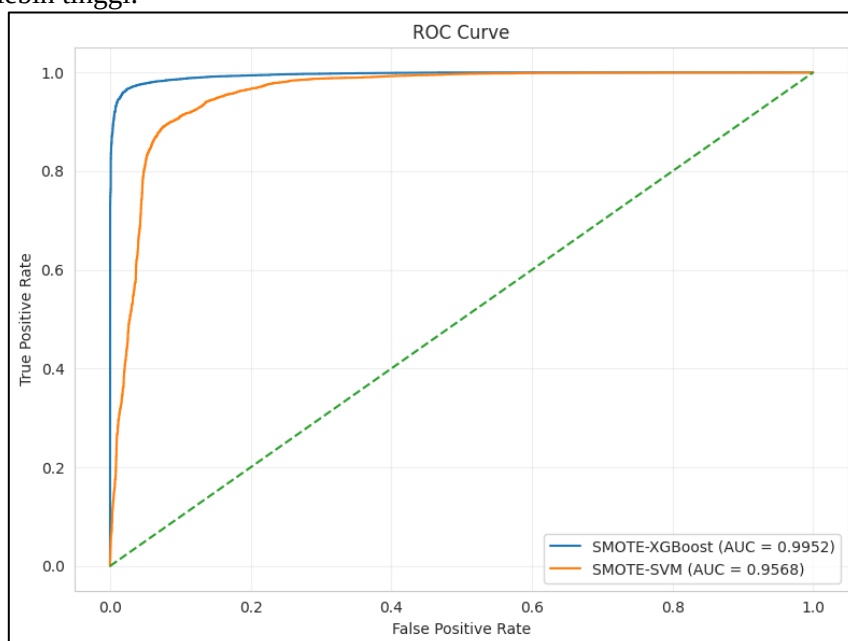
Berdasarkan *confusion matrix*, SMOTE-ENN dan XGBoost menunjukkan performa yang lebih baik dibandingkan SMOTE-ENN-SVM. XGBoost menghasilkan 8.155 *True Positive* (TP), 8.131 *True Negative* (TN), 243 *False Negative* (FN), dan 267 *False Positive* (FP), sedangkan SVM menghasilkan 7.796 TP, 7.342 TN, 602 FN, dan 1.056 FP. Jumlah *False Negative* yang lebih rendah pada XGBoost menunjukkan kemampuan yang lebih baik dalam mendeteksi pasien stroke sekaligus mengurangi risiko kasus stroke yang tidak teridentifikasi. Selain performa klasifikasi, XGBoost juga lebih efisien secara komputasi dengan waktu pelatihan 4,6781 detik, jauh lebih cepat dibandingkan SVM yang membutuhkan 1.273,1572 detik atau sekitar 21,2 menit. Dengan demikian, SMOTE-ENN dan XGBoost memberikan performa klasifikasi dan efisiensi komputasi yang lebih baik dibandingkan SMOTE-ENN dan SVM.



Gambar 10 Confusion Matrix dengan SMOTE-ENN

Kurva *Receiver Operating Characteristic* (ROC) digunakan untuk mengevaluasi kemampuan model dalam membedakan kelas stroke dan tidak stroke pada berbagai nilai *decision threshold*. Berdasarkan Gambar 11 kurva ROC, SMOTE-ENN dan XGBoost dengan nilai ROC-AUC 0,9952 menunjukkan kemampuan diskriminasi yang lebih baik dibandingkan SMOTE-ENN dan SVM

dengan ROC-AUC 0,9568, yang ditunjukkan oleh kurva yang lebih mendekati sudut kiri atas dan nilai AUC yang lebih tinggi.



Gambar 11 ROC-AUC dengan SMOTE ENN

Hasil tersebut menunjukkan bahwa XGBoost memiliki kemampuan yang lebih baik dalam membedakan pasien stroke dan tidak stroke secara konsisten pada berbagai ambang keputusan, dengan tingkat *False Positive* yang relatif lebih rendah. Dengan demikian, berdasarkan evaluasi ROC-AUC, SMOTE-ENN dan XGBoost memberikan kemampuan diskriminasi yang lebih optimal dibandingkan SMOTE-ENN dan SVM.

Keunggulan model dalam penelitian yang dilakukan dibandingkan dengan penelitian terdahulu yang relevan. Beberapa penelitian, seperti klasifikasi stroke [1] prediksi diabetes [6] [7] dan klasifikasi anemia [15] menggunakan SMOTE sebagai teknik *oversampling*. Berbeda dengan pendekatan tersebut, penelitian ini menggunakan SMOTE-ENN, yang tidak hanya meningkatkan jumlah data kelas minoritas melalui pembentukan data sintesis, tetapi juga menggunakan ENN untuk membersihkan *noise* dan data yang tumpang tindih pada batas antarkelas. Pendekatan ini menghasilkan performa SMOTE-ENN dan XGBoost dengan akurasi 96,96%, *Precision* 96,83%, *Recall* 97,11%, dan *F1-Score* 96,97%. Perbandingan dengan penelitian terdahulu menunjukkan bahwa penelitian ini memiliki karakteristik eksperimen yang berbeda. Beberapa penelitian sebelumnya menggunakan SMOTE yang dikombinasikan dengan *Random Forest* maupun *AdaBoost* untuk menangani ketidakseimbangan kelas dengan akurasi 82,07% dan *AdaBoost* dan SMOTE dengan akurasi 82,83% [1] [6]. Dengan demikian, perbandingan hasil tidak didasarkan pada nilai akurasi yang lebih tinggi tetapi pada kemampuan model dalam mengenali kelas minoritas, keseimbangan performa antar kelas, dan efisiensi komputasi.

Kelebihan penelitian ini juga terletak pada skala dataset dan pendekatan evaluasi yang digunakan. Penelitian ini menggunakan 43.400 observasi, yang relatif lebih besar dibandingkan beberapa penelitian terdahulu, seperti dataset diabetes Pima Indians sebanyak 768 data, dataset Parkinson sebanyak 195 data, serta penelitian anemia yang menggunakan 200 data. Penggunaan *dataset* yang lebih besar memberikan basis evaluasi yang lebih luas terhadap kemampuan model. Selain itu, penelitian ini tidak hanya mengevaluasi model berdasarkan akurasi, tetapi juga menggunakan *Precision*, *Recall*, *F1-Score*, *Specificity*, *Confusion Matrix*, ROC-AUC, dan *Precision-Recall Curve* dan membandingkan waktu komputasi kedua algoritma. Dengan demikian, penelitian ini memiliki keunikan pada kombinasi SMOTE-ENN, XGBoost, dan evaluasi multidimensi, sehingga tidak hanya menilai ketepatan klasifikasi, tetapi juga kemampuan model dalam mendeteksi kelas stroke dan efisiensi komputasinya. Hasil penelitian menunjukkan bahwa SMOTE-ENN dan XGBoost merupakan model terbaik dengan performa klasifikasi yang tinggi dan waktu pelatihan yang jauh lebih efisien dibandingkan SMOTE-ENN dan SVM.

5 Kesimpulan

Penelitian ini menunjukkan bahwa penerapan teknik penyeimbangan data SMOTE-ENN berhasil mengatasi masalah ketidakseimbangan kelas yang sangat ekstrem pada dataset stroke 1,8% stroke dan 98,2% tidak stroke menjadi seimbang. Berdasarkan pengujian yang dilakukan model SMOTE-ENN dan XGBoost memiliki performa terbaik dengan akurasi 96,96%, *recall* 97,11%, dan *F1-score* 96,97%, mengungguli model SMOTE-ENN dan SVM yang memperoleh akurasi 90,13% dan *recall* 92,83%. Selain itu, XGBoost lebih efisiensi komputasi dengan waktu pelatihan hanya 4,68 detik, sedangkan SVM membutuhkan waktu komputasi yang sangat lama hingga 21,2 menit. Dengan demikian, integrasi metode SMOTE-ENN dengan algoritma XGBoost menghasilkan performa klasifikasi terbaik dibandingkan SMOTE-ENN dan SVM dan model *baseline*, sekaligus mampu mengatasi ketidakseimbangan kelas dan memperbaiki kelemahan model dalam mendeteksi kelas minoritas.

Referensi

- [1] R. M. D. Saputra, Y. Chairul, D. Riana, A. S. Hewiz, and F. Aziz, "Stroke Prediction based on Random Forest with SMOTE," *2023 Int. Conf. Inf. Technol. Res. Innov. ICITRI 2023*, pp. 17–21, 2023, DOI: 10.1109/ICITRI59340.2023.10249261.
- [2] K. Sundaram, B. Lanitha, K. Kamaraj, and A. K. Ramamoorthy, "Enhanced Brain Stroke Prediction: An Ensemble of Random Forest, Logistic Regression and XGBoost," *2nd Int. Conf. Emerg. Res. Comput. Sci. ICERCS 2024*, pp. 1–5, 2024, DOI: 10.1109/ICERCS63125.2024.10895046.
- [3] Y. Wang and Y. Guo, "Forecasting Method of Stock Market Volatility in Time Series Data based on Mixed Model of ARIMA and XGBoost," *China Commun.*, Vol. 17, No. 3, pp. 205–221, 2020, DOI: 10.23919/JCC.2020.03.017.
- [4] Y. Zhou, X. Song, and M. Zhou, "Supply Chain Fraud Prediction based on XGBoost Method," *2021 IEEE 2nd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. ICBAIE 2021*, No. Icbaie, pp. 539–542, 2021, DOI: 10.1109/ICBAIE52039.2021.9389949.
- [5] K. Subramanian, J. B. Rawlings, and C. T. Maravelias, "Economic Model Predictive Control for Inventory Management in Supply Chains," *Comput. Chem. Eng.*, Vol. 64, pp. 71–80, 2014, DOI: 10.1016/j.compchemeng.2014.01.003.
- [6] S. Sidiq, Alfian, and N. Shobi Mabur, "Pengembangan Model Prediksi Risiko Diabetes menggunakan Pendekatan AdaBoost dan Teknik Oversampling SMOTE," *J. Ilm. Inform. Dan Ilmu Komput.*, Vol. 4, pp. 13–23, 2025, [Online]. Available: <https://doi.org/10.58602/jima-ilkom.v4i1.41>
- [7] L. Oktovianus, A. Linardi, B. A. Makalew, and I. K. Wairooy, "Comparative Analysis of KNN and SVM Algorithms for Diabetes Prediction," *Proc. - 6th Int. Conf. Informatics, Multimedia, Cyber Inf. Syst. ICIMCIS 2024*, pp. 631–636, 2024, DOI: 10.1109/ICIMCIS63449.2024.10957572.
- [8] P. Budi Utomo, M. Faruqziddan, E. Herdika Septa Aulia, and S. Dini Azzahra, "Perbandingan Skenario Balancing Oversampling dan Undersampling dalam Klasifikasi Risiko Kambuh Kanker Tiroid menggunakan Algoritma SVM Linear," *JAMI J. Ahli Muda Indones.*, Vol. 5, No. 2, pp. 172–182, 2024, DOI: 10.46510/jami.v5i2.320.
- [9] F. I. Kurniadi and H. A. Saputri, "Feature Selection using XGBoost on METABRIC Dataset for Survivability Breast Cancer Detection," *Proc. 2024 Int. Conf. Inf. Manag. Technol. ICIMTech 2024*, pp. 259–262, 2024, DOI: 10.1109/ICIMTech63123.2024.10780791.
- [10] T. N. Deepthi, S. Sharmila, M. Swarna, M. Gouthami, and C. L. Akshaya, "Prediction of Brain Stroke in Human Beings using Machine Learning," *Proc. 2023 2nd Int. Conf. Electron. Renew. Syst. ICEARS 2023*, pp. 1314–1320, 2023, DOI: 10.1109/ICEARS56392.2023.10085128.
- [11] U. K. Giri, M. Saxena, D. Bordoloi, R. Tripathy, S. K. Mohapatra, and P. Kumar Sarangi, "Machine Learning Model for Heart Disease Detection: A Comparative Analysis of SVM vs

- KNN,” 2024 *Int. Conf. Electr. Electron. Comput. Technol. ICEECT 2024*, Vol. 1, pp. 1–5, 2024, DOI: 10.1109/ICEECT61758.2024.10739169.
- [12] S. Rachel, S. Monicka, B. Abirami, and A. Kiruthika, “Heart Disease Prediction of an Individual using SVM Algorithm,” 2025 *Int. Conf. Comput. Commun. Technol. ICCCT 2025*, pp. 1–4, 2025, DOI: 10.1109/ICCCT63501.2025.11019575.
- [13] L. M. I. Leo Joseph, E. H. Reddy, M. Srinidhi, P. V. Saketh, D. Mohan, and C. Rajendra Prasad, “Predicting Real-Time House Prices: A Machine Learning Approach using XGBoost Algorithm,” 2024 *Asia Pacific Conf. Innov. Technol. APCIT 2024*, pp. 1–4, 2024, DOI: 10.1109/APCIT62007.2024.10673571.
- [14] J. K. Kanimozhi, G. Kalasurya, P. Anitha, R. Geetha, E. Loganathan, and C. A. Kandasamy, “Comparing SVM and KNN Algorithm for High Accuracy in Brain Tumor Prediction,” 2025 *Int. Conf. Comput. Technol. Data Commun. ICCTDC 2025*, pp. 1–6, 2025, DOI: 10.1109/ICCTDC64446.2025.11158881.
- [15] F. A. Kusuma and B. Prasetyo, “Pemodelan Klasifikasi Anemia Aplastik menggunakan Teknik Oversampling dan K-Nearest Neighbors,” *J. Inform. dan Tek. Elektro Terap.*, Vol. 12, No. 3, 2024, DOI: 10.23960/jitet.v12i3.4326.
- [16] K. Sulthana, F. Begum, G. B. Regulwar, S. K. Rout, V. Dasari, and B. Abhilash, “Parkinson’s Disease Prediction using XGBoost and SVM,” 2023 *2nd Int. Conf. Ambient Intell. Heal. Care, ICAIHC 2023*, 2023, DOI: 10.1109/ICAIHC59020.2023.10431437.
- [17] R. C. Hardika, F. Anggara, and M. A. M. Abrori, “Prediksi Safety Stock menggunakan Algoritma Support Vector Regression dengan Optimasi Hyperparameter,” *Progresif J. Ilm. Komput.*, Vol. 22, No. 1, p. 64, 2026, DOI: 10.35889/progresif.v22i1.3386.
- [18] F. Anggara, R. C. Hardika, A. P. Rahayu, and M. A. M. Abrori, “Efektivitas Kinerja Algoritma SVM dengan Teknik ADASYN untuk Deteksi Gangguan Kecemasan Mahasiswa,” *J. Nas. Komputasi dan Teknol. Inf.*, Vol. 9, No. 1, pp. 45–53, 2026, DOI: <https://doi.org/10.32672/jnkti.v9i1.10244>.