BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 532-541
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.612

Penerapan Metode ADASYN Dalam Mengatasi Imbalanced Data Untuk Klasifikasi Penyakit Stroke Menggunakan Support Vector Machine

Alwaliyanto, Siska Kurnia Gusti*, Iis Afrianty, Fadhilah Syafria

Fakultas Sains dan Teknologi, Prodi Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]12050116868@students.uin-suska.ac.id, [2,*]siskakurniagusti@uin-suska.ac.id, [3]iis.afrianty@uin-suska.ac.id, [3]fadhilah.syafria@uin-suska.ac.id
Email Penulis Korespondensi: siskakurniagusti@students.uin-suska.ac.id

Abstrak−Stroke merupakan salah satu penyebab utama kematian dan disabilitas di seluruh dunia, sehingga diperlukan model klasifikasi yang mampu membantu diagnosis secara dini dan akurat. Penelitian ini bertujuan untuk menerapkan *algoritma Support Vector Machine* (SVM) dengan tiga jenis kernel linear, polynomial, dan Radial Basis Function (RBF) untuk mengklasifikasikan data penyakit stroke. Metode *Adaptive Synthetic Sampling* (ADASYN) digunakan untuk mengatasi masalah ketidakseimbangan data, sementara proses pelatihan dan evaluasi model dilakukan menggunakan validasi silang 5-*Fold Cross Validation* untuk memastikan hasil yang stabil dan dapat diandalkan. Hasil penelitian menunjukkan bahwa ADASYN berhasil meningkatkan sensitivitas model terhadap kelas stroke (kelas minoritas), yang tercermin dari peningkatan nilai *recall* dan *F1-score*, meskipun disertai dengan sedikit penurunan pada akurasi keseluruhan sebuah *trade-off* yang umum dalam penanganan data tidak seimbang. Kernel linear (sesudah ADASYN) memberikan performa terbaik setelah penanganan ketidakseimbangan data, dengan rata-rata AUC-ROC sebesar 0.8333, *recall* 0.7827, dan *F1-score* 0.2181 untuk kelas stroke. Meskipun nilai F1-score masih tergolong rendah, nilainya meningkat dibanding sebelum penerapan ADASYN, menunjukkan adanya perbaikan dalam pendeteksian kasus stroke. Implementasi dilakukan menggunakan Google Colab, yang turut mendukung efisiensi dalam pengolahan dan visualisasi data. Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kombinasi SVM dan ADASYN efektif dalam meningkatkan sensitivitas model terhadap kelas minoritas, dan layak diterapkan dalam proses klasifikasi data medis untuk diagnosis penyakit stroke berbasis pembelajaran mesin.

Kata Kunci: Adaptive Synthetic Sampling Approach; Imbalanced Data; K-fold Cross Validation; Stroke; Support Vector Machine

Abstract−Stroke is one of the leading causes of death and disability worldwide, making it essential to develop classification models that can assist in early and accurate diagnosis. This study aims to implement the Support Vector Machine (SVM) algorithm with three types of kernels linear, polynomial, and Radial Basis Function (RBF) to classify stroke disease data. The Adaptive Synthetic Sampling (ADASYN) method is employed to address the class imbalance problem, while model training and evaluation are carried out using 5-Fold Cross-Validation to ensure stable and reliable results. The findings indicate that ADASYN successfully improves the model's sensitivity to stroke cases (the minority class), as reflected by an increase in recall and F1-score, despite a slight decrease in overall accuracy a common trade-off in handling imbalanced data. The linear kernel (after ADASYN) achieved the best performance after imbalance handling, with an average AUC-ROC of 0.8333, recall of 0.7827, and F1-score of 0.2181 for the stroke class. Although the F1-score remains relatively low, it improved compared to the pre-ADASYN results, indicating better detection of stroke cases. The implementation was conducted using Google Colab, which also contributed to efficient data processing and visualization. Overall, the results demonstrate that the combination of SVM and ADASYN is effective in enhancing the model's sensitivity to minority classes and is well-suited for medical data classification tasks, particularly in the early diagnosis of stroke using machine learning approaches.

Keywords: Adaptive Synthetic Sampling Approach; Imbalanced Data; K-fold Cross Validation; Stroke; Support Vector Machine

1. PENDAHULUAN

Stroke adalah penyakit saraf yang terjadi secara tiba-tiba disebabkan oleh tersumbat atau terputusnya pembuluh darah pada otak, sehingga mengakibatkan kelumpuhan sebelah tubuh atau bahkan kematian [1]. Stroke termasuk penyakit mematikan dengan angka kematian tinggi, setelah penyakit jantung dan kanker. Pencegahan dini sangat penting untuk mengurangi risiko kematian akibat stroke, namun deteksi awal sering sulit karena faktor risiko yang kompleks dan tidak terstruktur [2]. Di Indonesia, stroke menjadi penyebab kematian tertinggi, mencapai 21,1% dari kematian akibat penyakit tidak menular, terutama dipicu oleh hipertensi dan diabetes. Mengenali tanda awal stroke seperti kelemahan mendadak pada wajah, lengan, atau kaki, kebingungan, dan kesulitan berbicara sangat penting untuk penanganan cepat dan pemulihan optimal [3].

Pada bidang pengobatan stroke, biaya pemeriksaan kesehatan yang cukup mahal seringkali menjadi hambatan bagi masyarakat untuk melakukan deteksi dini risiko stroke. Oleh karena itu, menjaga pola hidup yang sehat dengan fokus pada aktivitas fisik menjadi langkah penting dalam mengurangi risiko stroke sejak awal. Namun, diagnosis medis tetap krusial untuk menilai kondisi kesehatan secara akurat dan memastikan pasien stroke, khususnya tipe 2, mendapatkan perawatan yang tepat [4]. Berbagai penelitian telah menggunakan teknik data mining untuk mengklasifikasikan penyakit stroke dengan tujuan meningkatkan akurasi diagnosis. Teknik ini menggabungkan pengenalan pola, statistik, dan matematika untuk mengekstrak pola serta wawasan baru dari data yang besar. Proses pengumpulan data melalui survei dan observasi mendukung berbagi pengetahuan yang lebih baik, sekaligus menegaskan pentingnya deteksi dini dan intervensi guna menurunkan risiko stroke berat [5].

Berbagai macam penelitian menggunakan algoritma *machine learning* untuk pemodelan klasifikasi penyakit stroke. Penelitian perbandingan ketidakseimbangan data pada *adaboost*, KNN, dan *Random Forest*, hasil klasifikasi gabungan metode ADASYN dan Random Forest menunjukkan hasil yang 5% hingga 10% lebih baik dibandingkan model-model lain [6]. Studi lain pada algortima SVM akurasi mencapai 100%, sedangkan KNN hanya mencapai 97%,

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 532-541
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.612

menghasilkan selisih 3% kurang dari SVM [7]. Pada Penggunaan Algoritma SVM dengan *Relief-f* dan Kernel Linear untuk mencapai hasil yang terbaik sebesar 100%. Namun, Akurasi 100% perlu dicermati karena berpotensi overfitting, penelitian ini difokuskan pada peningkatan *robustness* model melalui *k-fold cross validation* dan penanganan data *imbalance* dengan ADASYN [8]. Meskipun algoritma-algoritma ini menunjukkan hasil yang baik, banyak penelitian belum secara jelas menangani isu ketidakseimbangan data yang sering terjadi dalam dataset medis, termasuk dalam situasi diabetes. Masalah ini muncul ketika jumlah individu yang menderita diabetes jauh lebih sedikit dibandingkan yang tidak terkena, sehingga model menjadi condong kepada kelas mayoritas dan kurang memperhatikan kelas minoritas. Hal ini mengakibatkan kinerja model dalam mengidentifikasi pasien yang berisiko diabetes menjadi sangat rendah, terutama dari segi *recall* atau sensitivitas [9].

Dalam pengelolaan data sering kali terjadi isu *Imbalancing Data*, juga dikenal sebagai ketidakseimbangan kelas. Terjadi ketika sejumlah kelas dalam suatu dataset memiliki data yang tidak merata, dan ketidakseimbangan kelas ini berdampak negatif pada hasil klasifikasi. Metode ADASYN menghasilkan data sintetis dengan rasio oversampling 100%, 50%, dan 25%. Model yang menggunakan data hasil *oversampling* menunjukkan peningkatan kinerja AUC dan recall. Kinerja AUC tertinggi adalah 88,08% dengan data latih 70%, *oversampling* 100%, dan algoritma LGBM, dan kinerja recall tertinggi adalah 83,08% dengan data latih 70%, *oversampling* 100%, dan algoritma LGBM [10].

Kelebihan ADASYN dibanding metode *oversampling* lainnya seperti SMOTE adalah kemampuannya dalam secara adaptif memfokuskan sintesis data baru pada titik-titik data minoritas yang sulit diklasifikasikan. Hal ini memungkinkan ADASYN untuk menghasilkan data sintetis yang lebih relevan dan informatif, sehingga model dapat belajar secara lebih efektif terhadap karakteristik kelas minoritas. Dengan pendekatan ini, ADASYN tidak hanya meningkatkan akurasi secara keseluruhan, tetapi juga memperbaiki *recall* dan sensitivitas pada kasus medis seperti stroke yang sangat memerlukan deteksi pada kelas minoritas[11].

Penelitian ini membahas permasalahan *imbalanced data* pada klasifikasi penyakit stroke. *Imbalanced data* adalah kondisi dimana jumlah data pada kelas tertentu jauh lebih sedikit dibandingkan dengan kelas lain, sehingga dapat mempengaruhi kinerja model klasifikasi. Dalam konteks diagnosis penyakit stroke, ketidakseimbangan data ini sering terjadi karena jumlah pasien yang mengalami stroke biasanya lebih sedikit dibandingkan dengan pasien yang tidak mengalami stroke. Hal ini menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas, sehingga akurasi prediksi pada kelas minoritas (pasien stroke) menjadi kurang optimal [12].

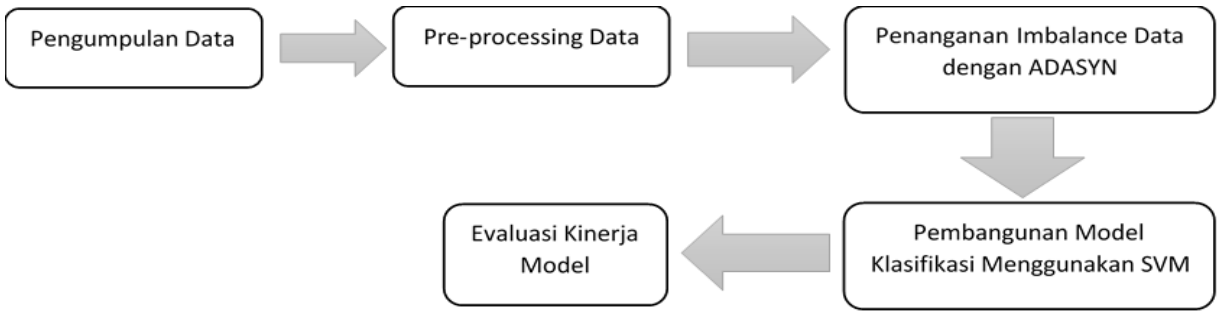
Untuk mengatasi permasalahan tersebut, penelitian ini menggunakan metode *Support Vector Machine* (SVM) sebagai algoritma klasifikasi yang sudah terbukti efektif dalam berbagai kasus klasifikasi. SVM akan diuji dengan menggunakan beberapa kernel yang berbeda untuk mencari performa terbaiknya dalam mengenali pola data stroke. Selain itu, untuk memastikan hasil evaluasi yang lebih valid, penelitian ini menerapkan metode validasi *K-fold cross-validation* yang membagi data menjadi beberapa bagian untuk pelatihan dan pengujian secara bergantian. Kemudian mengimplementasikan teknik penanganan data imbalanced, yaitu ADASYN (*Adaptive Synthetic Sampling*), yang bertujuan untuk menyeimbangkan distribusi data dengan cara menghasilkan data sintetis pada kelas minoritas. Dengan adanya ADASYN, diharapkan model SVM dapat belajar lebih baik terhadap data stroke yang selama ini kurang representatif [13].

Penelitian ini membandingkan hasil klasifikasi SVM dengan beberapa kernel sebelum dan sesudah penerapan ADASYN menggunakan validasi *K-fold*. Perbandingan ini bertujuan untuk mengetahui seberapa besar performa klasifikasi setelah data diseimbangkan menggunakan ADASYN. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam mengetahui performa diagnosis dan deteksi dini penyakit stroke melalui penanganan masalah *imbalanced data*.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Proses pembangunan rancangan model dilakukan dalam deteksi secara dini pada klasifikasi penyakit stroke. Dengan Teknik *data mining* dan penggunaan metode algoritma *machine learning*, hal ini diperlukan dalam menampilkan tahapan yang lebih sistematis dan saling berintegrasi secara terstruktur. Berikut tahapan utama dalam penelitian yang dilakukan pada gambar 1.



Gambar 1. Alur Penelitian

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 532-541
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.612

Sesuai pada Gambar 1, berikut penjelasan rinci dari masing-masing tahapan penelitian yang telah diimplementasikan.

a. Pengumpulan Data
Tahap awal dari penelitian ini dimulai dengan pengumpulan dataset yang relevan dengan proses pembangunan model klasifikasi penyakit stroke. Dataset diperoleh dari sumber website Kaggle dengan nama "healthcare-dataset-stroke-data" dengan format file csv (https://www.kaggle.com/code/saumyagupta2025/stroke-prediction-detailed-eda-7-ml models). Dataset penyakit stroke ini berjumlah 5110 data dengan mempunyai variabel data sebanyak 11 variabel yaitu stroke, *gender*, *age* (umur), *heypertension* (hipertensi), *heart disease* (riwayat jantung), *ever married* (status pernikahan), work *type* (tipe pekerjaan), *residence type* (tipe tempat tinggal), *avg glucose level* (kadar glukosa), BMI (*Body Masss Index*) dan *smoking status* (status merokok).

b. *Pre-processing* Data
Sebelum digunakan dalam pelatihan model, preprocessing dilakukan untuk memastikan kualitas dan konsistensi data. Beberapa atribut, seperti BMI, berisi nilai nol yang tidak dianggap valid secara medis, sehingga nilai-nilai ini diperlakukan sebagai data yang hilang dan diisi dengan median untuk menjaga stabilitas dalam kaitannya dengan outlier. Selanjutnya, transformasi data kategorikal menjadi numerik dilakukan untuk memungkinkan algoritma pembelajaran mesin memproses data dengan lebih efektif. Atribut seperti *Gender* dan *Ever_married* yang memiliki dua kategori dikonversi menggunakan metode label encoding, di mana masing-masing kategori diberikan nilai numerik biner [14]. Sementara itu, atribut dengan lebih dari dua kategori seperti *Work_type, Residence_type, dan Smoking_status* ditransformasikan menggunakan teknik *one-hot encoding* untuk menghindari asumsi ordinal yang salah dan memastikan representasi yang setara antar kategori [15]. Setelah itu, semua atribut numerik dinormalisasi menggunakan *MinMaxScaler* agar memiliki kisaran nilai seragam antara 0 dan 1. Normalisasi ini penting untuk menyeimbangkan skala antar fitur, sehingga mencegah dominasi fitur dengan rentang nilai besar terhadap hasil pembelajaran model. Selain itu, teknik ini juga dapat meningkatkan konvergensi dan akurasi model klasifikasi berbasis jarak atau margin, seperti *Support Vector Machine* (SVM), yang sensitif terhadap perbedaan skala antar fitur. Dalam penelitian ini, pembagian data pelatihan dan pengujian dilakukan menggunakan metode 5-*Fold cross validation*. Setiap lipatan dibentuk dengan teknik stratifikasi (*Stratified K-Fold*) untuk memastikan proporsi kelas stroke dan non-stroke tetap seimbang di seluruh subset pelatihan dan pengujian[16].

c. Penanganan Imbalance Data dengan ADASYN
Dataset mengalami ketidakseimbangan kelas yang signifikan, dimana jumlah sampel pasien tanpa stroke jauh lebih banyak dibandingkan pasien dengan stroke. Untuk mengatasi masalah ini, digunakan metode ADASYN (*Adaptive Synthetic Sampling*) yang secara adaptif menghasilkan sampel sintetis pada kelas minoritas sehingga distribusi kelas menjadi lebih seimbang [17]. Dengan demikian, model yang dibangun dapat belajar secara lebih efektif dan tidak bias terhadap kelas mayoritas.

d. Pembangunan Model Klasifikasi Menggunakan *Support Vector Machine* (SVM)
Pengembangan model klasifikasi menggunakan algoritma Support Vector Machine (SVM) dengan tiga fungsi kernel, yaitu Linear, Polynomial, dan *Radial Basis Function* (RBF). Untuk mendapatkan performa terbaik, pengujian dilakukan dengan metode *K-fold Cross Validation* pada masing-masing kernel. Proses ini memastikan model mampu generalisasi dengan baik dan meminimalisir *overfitting* dengan membagi data latih menjadi beberapa *fold* secara bergantian untuk pelatihan dan validasi [18].

e. Evaluasi Kinerja Model
Evaluasi dilakukan untuk menilai kemampuan model dalam mengidentifikasi pasien yang menderita stroke dan yang tidak. Metrik yang dipakai untuk evaluasi mencakup skor F1, *presisi*, *recall* (sensitivitas), matriks kebingungan, serta skor dari kurva ROC dan AUC. *Precision* dan *recall* dipakai untuk mengevaluasi akurasi dan komprehensivitas dalam mengidentifikasi kasus stroke. Skor F1 mencerminkan keseimbangan antara *presisi* dan *recall*. Kurva ROC digunakan untuk mengevaluasi kinerja model pada berbagai nilai ambang klasifikasi, sedangkan nilai AUC mencerminkan kemampuan model dalam membedakan kelas positif dari negatif. Di samping itu, analisis dilakukan dengan membandingkan dua model *Support Vector Machine* (SVM): model tanpa penanganan ketidakseimbangan data dan model dengan penerapan teknik oversampling ADASYN. Tujuan perbandingan ini adalah untuk menilai pengaruh penerapan ADASYN dalam meningkatkan kinerja model, terutama dalam mendeteksi lebih banyak contoh stroke yang tergolong dalam kelas minoritas.

2.2 Metode ADASYN

ADASYN (*Adaptive Synthetic Sampling*) merupakan metode *oversampling* yang digunakan untuk mengatasi ketidakseimbangan kelas pada dataset klasifikasi dengan cara menghasilkan sampel sintetis pada kelas minoritas. Berbeda dengan metode *oversampling* lain seperti SMOTE yang menghasilkan sampel sintetis secara merata, ADASYN secara adaptif menyesuaikan jumlah sampel sintetis yang dibuat berdasarkan tingkat kesulitan belajar pada masing-masing data minoritas. Data minoritas yang berada di area dengan banyak tetangga mayoritas (artinya sulit dipelajari) akan mendapatkan lebih banyak sampel sintetis, sehingga model dapat fokus pada pola data yang lebih kompleks dan meningkatkan performa klasifikasi [19].

Secara matematis, ADASYN menghitung rasio ketidakseimbangan dan tingkat kesulitan belajar untuk setiap sampel minoritas. Misalkan xi adalah sampel minoritas ke-i, dan k adalah jumlah tetangga terdekat, maka rasio kesulitan belajar ri dihitung sebagai:

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 532-541
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.612

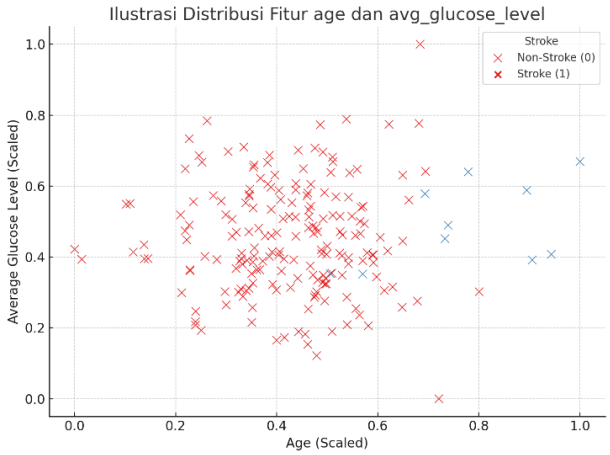
$$r_i = \frac{\Delta_i}{k} \quad (1)$$

Dengan Δi adalah jumlah tetangga dari kelas mayoritas di sekitar xi. Total jumlah sampel sintetis yang akan dihasilkan adalah G, dan proporsi sampel sintetis untuk setiap data minoritas dihitung sebagai:

$$g_i = \left(\frac{r_i}{\sum r_j}\right) \times G, \text{ untuk j = 1 hingga n} \quad (2)$$

Di mana n adalah jumlah data minoritas. Sampel sintetis dibuat dengan interpolasi antara xi dan tetangga minoritas terdekatnya Metode ini memungkinkan ADASYN untuk mengatasi ketidakseimbangan secara efektif dan adaptif. Ini membuatnya ideal untuk dataset klasifikasi penyakit stroke dengan distribusi kelas yang tidak seimbang. Sampel sintetis pada ADASYN dibuat dengan cara interpolasi linier antara data minoritas dan tetangga terdekatnya. Setelah menentukan jumlah sampel yang perlu ditambahkan berdasarkan tingkat kesulitan klasifikasi tiap data minoritas, ADASYN memilih salah satu tetangga minoritas terdekat, lalu menghasilkan data baru di antara keduanya menggunakan rumus x synthetic $= xi+\delta\cdot(xzi-xi)$, dengan δ adalah bilangan acak antara 0 dan 1. Proses ini menghasilkan distribusi data sintetis yang lebih fokus pada area yang sulit dipelajari oleh model.

Untuk memberikan gambaran yang lebih jelas mengenai cara kerja ADASYN, berikut disajikan ilustrasi perhitungan sederhana menggunakan dua fitur dari dataset *Stroke Prediction*, yaitu *age* dan *avg_glucose_level*.



Gambar 2. Tampilan Distribusi untuk 2 fitur

Ilustrasi pada Gambar 2, menggunakan dua fitur, *age* dan *avg_glucose_level*, bertujuan untuk menunjukkan bagaimana ADASYN secara adaptif memilih titik-titik minoritas (pasien stroke) yang berada di area berisiko tinggi karena memiliki banyak tetangga dari kelas mayoritas (non-stroke). Titik-titik ini kemudian dijadikan prioritas dalam proses sintesis data. Sebagai contoh, apabila satu titik data stroke memiliki lima tetangga terdekat dan empat di antaranya merupakan data non-stroke, maka titik ini dianggap sulit dipelajari oleh model. Oleh karena itu, ADASYN akan menghasilkan lebih banyak data sintetis di sekitar titik tersebut dibandingkan titik minoritas yang berada di area yang didominasi tetangga dari kelas yang sama [20].

Proses ini dilakukan dengan cara menginterpolasi linier antara titik data minoritas dan tetangga minoritas lainnya, tetapi dengan jumlah sintetis yang ditentukan oleh skor kesulitan belajar masing-masing. Hasilnya, distribusi kelas menjadi lebih seimbang secara kuantitatif sekaligus lebih representatif secara spasial. Hal ini diharapkan mampu meningkatkan sensitivitas model terhadap kasus stroke yang sebelumnya sulit dikenali. Dengan data pelatihan yang telah diperkuat oleh ADASYN, algoritma *Support Vector Machine* (SVM) kemudian digunakan untuk membangun model prediktif yang andal, dengan pengujian performa dilakukan melalui pendekatan *K-fold Cross Validation* pada masing-masing kernel (Linear, Polynomial, dan RBF) [21].

2.3 Metode *Support Vector Machine*

Support Vector Machine (SVM) adalah algoritma klasifikasi yang digunakan untuk mencari hyperplane terbaik yang dapat memisahkan dua kelas data dengan margin terbesar. Prinsip utama SVM adalah memaksimalkan margin antara dua kelas sehingga dapat meningkatkan kemampuan generalisasi model. SVM efektif dalam menangani data berdimensi tinggi dan mampu bekerja dengan baik pada data linear maupun non-linear melalui penggunaan fungsi kernel [22].
Secara matematis, SVM bertujuan untuk memecahkan permasalahan optimisasi berikut:

$$minimize: \ \frac{1}{2}\ \|w\|^2 \quad (3)$$

$$subject\ to\ y_i\ (w\ .x_i + b) \geq 1\ , \text{ untuk I – 1, 2, …, n}$$

Di mana w adalah vektor bobot, b adalah bias, xi adalah vektor fitur, dan yi adalah label kelas. Untuk kasus data yang tidak dapat dipisahkan secara linear, SVM menggunakan fungsi kernel untuk mentransformasikan data ke ruang

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 532-541
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.612

fitur yang lebih tinggi. Beberapa jenis kernel yang umum digunakan antara lain Linear, Polynomial, dan *Radial Basis Function* (RBF). Dalam penelitian ini, pengujian dilakukan terhadap ketiga kernel tersebut untuk membandingkan performa klasifikasi.

Untuk mengukur kinerja model secara menyeluruh, dilakukan pengujian dengan metode *K-fold Cross Validation*. Metode ini membagi data latih menjadi k subset (*fold*), di mana setiap subset bergiliran menjadi data validasi sementara yang lainnya digunakan untuk pelatihan. Pendekatan ini memberikan estimasi performa yang lebih stabil dan mengurangi risiko *overfitting*.

3. HASIL DAN PEMBAHASAN

3.1 *Data Preparation*

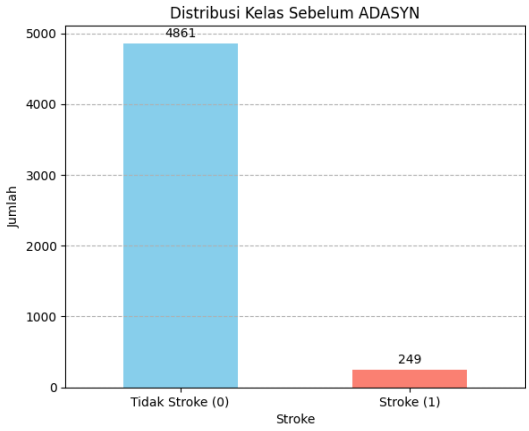
Pengembangan model prediksi penyakit stroke berbasis pendekatan ADASYN dan *Support Vector Machine* (SVM) diawali dengan penyiapan dataset yang digunakan dalam proses pelatihan dan pengujian. Dataset yang digunakan dalam penelitian ini berasal dari platform publik Kaggle dengan nama "Stroke Prediction Dataset" (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset). Dataset ini berisi total 5.110 data pasien dan mencakup 11 fitur, yaitu: jenis kelamin (*gender*), usia (*age*), apakah pasien mengalami hipertensi (*hypertension*), menderita penyakit jantung *(heart_disease*), status menikah (*ever_married*), jenis pekerjaan (*work_type*), tempat tinggal (*Residence_type*), kadar glukosa rata-rata (*avg_glucose_level*), indeks massa tubuh (bmi), status merokok (*smoking_status*), dan label diagnosis stroke (stroke), di mana nilai 1 menunjukkan pasien mengalami stroke dan 0 menunjukkan tidak.

3.2 *Pre-processing* dan Transformasi Data

Sebelum pelatihan model, dilakukan preprocessing untuk memastikan kualitas data. Nilai nol pada atribut BMI yang tidak valid secara medis diperlakukan sebagai data hilang dan diimputasi menggunakan median untuk menjaga kestabilan terhadap *outlier*. Data kategorikal seperti Gender dan Ever_married ditransformasikan dengan label encoding, sementara atribut dengan lebih dari dua kategori seperti *Work_type*, *Residence_type*, dan *Smoking_status* dikonversi menggunakan *one-hot encoding*. Seluruh atribut numerik dinormalisasi menggunakan *MinMaxScaler* agar memiliki skala seragam (0 - 1), yang penting untuk meningkatkan performa algoritma seperti SVM yang sensitif terhadap perbedaan skala. Pembagian data latih dan uji dilakukan menggunakan stratified split guna menjaga proporsi kelas tetap seimbang.

3.3 Penanganan *Imbalanced Data* dengan ADASYN

Setelah preprocessing selesai, langkah penting berikutnya adalah menangani ketidakseimbangan kelas yang signifikan pada dataset. Dataset ini bersifat *imbalanced*, di mana jumlah data pasien non-stroke jauh lebih dominan dibandingkan pasien yang mengalami stroke. Untuk mengatasi ketidakseimbangan ini, metode *oversampling* ADASYN digunakan guna meningkatkan jumlah sampel sintetis pada kelas minoritas (pasien stroke) secara adaptif. Hal ini bertujuan untuk memperbaiki performa model klasifikasi, khususnya dalam mendeteksi kasus stroke yang jumlahnya relatif sedikit dalam data latih. Dataset kemudian dibagi menjadi data latih dan data uji menggunakan teknik *Stratified Split*, sehingga proporsi kelas tetap terjaga di setiap subset data. Gambar 4 berikut tampilan dari dataset stroke yang tidak Seimbang:

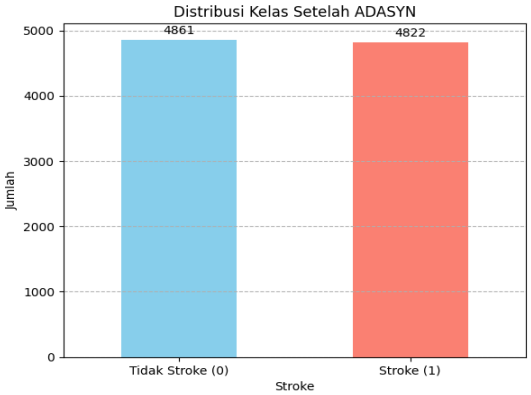


Gambar 3. Tampilan Grafik Distribusi Kelas Tidak Seimbang Sebelum ADASYN

Gambar 3, yang menampilkan distribusi kelas target sebelum dilakukan teknik *oversampling*, di mana kelas 0 (non-stroke) mendominasi sebesar 4861 data dari total 5.110 data, sementara kelas 1 (stroke) hanya sebesar 249 data. Ketidakseimbangan ini dapat mengakibatkan model klasifikasi cenderung bias terhadap kelas mayoritas dan gagal mengenali pola pada kelas minoritas. Untuk mengatasi permasalahan ini, diterapkan teknik ADASYN (*Adaptive Synthetic Sampling*) pada data pelatihan.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 532-541
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.612

Sebelum proses oversampling dilakukan, data terlebih dahulu melalui tahap imputasi nilai kosong dan normalisasi menggunakan *MinMaxScaler* untuk memastikan semua fitur berada pada skala yang sama. Selanjutnya, ADASYN menghasilkan sampel sintetis pada area di mana data minoritas sulit dipelajari, yaitu di sekitar data stroke yang dikelilingi oleh tetangga dari kelas mayoritas. Teknik ini bekerja secara adaptif dengan menghitung tingkat kesulitan belajar dari masing-masing sampel minoritas dan secara proporsional menghasilkan data sintetis yang lebih banyak di wilayah yang kompleks. Dengan pendekatan ini, ADASYN mampu merepresentasikan distribusi lokal dari kelas stroke secara lebih baik dan meminimalkan kemungkinan pembentukan sampel sintetis pada area yang berisiko tinggi terhadap noise.



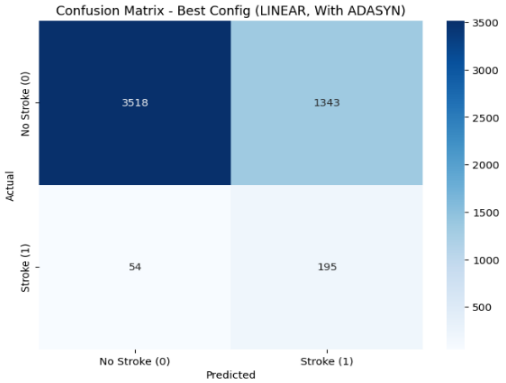
Gambar 4. Tampilan Grafik Distribusi Kelas Seimbang Setelah ADASYN

Gambar 4, menunjukkan distribusi kelas target setelah dilakukan proses *oversampling* menggunakan metode ADASYN. Terlihat bahwa jumlah data dari kelas minoritas, yaitu pasien dengan label stroke (kelas 1), telah meningkat secara signifikan dan menjadi lebih seimbang dengan kelas mayoritas (non-stroke). Sebelumnya, jumlah data stroke hanya mencakup sekitar 249 dari total data, namun setelah proses ADASYN, proporsi kelas stroke meningkat mendekati proporsi kelas non-stroke sebanyak 4822 data.

Distribusi yang lebih seimbang ini menunjukkan bahwa ADASYN berhasil menghasilkan sampel sintetis secara adaptif, khususnya di area di mana data minoritas lebih sulit dipelajari oleh model. Hal ini sangat penting dalam konteks klasifikasi medis, karena memungkinkan model untuk lebih baik dalam mengenali pola-pola dari pasien stroke, yang sebelumnya tersembunyi di bawah dominasi data non-stroke. Dengan perbandingan distribusi sebelum dan sesudah ADASYN yang tergambar jelas pada grafik, dapat disimpulkan bahwa ADASYN tidak hanya meningkatkan jumlah data minoritas, tetapi juga memperbaiki representasi keseluruhan dalam ruang fitur, yang pada akhirnya berkontribusi terhadap peningkatan performa klasifikasi.

3.4 Evaluasi Kinerja Model SVM

Proses selanjutnya adalah membangun model klasifikasi menggunakan algoritma *Support Vector Machine* (SVM) yang dilatih dengan tiga jenis kernel, yaitu Linear, Polynomial, dan *Radial Basis Function* (RBF). Data yang telah diseimbangkan menggunakan ADASYN kemudian dinormalisasi dengan *MinMaxScale*r agar setiap fitur berada dalam skala yang sama. Pelatihan dan evaluasi model dilakukan menggunakan metode 5-*Fold Cross Validation*, di mana pengujian dilakukan secara terpisah untuk masing-masing kernel SVM tersebut. Metode ini membagi data menjadi lima fold secara stratifikasi agar proporsi kelas tetap seimbang di setiap *fold*. Setiap *fold* secara bergantian digunakan sebagai data uji, sementara sisanya sebagai data latih, sehingga performa model dapat diukur dengan lebih stabil dan representatif untuk setiap kernel. Evaluasi dilakukan dengan menggunakan metrik seperti confusion matrix, classification report, dan ROC-AUC score untuk menilai kemampuan model dalam membedakan antara kelas stroke dan non-stroke secara akurat. Hasil evaluasi berupa confusion matrix dan kurva ROC untuk masing-masing kernel SVM disajikan pada bagian berikutnya.



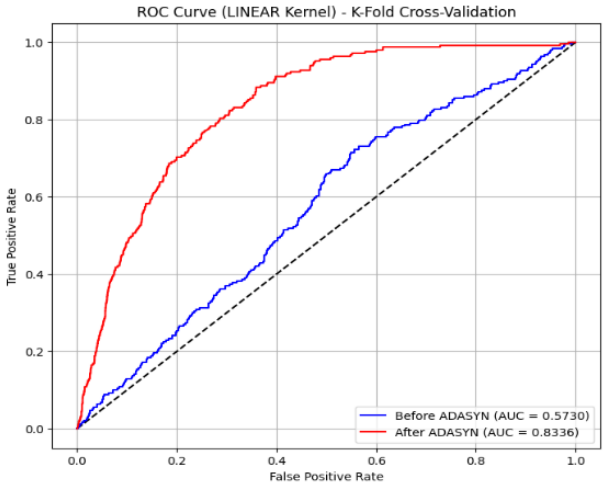
Gambar 5 *Confussion Matrix*

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 532-541
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.612

Gambar 5, menunjukkan hasil confusion matrix dari model *Support Vector Machine* (SVM) dengan kernel Linear yang telah dilakukan penanganan ketidakseimbangan data menggunakan metode Adaptive Synthetic Sampling (ADASYN). Berdasarkan confusion matrix tersebut, diketahui bahwa model berhasil mengklasifikasikan 3.518 data sebagai tidak stroke secara benar (*true negative*) dan 195 data sebagai stroke secara benar (*true positive*). Namun demikian, terdapat 1.343 data yang diklasifikasikan sebagai stroke padahal sebenarnya tidak stroke (*false positive*), serta 54 data yang diklasifikasikan sebagai tidak stroke padahal sebenarnya stroke (*false negative*).

Dari hasil ini, diketahui bahwa model memiliki akurasi sebesar 72,7% terhadap data uji, yang menunjukkan bahwa sebagian besar prediksi model tergolong benar. Nilai *recall* untuk kelas stroke sebesar 78,3%, menandakan bahwa model cukup baik dalam mengenali pasien yang benar-benar mengalami stroke. Namun, precision untuk kelas stroke hanya sebesar 12,7%, yang menunjukkan bahwa sebagian besar data yang diprediksi sebagai stroke ternyata merupakan kesalahan (*false positive*). Tingginya *recall* namun rendahnya precision ini menunjukkan bahwa model cenderung memprioritaskan deteksi kasus stroke sebanyak mungkin, meskipun dengan risiko menghasilkan banyak peringatan palsu.

Strategi ini dapat dikatakan relevan dalam konteks sistem skrining medis, di mana kegagalan mendeteksi pasien yang benar-benar stroke (*false negative*) dapat menimbulkan konsekuensi serius dibandingkan dengan memberikan peringatan palsu (*false positive*). Oleh karena itu, meskipun precision rendah, peningkatan *recall* akibat penerapan ADASYN memperlihatkan bahwa metode ini mampu membantu model dalam mengenali kelas minoritas (stroke) dengan lebih baik, walaupun masih terdapat ruang untuk perbaikan terutama dalam menurunkan tingkat kesalahan prediksi positif palsu.



Gambar 6. *ROC-AUC Curve* dari Model SVM-ADASYN pada Kernel Linear

Gambar 6, merupakan kurva ROC (*Receiver Operating Characteristic*) yang dihasilkan dari proses klasifikasi menggunakan algoritma SVM dengan kernel Linear sebelum dan sesudah penanganan data imbalance menggunakan metode ADASYN (*Adaptive Synthetic Sampling*). Evaluasi dilakukan dengan pendekatan *K-fold cross-validation* untuk mendapatkan hasil yang lebih generalisasi.

Kurva ROC ini menggambarkan perbandingan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai ambang batas (*threshold*). Semakin mendekati pojok kiri atas, maka performa model dikatakan semakin baik. Nilai yang digunakan untuk mengukur performa keseluruhan model adalah AUC (*Area Under the Curve*).

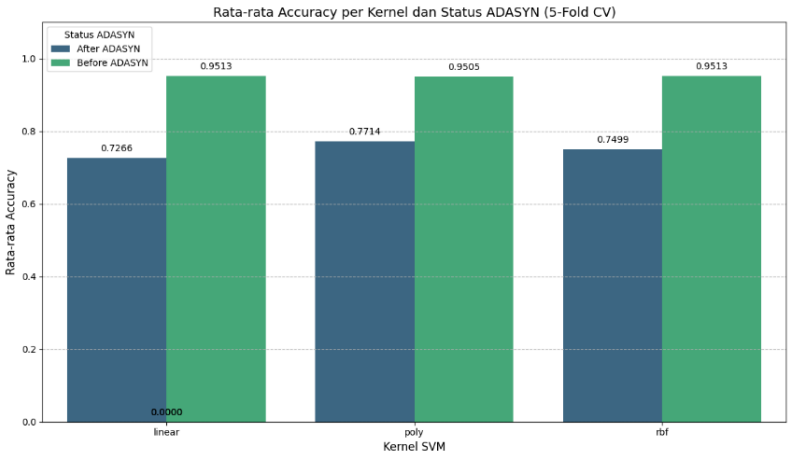
a. Garis berwarna biru menunjukkan performa model sebelum dilakukan penanganan imbalance data. Model menghasilkan AUC sebesar 0.5730, yang menunjukkan bahwa model hampir tidak lebih baik dari tebakan acak (AUC = 0.5).
b. Garis berwarna merah menunjukkan performa model setelah data ditangani menggunakan metode ADASYN. Nilai AUC meningkat signifikan menjadi 0.8336, yang menandakan bahwa model memiliki kemampuan klasifikasi yang jauh lebih baik setelah dilakukan penyeimbangan data.

Nilai *Area Under the Curv*e (AUC) yang diperoleh sebesar 0,8336 menandakan bahwa model memiliki performa yang baik dalam membedakan antara kelas positif (stroke) dan kelas negatif (tidak stroke). AUC dengan nilai di atas 0,8 secara umum dianggap sebagai indikator bahwa model memiliki kapabilitas diskriminatif yang tinggi. Hal ini menunjukkan bahwa probabilitas model untuk mengidentifikasi pasien yang benar-benar stroke lebih tinggi dibandingkan untuk pasien yang tidak stroke, pada sebagian besar ambang batas klasifikasi yang mungkin.

Secara keseluruhan, hasil pada kurva ROC menunjukkan bahwa pendekatan penanganan ketidakseimbangan data menggunakan metode ADASYN berhasil meningkatkan performa model SVM dalam mendeteksi kelas minoritas (stroke) dan memperkuat keandalan model dalam pengambilan keputusan klasifikasi pada studi ini. Peningkatan nilai AUC dari 0.5730 sebelum ADASYN menjadi 0.8336 setelah ADASYN menunjukkan bahwa metode ini efektif dalam mengatasi permasalahan data imbalance. Selain itu, peningkatan metrik recall dari 0.000 menjadi 0.7828 menandakan bahwa model yang sebelumnya sama sekali tidak mampu mengenali kasus stroke kini dapat mengidentifikasi sebagian besar kasus tersebut dengan cukup baik. Hal ini turut didukung oleh peningkatan F1-score dari 0.000 menjadi 0.2181, yang

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 532-541
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.612

menunjukkan adanya perbaikan keseimbangan antara precision dan recall. Meskipun F1-score belum optimal, perubahan ini tetap signifikan karena mengindikasikan bahwa model mulai mampu menangkap pola pada kelas minoritas. Dengan demikian, model klasifikasi menjadi lebih sensitif terhadap kasus stroke, yang sangat krusial dalam konteks diagnosis medis, serta memberikan hasil prediksi yang lebih andal dan informatif.



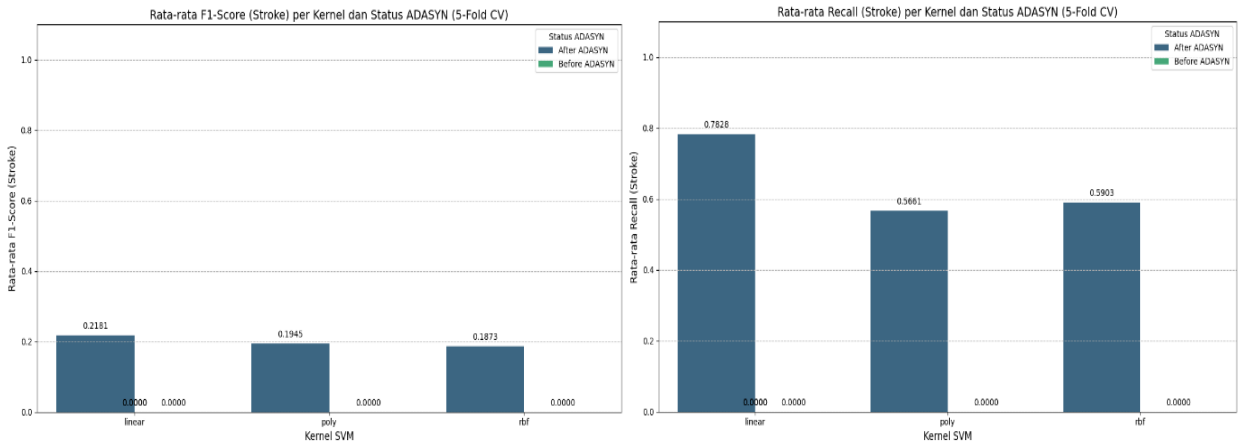
Gambar 7. Perbandingan Rata-rata Akurasi per Kernel

Gambar 7, menunjukkan perbandingan rata-rata akurasi dari masing-masing kernel algoritma *Support Vector Machine* (SVM), yaitu linear, polynomial, dan *Radial Basis Function* (rbf), baik sebelum maupun sesudah penerapan metode ADASYN. Evaluasi dilakukan menggunakan teknik 5-*Fold Cross Validation*.

Sebelum dilakukan penanganan ketidakseimbangan data, seluruh kernel menunjukkan akurasi yang sangat tinggi, yaitu di atas 95%, dengan kernel polynomial mencatat nilai tertinggi sebesar 0.9505, diikuti oleh linear dan rbf yang sama-sama berada di kisaran 0.9513. Namun, akurasi ini perlu dikritisi karena model cenderung bias terhadap kelas mayoritas (non-stroke), mengingat distribusi data yang tidak seimbang.

Setelah ADASYN diterapkan, terjadi penurunan akurasi secara signifikan di semua kernel. Kernel linear mencatat penurunan terbesar hingga 0.7266, sementara polynomial dan rbf masing-masing menurun ke 0.7714 dan 0.7499. Penurunan akurasi ini merupakan trade-off yang umum terjadi dalam penanganan ketidakseimbangan data, di mana model menjadi lebih seimbang dalam mengenali kedua kelas (stroke dan non-stroke), meskipun mengorbankan tingkat akurasi secara keseluruhan.

Penurunan ini justru menandakan keberhasilan ADASYN dalam mendorong model agar tidak hanya fokus pada kelas mayoritas, tetapi juga mampu mendeteksi kasus stroke yang sebelumnya terabaikan. Dengan demikian, meskipun akurasi turun, keseimbangan klasifikasi meningkat, yang sangat penting dalam konteks sistem pendukung keputusan medis.



(a) (b)

Gambar 8 (a) Perbandingan Rata-rata F1-*Score* per Kernel (b) Perbandingan Rata-rata *Recall* per Kernel

Gambar 8, menyajikan analisis perbandingan performa model klasifikasi stroke berdasarkan nilai metrik *recall* dan F1-*score* untuk kelas stroke (positif) pada tiga jenis kernel *Support Vector Machine* (SVM), yakni linear, polynomial, dan *Radial Basis Function* (RBF). Evaluasi dilakukan terhadap dua skenario, yaitu sebelum dan sesudah penerapan metode oversampling ADASYN (*Adaptive Synthetic Sampling*). Metrik *recall* menggambarkan kemampuan model dalam mengenali semua kasus stroke yang benar, sedangkan F1-*score* mempertimbangkan keseimbangan antara precision dan *recall*, sehingga lebih mencerminkan performa keseluruhan dalam konteks data yang tidak seimbang.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 532-541
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.612

Sebelum diterapkannya ADASYN, seluruh jenis kernel menghasilkan nilai *recall* dan F1-*score* sebesar 0.0000, yang menunjukkan bahwa model sama sekali tidak mampu mengenali kasus stroke. Hal ini mengindikasikan bahwa ketidakseimbangan data menyebabkan model sangat bias terhadap kelas mayoritas (non-stroke), sehingga seluruh prediksi diarahkan ke kelas tersebut.

Setelah dilakukan penanganan data tidak seimbang menggunakan ADASYN, peningkatan signifikan terjadi pada kedua metrik. Kernel linear menghasilkan nilai *recall* tertinggi sebesar 0.7828 dan F1-*score* sebesar 0.2181, menunjukkan bahwa model mampu mengenali sebagian besar kasus stroke dengan cukup baik dan memberikan keseimbangan yang wajar antara sensitivitas dan presisi. Kernel polynomial dan rbf juga mengalami peningkatan *recall* menjadi masing-masing 0.5661 dan 0.5903, serta F1-*score* sebesar 0.1945 dan 0.1873, walaupun performanya tidak sebaik kernel linear.

Peningkatan ini menunjukkan bahwa ADASYN sangat efektif dalam memperkuat sensitivitas model terhadap kelas minoritas, yang dalam konteks medis seperti diagnosis stroke sangatlah penting. Meskipun terdapat penurunan pada metrik akurasi keseluruhan, namun hal tersebut merupakan *trade-off* yang dapat diterima mengingat pentingnya deteksi yang tepat pada kondisi medis berisiko tinggi. Untuk mencapai performa yang lebih seimbang dan dapat diterima secara klinis, pendekatan lanjutan seperti *cost-sensitive learning*, perbaikan fitur, atau penggunaan model yang lebih kompleks dapat dipertimbangkan. Dengan demikian, kombinasi antara algoritma SVM dan teknik ADASYN tidak hanya mampu meningkatkan deteksi kasus stroke, tetapi juga memberikan performa klasifikasi yang lebih seimbang, menjadikannya solusi yang layak untuk diterapkan dalam model klasifikasi diagnosis penyakit stroke.

Tabel 1. Perbandingan Performa Model SVM pada Kelas Stroke (Sebelum dan Sesudah ADASYN)

Kernel	ADASYN	Akurasi	Precision (Stroke)	*Recall* (Stroke)	F1-*Score* (Stroke)	AUC-ROC
Linear	Sebelum	0.9513	0.0000	0.0000	0.0000	0.6030
Linear	Sesudah	0.7266	0.1267	0.7828	0.2181	0.8333
Polynomial	Sebelum	0.9505	0.0000	0.0000	0.0000	0.6336
Polynomial	Sesudah	0.7714	0.1174	0.5661	0.1945	0.7351
RBF	Sebelum	0.9513	0.0000	0.0000	0.0000	0.6214
RBF	Sesudah	0.7499	0.1113	0.5903	0.1873	0.7615

Keterangan:

a. Precision, *Recall*, dan F1-*Score* pada tabel ini ditampilkan untuk kelas stroke (label 1), karena fokus utama dari penelitian adalah pada keberhasilan deteksi kasus stroke.
b. AUC-ROC menunjukkan area under the ROC curve yang mengindikasikan keseimbangan kemampuan model dalam mendeteksi kedua kelas.
c. Berdasarkan hasil tersebut, ADASYN memberikan peningkatan signifikan terutama pada nilai *recall* dan F1-*score*, yang sangat penting dalam konteks pendeteksian penyakit stroke.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan sistem klasifikasi untuk diagnosis penyakit stroke menggunakan algoritma *Support Vector Machine* (SVM) yang diuji dengan tiga jenis kernel, yaitu linear, polynomial, dan *Radial Basis Function* (RBF). Evaluasi performa model dilakukan menggunakan metode validasi silang sebanyak lima kali lipat (5-*Fold Cross Validation*) untuk memastikan hasil yang diperoleh bersifat konsisten dan representatif terhadap keseluruhan data. Dalam konteks masalah klasifikasi yang dihadapi data, sampel sintetis *Adaptive Synthetic Sampling* (ADASYN) digunakan untuk meningkatkan representasi kelas minoritas, yaitu pasien yang menderita stroke. Hasil eksperimen menunjukkan bahwa ADASYN mampu meningkatkan sensitivitas model terhadap kelas minoritas, yang tercermin dari meningkatnya nilai *recall* dan F1-*score*, khususnya untuk kelas stroke. Meskipun akurasi keseluruhan menurun setelah ADASYN diterapkan, hal ini merupakan *trade-off* yang dapat diterima dalam konteks medis, di mana kemampuan mendeteksi kasus positif lebih diutamakan. Namun demikian, F1-score yang dihasilkan (0.2181) masih tergolong rendah, mengindikasikan precision yang lemah akibat tingginya jumlah false positive, sehingga performa model belum sepenuhnya seimbang untuk aplikasi klinis. Oleh karena itu, pendekatan lanjutan seperti *cost-sensitive learning*, pengayaan fitur, atau penggunaan model yang lebih kompleks disarankan untuk meningkatkan keseimbangan performa klasifikasi. Kernel linear memberikan hasil terbaik dengan AUC-ROC sebesar 0.8333, recall sebesar 0.7827, dan F1-score sebesar 0.2181. Implementasi sistem dilakukan pada platform Google Colab yang menyediakan berbagai pustaka pemrograman Python dan kemudahan dalam pemrosesan data secara *cloud-based*, sehingga mempercepat proses eksperimen dan visualisasi hasil. Secara keseluruhan, kombinasi algoritma SVM dan penanganan ketidakseimbangan data dengan ADASYN terbukti dapat meningkatkan deteksi kasus stroke secara signifikan dan layak digunakan sebagai bagian dari model klasifikasi medis dalam diagnosis penyakit stroke. Untuk penelitian selanjutnya, disarankan untuk membandingkan metode ADASYN dengan teknik penyeimbangan data lainnya seperti SMOTE, Borderline-SMOTE, atau kombinasi *undersampling* dan *oversampling*. Selain itu, pengujian terhadap dataset yang lebih besar atau berbasis data klinis riil dapat memberikan gambaran yang lebih akurat terhadap performa sistem dalam praktik dunia nyata. Integrasi sistem ini

ke dalam aplikasi berbasis web atau mobile sebagai alat bantu diagnosis juga merupakan arah pengembangan yang potensial dan bermanfaat dalam konteks pelayanan kesehatan.

REFERENCES

[1] D. E. Cahyani, "Penerapan Machine Learning Untuk Prediksi Penyakit Stroke," *J. Kaji. Mat. dan Apl.*, 2022, doi: 10.17977/um055v3i12022p15-22.
[2] Y. Azhar, A. K. Firdausy, and P. J. Amelia, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," *SINTECH (Science Inf. Technol. J.*, 2022, doi: 10.31598/sintechjournal.v5i2.1222.
[3] E. Firmawati, E. Rochmawati, and I. Setyopranoto, "Deteksi Risiko Stroke Dan Edukasi Sebagai Upaya Pencegahan Primer Terjadinya Stroke," *J. SOLMA*, 2023, doi: 10.22236/solma.v12i2.11834.
[4] A. M. Ramadhan, J. S. Zahra, K. Al Rasyid, and D. O. W. Nugroho, "Aplikasi Forecasting Risiko Terkena Penyakit Stroke Menggunakan Program R-Shiny," *J. Sains dan Seni ITS*, 2022, doi: 10.12962/j23373520.v11i3.62543.
[5] Ardi Ramdani, Christian Dwi Sofyan, Fauzi Ramdani, Muhamad Fauzi Arya Tama, and Muhammad Angga Rachmatsyah, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Masyarakat Dalam Menerima Bantuan Sosial," *J. Ilm. Sist. Inf.*, 2022, doi: 10.51903/juisi.v1i2.363.
[6] K. Fithriasari, I. Hariastuti, and K. S. Wening, "Handling Imbalance Data in Classification Model with Nominal Predictors," *Int. J. Comput. Sci. Appl. Math.*, 2020, doi: 10.12962/j24775401.v6i1.6643.
[7] Rahel Lina Simanjuntak, Rizki Agung Ramadhan, Theresia Romauli Siagian, and Vina Anggriani, "Komparasi Algoritma KNN dan SVM dalam Memprediksi Penyakit Stroke," *J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 3, no. 3, pp. 60–74, 2023, doi: 10.55606/teknik.v3i3.2474.
[8] U. Amelia, J. Indra, and A. F. N. Masruriyah, "Implementasi Algoritma Support Vector Machine (Svm) Untuk Prediksi Penyakit Stroke Dengan Atribut Berpengaruh," *Sci. Student J. Information, Technol. Sci.*, vol. III, no. 2, pp. 254–259, 2022.
[9] M. Khushi *et al.*, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3102399.
[10] I. W. Dharmana, I. G. A. Gunadi, and L. J. E. Dewi, "Deteksi Transaksi *Fraud* Kartu Kredit Menggunankan *Oversampling* ADASYN dan Seleksi Fitur SVM-RFECV," *J. Teknol. Inf. dan Ilmu Komput.*, 2024, doi: 10.25126/jtiik.20241117640.
[11] R. M. Munshi, "Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction," *PLoS One*, 2024, doi: 10.1371/journal.pone.0296107.
[12] I. Pratama, A. Y. Chandra, and P. T. Presetyaningrum, "Seleksi Fitur dan Penanganan Imbalanced Data menggunakan RFECV dan ADASYN," *J. Eksplora Inform.*, 2022, doi: 10.30864/eksplora.v11i1.578.
[13] A. A. Rahman, S. S. Prasetiyowati, and Y. Sibaroni, "Performance Analysis Of The Imbalanced Data Method On Increasing The Classification Accuracy Of The Machine Learning Hybrid Method," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, 2023, doi: 10.29100/jipi.v8i1.3286.
[14] B. J. Jansen, K. K. Aldous, J. Salminen, H. Almerekhi, and S. gyo Jung, "Data Preprocessing," in *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2024. doi: 10.1007/978-3-031-41933-1_6.
[15] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technol. J. Ilm.*, 2024, doi: 10.31602/tji.v15i1.13457.
[16] V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, P. R. da Silva Pereira, and J. L. Victória Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowl. Inf. Syst.*, 2023, doi: 10.1007/s10115-022-01772-8.
[17] R. Mia *et al.*, "Exploring Machine Learning for Predicting Cerebral Stroke: A Study in Discovery," *Electron.*, 2024, doi: 10.3390/electronics13040686.
[18] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput. Oper. Res.*, 2023, doi: 10.1016/j.cor.2022.106131.
[19] F. O. Awalullaili, D. Ispriyanti, and T. Widiharih, "Klasifikasi Penyakit Hipertensi Menggunakan Metode Svm Grid Search Dan Svm Genetic Algorithm (Ga)," *J. Gaussian*, 2023, doi: 10.14710/j.gauss.11.4.488-498.
[20] Y. A. Sir and A. H. H. Soepranoto, "Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas," *J. Komput. dan Inform.*, 2022, doi: 10.35508/jicon.v10i1.6554.
[21] G. Abdurrahman, "Klasifikasi Kanker Payudara Menggunakan Algoritma SVM dengan Kernel RBF, Linier, dan Sigmoid," *JUSTIFY J. Sist. Inf. Ibrahimy*, 2023, doi: 10.35316/justify.v2i1.3370.
[22] D. Nurlaily, Y. P. Irfandi, N. Santoso, S. Qomariyah, and D. Wibowo, "Classification of Hepatitis Patients Using Logistic Regression and Support Vector Machines Methods," *J. Pendidik. Mat.*, 2022, doi: 10.21043/jpmk.v5i2.17052.