ZONAsi Jurnal Sistem Informasi
ZONAsi VOL. 7 NO. 1 JANUARI 2025
Page: 61 - 73
ISSN: 2656-7407 (Online)
2656-7393 (Print)

PENERAPAN TEKNIK SMOTE PADA KLASIFIKASI PENYAKIT STROKE DENGAN ALGORITMA SUPPORT VECTOR MACHINE

Lugas Pasiolo[1], Iis Afrianty [*2] , Elvia Budianita [3], Rahmad Abdillah [4]
[1,2,3,4]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau
[1,2,3,4] Jl.HR. Soebrantas RW 15, Simpang Baru, Pekanbaru, Riau, telp. (0761) 56222
e-mail: [1]12050113246@students.uin-suska.ac.id, [*2]iis.afrianty@uin-suska.ac.id, [3]elvia.budianita@uin-suska.ac.id, [4]rahmad.abdillah@uin-suska.ac.id

Abstrak

Stroke adalah kondisi darurat medis yang dapat menyebabkan kerusakan otak atau kematian. Deteksi dini dan klasifikasi risiko stroke sangat penting untuk pencegahan dan penanganannya. Penelitian ini menggunakan dataset sebanyak 5110 data untuk meningkatkan akurasi klasifikasi stroke dengan algoritma Support Vector Machine (SVM) pada data tidak seimbang. Teknik Synthetic Minority Over-sampling Technique (SMOTE) diterapkan untuk menyeimbangkan data stroke dan non-stroke, yang dapat meningkatkan performa model. SVM diuji dengan berbagai kernel, yaitu Linear, RBF, Polynomial, dan Sigmoid, serta variasi parameter pada masing-masing kernel untuk mencari konfigurasi optimal. Hasil pengujian menunjukkan penerapan SMOTE meningkatkan akurasi, presisi, dan recall, dengan kernel RBF mencapai akurasi tertinggi 92% pada parameter Cost 100 dan Gamma 1. Temuan ini menunjukkan bahwa penggunaan SMOTE dan optimasi parameter SVM dapat menghasilkan model klasifikasi yang lebih efektif dalam mendeteksi risiko stroke pada data tidak seimbang.

Kata kunci: *Stroke, Support Vector Machine, SMOTE, Klasifikasi, Ketidakseimbangan Data.*

Abstract

Stroke is a medical emergency that can lead to significant brain damage or even death. Early detection and risk classification of stroke are crucial for prevention and management. This study uses a dataset of 5110 samples to improve the accuracy of stroke classification using the Support Vector Machine (SVM) algorithm on imbalanced data. The Synthetic Minority Over-sampling Technique (SMOTE) is applied to balance the stroke and non-stroke data, enhancing model performance. SVM is tested with various kernels, including Linear, RBF, Polynomial, and Sigmoid, along with parameter variations for each kernel to find the optimal configuration. The results show that applying SMOTE significantly improves accuracy, precision, and recall, with the RBF kernel achieving the highest accuracy of 92% at Cost 100 and Gamma 1 parameters. These findings suggest that using SMOTE and optimizing SVM parameters can produce a more effective classification model for detecting stroke risk in imbalanced datasets.

Keywords: *Stroke, Support Vector Machine, SMOTE, Classification, Imbalanced Data.*

1. PENDAHULUAN

Stroke adalah kondisi darurat medis yang setara dengan serangan jantung [1]. Penyakit ini disebabkan oleh gangguan sirkulasi darah di otak yang terjadi secara cepat, progresif, dan mendadak [2]. Kondisi ini sering kali menyebabkan kerusakan otak yang signifikan atau bahkan kematian. Faktor utama yang menyebabkan kematian akibat stroke antara lain perdarahan hebat (34%), infark serebral hebat (25%), dan pneumonia aspirasi (22%) [3]. Dengan tingginya risiko yang ditimbulkan, penting untuk meningkatkan kesadaran masyarakat akan pentingnya pencegahan dan pengobatan stroke, guna menurunkan risiko penyakit tersebut.

Pada tahun 2022, *World Stroke Organization* melaporkan bahwa stroke menjadi tantangan kesehatan global dengan lebih dari 101 juta orang di seluruh dunia hidup dengan riwayat stroke. Setiap tahun, terdapat lebih dari 7,6 juta kasus stroke iskemik baru, di mana 62% terjadi di seluruh dunia, dan lebih dari 12,2 juta orang atau sekitar satu dari empat orang berusia di atas 25 tahun berisiko mengalami stroke. Di negara-negara berpenghasilan rendah dan menengah ke bawah, terjadi peningkatan kasus stroke sebesar 70%, kematian sebesar 43%, dan morbiditas sebesar 143% antara tahun 1990 hingga 2019 [4]. Fakta ini menegaskan perlunya upaya preventif dan intervensi yang tepat dalam menangani penyakit stroke.

Di Indonesia, hasil Riset Kesehatan Dasar (Riskesdas) yang dilakukan antara tahun 2007 hingga 2018 menunjukkan peningkatan prevalensi stroke. Data menunjukkan bahwa prevalensi stroke meningkat dari 7% pada tahun 2013 menjadi 10,9% pada tahun 2018 (Badan Penelitian dan Pengembangan Kesehatan, 2021). Hal ini menandakan pentingnya pencegahan dan penanganan yang lebih intensif untuk menurunkan angka kematian dan kecacatan akibat stroke. Namun, biaya pemeriksaan dan perawatan medis untuk stroke yang cukup mahal menjadi penghalang bagi sebagian masyarakat untuk melakukan pemeriksaan rutin (Sandy et al., 2022). Oleh karena itu, pendekatan yang lebih terjangkau dan dapat diakses masyarakat luas sangat dibutuhkan.

Kemajuan teknologi membuka peluang untuk meningkatkan deteksi dini stroke melalui teknologi kecerdasan buatan (AI). Teknologi ini memungkinkan deteksi risiko stroke sejak dini, sehingga peluang pemulihan dan penanganan segera dapat ditingkatkan [5]. Salah satu aplikasi AI yang mulai banyak digunakan adalah sistem berbasis machine learning, yang memungkinkan sistem komputer untuk membantu pasien memahami kondisi mereka [6]. Beberapa metode *machine learning* yang sering digunakan dalam diagnosis dan klasifikasi penyakit meliputi *Decision Tree*, *Backpropagation Neural Network*, *Learning Vector Quantization*, dan *Support Vector Machine* (SVM) yang memiliki potensi untuk memprediksi risiko stroke secara lebih akurat.

Penelitian sebelumnya menunjukkan hasil yang beragam dalam penggunaan berbagai algoritma untuk klasifikasi penyakit stroke. Penelitian oleh [7]menggunakan algoritma *Decision Tree* C.45 dengan data dari Kaggle yang terdiri dari 11 atribut, menghasilkan akurasi klasifikasi sebesar 96,05%. Penelitian lainnya oleh [8] menggunakan metode *Backpropagation Neural Network* dengan 4981 data stroke dan menghasilkan rata-rata akurasi sebesar 96,14%. Sementara itu, (Melani et al., 2023) menggunakan metode *Learning Vector Quantization* pada dataset yang sama, menghasilkan akurasi tertinggi sebesar 70%, dengan presisi 0,72, recall 0,70, dan F1-score 0,69. Meskipun hasilnya cukup baik, akurasi metode ini masih perlu ditingkatkan untuk memenuhi kebutuhan klinis.

Penelitian yang menggunaka metode *Support Vector Machine* (SVM) yang dilakukan oleh penelitian lain dengan metode *Support Vector Machine* (SVM) pada penyakit cacar monyet oleh [10] mencapai akurasi sebesar 65%. Penelitian yang membandingkan SVM dan Naïve Bayes oleh [11] pada dataset penyakit diabetes menunjukkan bahwa *Support Vector Machine* (SVM) dengan kernel polynomial memiliki akurasi 96,27%, dibandingkan dengan 92,07% pada Naïve Bayes. Kesimpulannya, metode *Support Vector Machine* (SVM) cukup baik dalam klasifikasi, namun performa model dapat bervariasi tergantung pada dataset dan pengaturan parameter yang digunakan.

Pada penelitian yang dilakukan oleh [12] pada penyakit stroke dengan menggunakan metode Support Vector Machine (SVM) dengan kernel linear, polynomial, RBF dan sigmoid pada dataset yang terdiri dari 5110 data dengan 12 atribut. Hasil penelitian menunjukkan bahwa kernel polynomial menghasilkan akurasi tertinggi sebesar 78,86%, dengan presisi 73,98% dan recall 56,75% pada rasio pembagian data 80:20. Namun, akurasi ini masih tergolong rendah, kemungkinan disebabkan oleh masalah ketidakseimbangan data (imbalanced data), yang mengurangi kemampuan model dalam mendeteksi kasus stroke secara akurat.

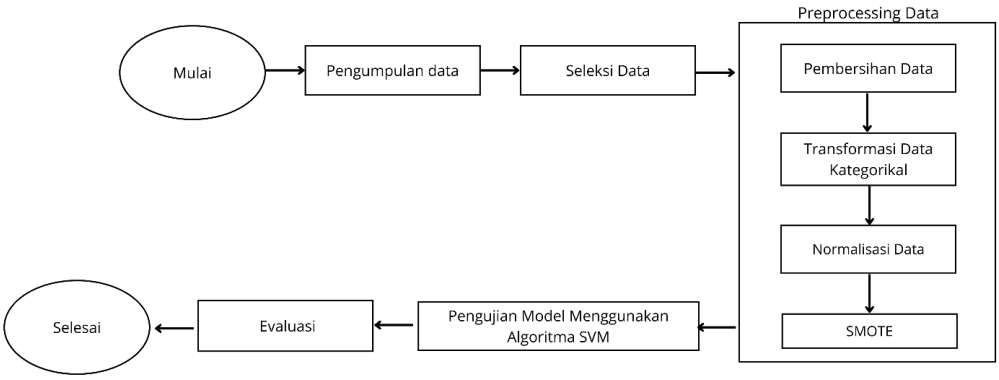
Sebagai solusi terhadap kendala seperti ketidakseimbangan data (imbalanced data), penelitian ini akan menggunakan dataset yang sama dan metode *Support Vector Machine* (SVM), namun dengan penambahan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk menyeimbangkan distribusi data stroke dan non-stroke. Transformasi data serta penyesuaian parameter juga akan dilakukan guna meningkatkan performa model. Dengan demikian, penelitian ini diharapkan dapat menghasilkan model klasifikasi stroke yang lebih cepat, efisien, dan akurat, khususnya dalam menangani masalah ketidakseimbangan data yang sering kali mengurangi kemampuan model dalam mendeteksi kasus stroke dengan akurat.

2. METODE PENELITIAN

Metode penelitian ini menggunakan pendekatan eksperimen dengan tahapan utama berupa pengumpulan data stroke, preprocessing data untuk menangani ketidakseimbangan kelas menggunakan SMOTE, serta penerapan algoritma SVM untuk klasifikasi. Evaluasi model dilakukan dengan mengukur akurasi, presisi, recall, dan F1-score untuk menilai kinerja sistem dalam mendeteksi penyakit stroke.

2.1. Tahapan Penelitian

Tahapan penelitian pada Gambar 1 menunjukkan alur proses pengolahan data dan pengujian model menggunakan algoritma *Support Vector Machine* (SVM). Dimulai dari pengumpulan data, seleksi data, preprocessing (pembersihan, transformasi, normalisasi, dan SMOTE), hingga pengujian dan evaluasi model sebelum proses selesai.



Gambar 1. Tahapan Penelitian

2.2. Data Penelitian

Data penelitian yang di ambil dari platform kaggle. Data dapat diakses melalui https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. Dataset penyakit stroke ini berjumlah 5110 data dengan mempunyai variabel data sebanyak 12 variabel yaitu id, 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). Pada Tabel 1 dijelaskan dataset awal dari penyakit stroke.

Tabel 1. Data Penelitian

id	Gender	Age	Hypertension	Heart_diease	Ever_married	Work_type	Residence_type	Avg_glucose_level	BMI	Smoking_status	Stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
…	…	…	…	…	…	…	…	…	…	…	…
19723	Female	35	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0

37544	Male	51	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
44679	Female	44	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

2.3. Seleksi Data

Pada tahap ini, dilakukan pemilihan data yang relevan dari dataset mentah. Data yang tidak relevan atau tidak dibutuhkan akan dibuang. Tujuan utama dari tahap ini adalah memastikan hanya data yang berguna yang dipertahankan untuk analisis lebih lanjut. Data yang awalnya memiliki 12 variabel diseleksi sehingga hanya digunakan 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). Terdapat 1 variabel target atau label yaitu stroke dan 10 atribut.

2.4. Pembersihan Data

Pembersihan data bertujuan untuk menghapus data yang ambigu sehingga hanya atribut-atribut yang relevan dan berguna untuk penelitian yang tersisa seperti data yang bernilai NaN [13]. Pada kolom BMI, beberapa nilai hilang atau NaN diganti dengan rata-rata (mean) kolom tersebut. Hal ini dilakukan untuk menghindari bias yang dapat ditimbulkan oleh data yang hilang serta untuk memastikan bahwa semua fitur siap digunakan dalam proses pelatihan model

2.5. Transformasi Data Kategorikal

Pada tahapan transformasi data kategorikal dilakukan proses Encoding, yaitu proses *Label Encoding* dan *One-Hot Encoding*. Label encoding merupakan teknik dalam machine learning yang digunakan untuk mengonversi data kategori menjadi data numerik agar dapat digunakan oleh algoritma pemodelan [13]. *One hot encoding* merupakan proses mengonversi variabel menjadi format yang dapat digunakan oleh algoritma machine learning untuk meningkatkan kinerja klasifikasi, hasil encoding ini berupa representasi biner, yaitu 1 dan 0 [14]. Proses Encoding dilakukan pada fitur kategorikal seperti *Gender*, *Ever_married*, Work_type, *Residence_type*, dan *Smoking_status*. Karena nilai kategori ini tidak kompatibel dengan algoritma *Support Vector Machine* (SVM), diperlukan encoding untuk mengonversinya menjadi format numerik atau biner agar dapat diproses oleh model *Support Vector Machine* (SVM). Pada Tabel 3 menunjukkan data setelah *Label Encoding* dan *One-Hot Encoding*. Fitur *Gender*, *Work_type*, dan *Smoking_status* dilakukan proses *One-Hot Encodiing*, seperti *gender_male* dan *work_type_private*, yang menunjukkan kategori dengan nilai 1 atau 0. Sedangkan fitur *Ever_maried* dan *Residence_type* telah dilakukan Label Encoding.

Tabel 2. Transformasi Data Kategorikal

gender_male	**gender_female**	**gender_other**	**Ever_married**	**work_type_private**	**work_type_Self-employed**	**work_type_Self-employed**	**work_type_Self-employed**	**Residence_type**	**smoking_status_never**	**smoking_status_fomerly smoked**	**smokes**
1	0	0	1	1	0	0	0	1	0	1	0
0	1	0	1	0	1	0	0	0	1	0	0
1	0	0	1	1	0	0	1	0	1	0	0

2.6. Normalisasi Data

Tujuan normalisasi data dalam dataset adalah untuk menyelaraskan nilai-nilai data agar berada dalam rentang yang sama, sehingga memudahkan analisis dan meningkatkan kinerja algoritma machine learning [15]. Teknik yang digunakan adalah *Min-Max Scaling* yaitu teknik normalisasi yang digunakan untuk mengubah nilai numerik menjadi skala yang lebih seragam, biasanya dalam rentang [0, 1] [16]. Proses normalisasi pada fitur numerik seperti *age*, *avg glucose level*, dan BMI dinormalisasi menggunakan *Min-Max Scaling* untuk rentang nilai antara 0 dan 1, sehingga semua fitur memiliki kontribusi yang proporsional dalam model *Support Vector Machine* (SVM). berikut merupakan persamaan dari *Min-Max Scalling.*

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \textbf{(1)}$$

Keterangam

X = merupakan nilai asli dari data yang akan dinormalisasi.

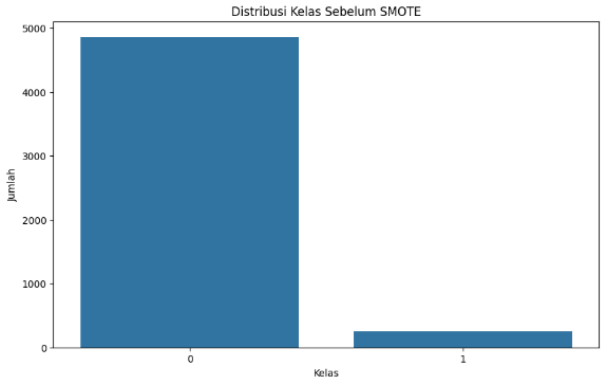
X_{min} = merupakan nilai terkecil (minimum) dari data dalam fitur tersebut.

X_{max} = merupakan nilai terbesar (maksimum) dari data dalam fitur tersebut.

X_{norm} = merupakan nilai setelah normalisasi, yang berada dalam rentang 0 hingga 1 (jika data berada dalam rentang normal).

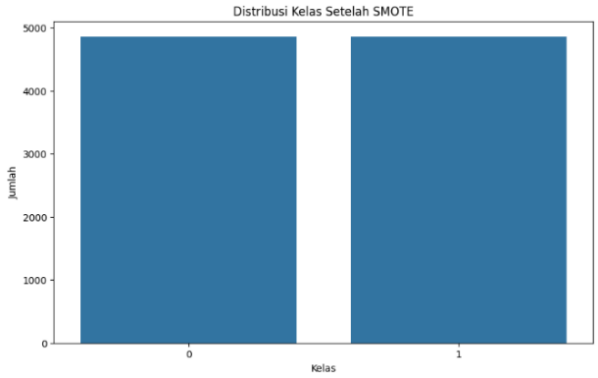
2.7. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) adalah metode *oversampling* di mana data pada kelas minoritas ditingkatkan dengan menggunakan data sintetik yang dihasilkan dari replikasi data kelas minoritas [17]. Dalam penelitian ini, kasus stroke merupakan kelas minoritas [18]. Pada Gambar 2 adalah data sebelum SMOTE. Diagram ini menunjukkan distribusi data antara kelas 0 (tidak stroke) dan 1 (stroke) sebelum diterapkan SMOTE, dengan jumlah yang tidak seimbang: 4.861 data pada kelas 0 dan 249 pada kelas 1.



Gambar 2. Data Sebelum dilakukan SMOTE

Gambar 3 merupakan data setelah SMOTE. Setelah menerapkan SMOTE, data menjadi seimbang dengan total 9.722 data, yaitu masing-masing 4.861 untuk kelas 0 dan kelas 1, sehingga model dapat mengenali pola pada kedua kelas dengan lebih baik.



Gambar 3. Data Setelah dilakukan SMOTE

2.7. Pengujian Model Menggunakan Algoritma Support Vector Machine (SVM)

Pemodelan dilakukan menggunakan algoritma *Support Vector Machine* (SVM) dengan empat jenis kernel, yaitu RBF , Linear, Polynomial, dan Sigmoid. Tahap pengujian pertama bertujuan untuk membandingkan hasil penelitian sebelumnya oleh [12], di mana hanya diuji akurasi tertinggi dari setiap kernel pada rasio data latih dan uji tertentu. Hasil penelitian tersebut menunjukkan akurasi terbaik pada kernel Linear dan Polynomial dengan rasio 80:20, kernel RBF dengan rasio 90:10, dan kernel Sigmoid dengan rasio 10:90.

Untuk meningkatkan hasil pengujian dan mengevaluasi performa model secara lebih mendalam, dilakukan pengujian tambahan menggunakan parameter tertentu pada kernel RBF , Linear, Polynomial, dan Sigmoid. Pertama, model diuji dengan kernel Linear menggunakan persamaan yang telah dijelaskan sebelumnya, dengan variasi nilai C = 1, 2, dan 3 untuk mengamati pengaruh regularisasi terhadap margin dan kesalahan klasifikasi. Kedua, model diuji dengan kernel RBF, menggunakan variasi nilai C = 1, 2, dan 3, serta gamma (γ) = 1, 2, dan 3 untuk menentukan sejauh mana pengaruh dari setiap titik data terhadap klasifikasi. Terakhir, model diuji menggunakan kernel Polynomial, dengan variasi nilai C = 1, 2, dan 3, degree = 1, 2, dan 3, serta gamma (γ) = 1, untuk menyesuaikan pengaruh bias dalam proses klasifikasi.

Untuk meningkatkan validitas dan keakuratan hasil pengujian model, dilakukan evaluasi lebih lanjut menggunakan teknik cross-validation pada masing-masing kernel SVM yang diuji. Cross-validation ini bertujuan untuk memberikan gambaran yang lebih baik tentang kinerja model dengan meminimalkan kemungkinan overfitting dan memastikan bahwa model dapat menggeneralisasi dengan baik pada data yang belum terlihat.

2.8. Evaluasi

Evaluasi dilakukan untuk mengukur validitas dan manfaat pola yang ditemukan menggunakan Confusion Matrix, yang membandingkan hasil prediksi model dengan nilai sebenarnya. Matriks ini terdiri dari True Positive (TP), True Negative (TN), False Positive (FP) (kesalahan Tipe I), dan False Negative (FN) (kesalahan Tipe II). Dari matriks ini, metrik seperti akurasi, presisi, recall, dan F1-score dihitung untuk menilai kualitas model, memberikan gambaran tentang kemampuan model dalam mengklasifikasikan data secara akurat dan mengidentifikasi kelemahannya.

Tabel 3. Confusion Matrix

Kelas Benar	Stroke	Tidak Stroke
Stroke	True Positif	False Positif
Tidak Stroke	False Negatif	True Negatif

Akurasi adalah persentase dari prediksi yang benar terhadap keseluruhan data:

$$\boldsymbol{Akurasi} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \qquad \textbf{(2)}$$

Presisi mengukur proporsi prediksi positif yang benar-benar benar:

$$\boldsymbol{Presisi = \frac{TP}{TP + FP}} \quad \textbf{(3)}$$

Recall mengukur kemampuan model untuk menemukan semua contoh positif dalam data:

$$\boldsymbol{Recall = \frac{TP}{TP + FN}} \quad \textbf{(5)}$$

F1-Score adalah rata-rata harmonis antara presisi dan recall:

$$\boldsymbol{F1 = 2 . \frac{Presisi\ .\ Recall}{Presisi + Recall}} \quad \textbf{(6)}$$

3. HASIL DAN PEMBAHASAN

Pada hasil dan pembahasan, akan dibahas mengenai hasil eksperimen yang dilakukan untuk mengukur efektivitas penerapan teknik SMOTE dalam meningkatkan performa klasifikasi penyakit stroke menggunakan algoritma SVM. Pembahasan mencakup analisis perbandingan antara model dengan dan tanpa penggunaan SMOTE, serta evaluasi terhadap kinerja model berdasarkan metrik-metrik klasifikasi seperti akurasi, presisi, recall, dan F1-score.

3.1. Hasil Pemodelan

Tabel 4 menunjukkan hasil pengujian tanpa SMOTE yang dilakukan oleh [12] pada berbagai kernel (Linear, Polynomial, RBF, dan Sigmoid) menunjukkan bahwa kernel Polynomial memiliki akurasi tertinggi sebesar 78.86% pada rasio data latih dan uji 80:20, sementara kernel Sigmoid hanya mencapai akurasi 47.14%.

Tabel 4. Hasil Pengujian Penelitian Sebelumnya Tanpa SMOTE

Pengujian	Kernel	Akurasi	Presisi	Recall
80:20	Linear	75.73	74.07	56.16
	Polynomial	78.86	73.98	56.75
90:10	RBF	73.38	71.79	55.85
10:90	Sigmoid	47.14	39.84	48.13

Tabel 5, setelah menerapkan SMOTE dengan 3 kali pengujian , kemudian hasil rata-ratanya dihitung dan ditampilkan, hasil pengujian meningkat, dengan kernel RBF mencapai akurasi 83,85%, presisi 79,21%, dan recall 91,28% pada rasio 90:10. Peningkatan ini menunjukkan bahwa SMOTE secara signifikan meningkatkan performa klasifikasi pada semua kernel. Selanjutnya akan dilakukan pengujian dengan menggunakan parameter pada kernel linear, RBF dan polynomial.

Tabel 5. Hasil Rata-Rata Pengujian Penelitian Menggunakan SMOTE

Pengujian	Kernel	Akurasi	Presisi	Recall
80:20	Linear	79,75	76,03	88,62
	Polynomial	83,17	79,83	87,79
90:10	RBF	83,85	79,21	91,28
10:90	Sigmoid	58,62	58,10	59,37

Tabel 6 menunjukkan hasil pengujian kernel Linear dengan variasi parameter Cost dengan 3 kali pengujian , kemudian hasil rata-ratanya dihitung dan ditampilkan, menunjukkan stabilitas performa,

dengan akurasi konsisten pada 79%,presisi 84-85%, recall 88% dan F1-Score 81%, menunjukkan bahwa kernel Linear kurang sensitif terhadap perubahan Cost.

Tabel 6. Hasil Rata-Rata Pengujian Kernel Linear Menggunakan Parameter

Pengujian	Cost	Akurasi	Presisi	Recall	F1-Score
80:20	0.1	79,25	74,67	88,39	81,31
	1	79,35	75,08	87,7	81,90
	10	79,33	75,12	87,56	80,86
	100	79,33	75,12	87,56	80,86
90:10	0.1	78,83	74,99	88,37	81,14
	1	79,20	75,54	88,16	81,36
	10	79,24	75,62	88,09	81,38
	100	79,24	75,62	88,09	81,38

Tabel 7 pada kernel RBF, pengujian kernel RBF dengan variasi Cost dan Gamma dengan 3 kali pengujian , kemudian hasil rata-ratanya dihitung dan ditampilkan, menunjukkan hasil terbaik pada pengujian 90:10 Cost 100 dan Gamma 1, dengan akurasi 92,12%, presisi 88,65%, recall 97,14% dan F1-Score 89,26%. Variasi Gamma berdampak signifikan terhadap performa, terutama pada akurasi dan recall.

Tabel 7. Hasil Rata-Rata Pengujian Kernel RBF Menggunakan Parameter

Pengujin	Cost	Gamma	Akurasi	Presisi	Recall	F1-Score
80:20	0.1	0.01	71,65	64,86	94,22	72,77
		0.1	79,09	73,58	90,61	81,21
		1	81,61	76,85	90,35	83,05
		scale	80,46	75,01	91,20	82,32
	1	0.01	79,15	73,88	90,04	81,17
		0.1	81,87	77,43	89,83	83,17
		1	85,89	81,46	92,85	86,78
		scale	84,04	79,53	91,58	85,13
	10	0.01	79,98	75,88	87,77	81,39
		0.1	84,17	79,88	91,23	85,18
		1	88,79	85,03	90,76	89,33
		scale	86,10	81,64	93,06	86,98
	100	0.01	81,78	77,64	89,14	83
		0.1	85,55	81,05	92,72	86,49
		1	91,42	87,91	95,98	91,77
		scale	88,57	84,37	94,60	89,2
90:10	0.1	0.01	72,18	66,01	94,81	77,82
		0.1	78,66	73,5	90,69	81,40
		1	81,57	77,18	91,15	83,58
		scale	80,23	75,36	91,55	82,67
	1	0.01	78,35	73,76	89,95	81,05
		0.1	82,02	78,37	89,89	83,73
		1	86,06	82,24	93,01	87,29
		scale	84,75	80,87	92,22	86,17
	10	0.01	79,35	76,12	88,22	81,73
		0.1	84,76	81,23	91,55	86,08
		1	89,62	86,28	94,94	90,4

		scale	86,84	83,13	93,41	87,97
	100	0.01	81,47	78,03	89,09	83,19
		0.1	86,43	82,59	93,35	87,63
		1	92,12	88,65	97,14	89,36
		scale	89,62	86,10	95,21	90,42

Tabel 8 yaitu pengujian kernel Polynomial diuji dengan variasi Cost, Degree, dan Coef0 dengan 3 kali pengujian , kemudian hasil rata-ratanya dihitung dan ditampilkan. Kombinasi terbaik diperoleh pada pengujian 90:10 Cost 100, Degree 3, dan Coef0 1, dengan akurasi tertinggi 88,56%, presisi 84,7%, recall 94,94%, F1-Score 84,58. Secara umum, nilai Degree dan Cost yang lebih tinggi cenderung meningkatkan performa.

Tabel 8. Hasil Rata- RataPengujian Kernel Polynomial Menggunakan Parameter

Pengujian	**cost**	**degree**	**coef0**	**Akurasi**	**Presisi**	**Recall**	**F1-Score**
80:20	0,1	1	0	79,11	74,14	89,24	80,99
		2		80,14	76,22	87,46	81,45
		3		81,95	78,44	88,01	82,95
		1	1	79,13	74,14	89,28	81,01
		2		80,38	76,45	87,66	81,67
		3		82,5	78,63	89,14	83,56
	1	1	0	79,52	75,24	87,83	81,05
		2		81,71	78,12	87,97	82,76
		3		84,32	80,52	90,45	85,19
		1	1	79,2	75,24	87,83	81,05
		2		81,49	77,58	88,47	82,66
		3		84,57	80,63	91,55	85,55
	10	1	0	79,55	75,38	87,59	81,04
		2		82,74	78,72	89,62	83,82
		3		85,71	81,55	92,2	86,55
		1	1	79,54	75,37	87,55	81,02
		2		82,67	78,39	90,11	83,84
		3		85,81	81,32	92,89	86,72
	100	1	0	79,54	75,39	87,56	81,02
		2		83,15	78,52	91,17	84,37
		3		87,16	83,04	93,33	87,88
		1	1	79,52	75,37	87,56	81,00
		2		83,12	78,33	91,44	84,38
		3		87,54	83,38	93,71	91,57
90:10	0,1	1	0	78,69	74,3	89,62	81,24
		2		80,09	76,87	87,76	81,95
		3		82,29	79,59	88,22	83,68
		1	1	78,69	74,3	89,62	81,24
		2		80,12	76,97	87,82	82,02
		3		82,6	79,66	88,89	84,02
	1	1	0	79,07	75,37	88,16	81,26
		2		81,57	78,67	88,09	83,11
		3		84,79	81,76	90,69	85,99
		1	1	79,07	75,37	88,16	81,26
		2		81,19	77,93	88,55	82,90
		3		84,1	81,38	91,88	86,31
	10	1	0	79,17	75,56	88,02	81,31
		2		82,22	78,67	89,82	83,72

	3		86,67	83,16	92,95	87,78
10	1		79,14	75,54	87,96	81,28
	2	1	82,08	78,29	90,22	83,83
	3		86,74	82,75	93,74	87,9
	1		79,1	75,53	87,89	81,24
	2	0	81,64	77,68	90,29	83,51
	3		86,06	82,10	93,35	87,36
100	1		79,1	75,53	87,89	81,24
	2	1	82,84	78,69	91,42	84,58
	3		88,56	84,7	94,94	89,52

Tabel 9 menunjukkan hasil rata-rata cross-validation untuk tiga jenis kernel yang digunakan dalam model, yaitu kernel Linear, RBF, dan Polynomial, dengan parameter yang ditentukan. Kernel Linear dengan parameter C=100 menghasilkan akurasi sebesar 79,84%, presisi 76,1%, recall 87,05%, dan F1-Score 81,2%. Kernel RBF, dengan parameter C=100 dan Gamma=1, menunjukkan performa yang lebih baik, dengan akurasi mencapai 89,23%, presisi 85,36%, recall 94,72%, dan F1-Score 89,79%. Sementara itu, kernel Polynomial dengan parameter C=100, Degree=3, dan Coef0=1 juga memberikan hasil yang signifikan, dengan akurasi 86,68%, presisi 82,49%, recall 93,14%, dan F1-Score 87,47%. Dari hasil tersebut, kernel RBF menunjukkan kinerja terbaik dalam hal akurasi, presisi, recall, dan F1-Score, diikuti oleh kernel Polynomial, sementara kernel Linear menghasilkan performa yang lebih rendah dibandingkan kedua kernel lainnya.

Tabel 9. Hasil Rata-Rata Cross-Validation

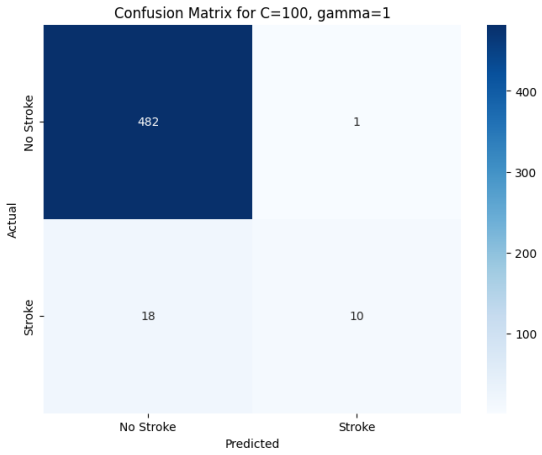
Kernel	**Akurasi**	**Presisi**	**Recall**	**F1-Score**
Linear	79,84	76,1	87,05	81,2
RBF	89,23	85,36	94,72	89,79
Polynomial	86,68	82,49	93,14	87,47

Secara keseluruhan, hasil pengujian yang disajikan dalam tabel-tabel tersebut menunjukkan bahwa penerapan SMOTE secara signifikan meningkatkan kinerja model pada semua jenis kernel yang diuji. Tanpa SMOTE, kernel Polynomial menunjukkan akurasi tertinggi sebesar 78,86% pada rasio data latih dan uji 80:20 , sementara kernel Sigmoid memiliki akurasi yang jauh lebih rendah, yaitu 47,14%. Setelah menggunakan SMOTE, terdapat peningkatan kinerja yang signifikan, dengan kernel RBF mencapai akurasi 92,12% pada rasio data 90:10 dan parameter Cost 100 serta Gamma 1. Penerapan SMOTE juga meningkatkan akurasi kernel Polynomial menjadi 83%, menunjukkan bahwa teknik ini dapat mengatasi masalah ketidakseimbangan data dengan efektif. Hasil dari pengujian dengan berbagai parameter, seperti variasi Cost pada kernel Linear dan RBF, serta parameter Degree dan Coef0 pada kernel Polynomial, menunjukkan bahwa pengaturan parameter yang tepat dapat lebih meningkatkan kinerja model. Secara keseluruhan, kernel RBF menunjukkan kinerja terbaik dalam hal akurasi dan recall, diikuti oleh kernel Polynomial, sementara kernel Linear memiliki performa yang lebih stabil namun lebih rendah dibandingkan keduanya. Dengan demikian, penerapan SMOTE dan pemilihan parameter yang tepat sangat penting untuk meningkatkan akurasi dan deteksi risiko stroke yang lebih baik.

3.2. Evaluasi

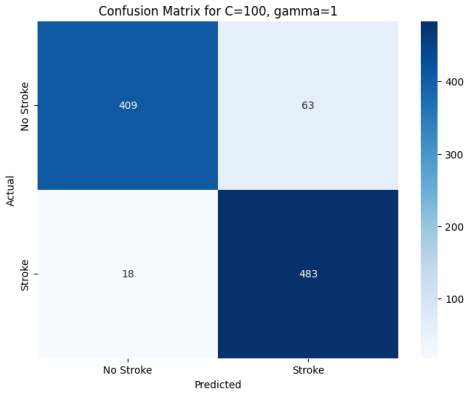
Pengujian kernel RBF ditunjukkan pada Gambar 4 Confusion Matrix pada kernel RBF menggunakan Cost 100 dan Gamma 1 tanpa menggunakan SMOTE. Model berhasil mengklasifikasikan 482 data yang sebenarnya No Stroke dengan benar sebagai No Stroke (True Negative), dan hanya ada 1 data yang salah diklasifikasikan sebagai Stroke (False Positive). Di sisi lain, model berhasil mengklasifikasikan 10 data yang sebenarnya Stroke dengan benar sebagai Stroke (True Positive), namun terdapat 18 data Stroke yang salah diklasifikasikan sebagai No Stroke (False Negative). Hasil ini menunjukkan bahwa meskipun model mampu mengidentifikasi sebagian besar data

yang tidak mengalami stroke dengan baik, namun model memiliki kelemahan dalam mendeteksi beberapa kasus stroke, yang mengarah pada false negatives yang cukup signifikan.



Gambar 4. Confusion Matrix Tanpa SMOTE

Pengujian kernel RBF ditunjukkan pada Gambar 5 Confusion Matrix menggunakan SMOTE dengan variasi Cost dan Gamma menunjukkan hasil terbaik pada Cost 100 dan Gamma 1, dengan akurasi 92,12%, presisi 88,65%, recall 97,14% dan F1-Score 89,26%. Matriks menunjukkan bahwa model berhasil mengklasifikasikan 409 data yang sebenarnya No Stroke dengan benar sebagai No Stroke (True Negative), namun ada 63 data yang salah diklasifikasikan sebagai Stroke (False Positive). Model juga mengklasifikasikan 483 data yang sebenarnya Stroke dengan benar sebagai Stroke (True Positive), tetapi terdapat 18 data Stroke yang salah diklasifikasikan sebagai No Stroke (False Negative). Matriks ini menggambarkan kinerja model dalam mendeteksi kasus stroke, dengan performa yang cukup baik dalam mengidentifikasi kasus Stroke namun masih terdapat kesalahan dalam mendeteksi No Stroke.



Gambar 5. Confusion Matrix Menggunakan SMOTE

Pengujian kernel RBF tanpa SMOTE pada Gambar 4 menunjukkan model dengan akurasi tinggi, tetapi recall yang rendah (35,71%), mengindikasikan bahwa meskipun model berhasil mengklasifikasikan sebagian besar data No Stroke dengan benar, model kesulitan dalam mendeteksi banyak kasus Stroke (18 false negatives). Sebaliknya, pengujian dengan SMOTE pada Gambar 5 meningkatkan recall menjadi 97,14%, yang berarti model lebih baik dalam mendeteksi kasus Stroke, meskipun ini mengorbankan presisi yang sedikit menurun dan menyebabkan false positives meningkat (63 data No Stroke diklasifikasikan sebagai Stroke). Meskipun ada lebih banyak kesalahan pada false positives, penggunaan SMOTE berhasil membuat model lebih sensitif dalam mendeteksi kasus positif, yang lebih diutamakan dalam aplikasi medis, di mana mendeteksi sebanyak mungkin kasus Stroke lebih penting daripada menghindari false positives.

3.3. Pembahasan

Dalam pembahasan ini, dapat didentifikasi beberapa gap penelitian yang membedakan penelitian ini dengan penelitian sebelumnya. Penelitian [12] hanya menggunakan berbagai kernel (Linear, Polynomial, RBF, dan Sigmoid) tanpa ada nya parameter dan tidak menerapkan SMOTE, sehingga akurasi maksimal yang dicapai adalah 78,86% pada rasio data 80:20 di kernel Polynomial. Selain itu, pengujian pada berbagai parameter menunjukkan bahwa kombinasi Cost dan Gamma pada kernel RBF memberikan hasil yang optimal dibandingkan pendekatan sebelumnya yang cenderung menggunakan parameter default. Dalam studi terdahulu, efek dari parameter Degree dan Coef0 pada kernel Polynomial kurang diperhatikan, yang dalam penelitian ini terbukti berkontribusi terhadap performa model.

4. KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa penggunaan algoritma *Support Vector Machine* (SVM) dengan penerapan teknik Synthetic Minority Over-sampling Technique (SMOTE) secara signifikan meningkatkan performa klasifikasi dalam mendeteksi risiko stroke, terutama pada data yang tidak seimbang. Pengujian tanpa SMOTE menunjukkan performa terbaik pada kernel Polynomial dengan akurasi tertinggi sebesar 78,86%, sementara setelah SMOTE diterapkan, akurasi model meningkat secara signifikan, terutama pada kernel RBF yang mencapai akurasi optimal 92,12%, presisi 88,65%, recall 97,14% dan F1-Score 89,26%dengan parameter Cost 100 dan Gamma 1. Hasil ini menunjukkan bahwa penggunaan SMOTE berhasil menyeimbangkan data kelas minoritas dan mayoritas, sehingga meningkatkan sensitivitas model terhadap kasus stroke yang sebelumnya cenderung terabaikan. Secara keseluruhan, kernel Polynomial dan RBF menunjukkan hasil yang konsisten dan akurat dalam mengklasifikasikan data stroke, menegaskan bahwa kombinasi SMOTE dengan *Support Vector Machine* (SVM) dan pengaturan parameter yang tepat mampu menghasilkan model prediktif yang lebih andal dalam mendeteksi risiko stroke. Penelitian selanjutnya disarankan dapat mengeksplorasi kombinasi algoritma lain dengan SMOTE untuk meningkatkan deteksi risiko stroke, serta mencoba teknik sampling lain seperti ADASYN.

Daftar Pustaka

[1] M. T. N. Rosmary and F. Handayani, "Hubungan Pengetahuan Keluarga dan Perilaku Keluarga pada Penanganan Awal Kejadian Stroke," *Journal of Holistic Nursing and Health Science*, vol. 3, no. 1, pp. 32–39, 2020.

[2] B. P. Tomasouw and F. Y. Rumlawang, "Penerapan Metode SVM Untuk Deteksi Dini Penyakit Stroke (Studi Kasus : RSUD Dr. H. Ishak Umarella Maluku Tengah dan RS Sumber Hidup-GPM)," *Tensor: Pure and Applied Mathematics Journal*, vol. 4, no. 1, pp. 37–44, Jun. 2023, doi: 10.30598/tensorvol4iss1pp37-44.

[3] M. I. Tew, L. C. Goo, M. S. Said, H. I. Zahari, and N. A. Ali, "Oral health related quality of life in stroke survivors at community-based rehabilitation centre: A pilot study," *Makara Journal of Health Research*, vol. 24, no. 1, pp. 21–26, 2020, doi: 10.7454/msk.v24i1.1181.

[4] V. L. Feigin *et al.*, "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," Jan. 01, 2022, *SAGE Publications Inc.* doi: 10.1177/17474930211065917.

[5] V. Adelina, D. E. Ratnawati, and M. A. Fauzi, "Klasifikasi Tingkat Risiko Penyakit Stroke Menggunakan Metode GA-Fuzzy Tsukamoto," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 9, pp. 3015–3021, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[6] A. W. Mucholladin, F. Abdurrachman Bachtiar, and M. T. Furqon, "Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 2, pp. 622–633, 2021, [Online]. Available: http://j-ptiik.ub.ac.id

[7] R. Estian Pambudi, Sriyanto, and Firmansyah, "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C.45," *Jurnal Teknika*, vol. 16, no. 02, pp. 221–226, 2022.

[8] M. Azhima, I. Afrianty, E. Budianita, and S. Kurnia Gusti, "Penerapan Metode Backpropagation Neural Network untuk Klasifikasi Penyakit Stroke," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 6, pp. 3013–3021, 2024, doi: 10.30865/klik.v4i6.1956.

[9] P. Melani Almahmuda Batubara, I. Afrianty, S. Sanjaya, and F. Syafria, "Klasifikasi Penyakit Stroke Jaringan Syaraf Tiruan Menerapkan Metode Learning Vector Quantization," *Jurnal Informatika Universitas Pamulang*, vol. 8, no. 2, pp. 223–228, 2023.
[10] W. Anugrah, E. Haerani, Yusra, and L. Oktavia, "Klasifikasi Penyakit Cacar Monyet Menggunakan Metode Support Vector Machine," *Journal of Computer System and Informatics (JoSYC)*, vol. 5, no. 3, pp. 558–566, 2024, doi: 10.47065/josyc.v5i3.5149.
[11] H. Apriyani and kurniati, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," *Journal of Information Technology Ampera*, vol. 1, no. 3, pp. 2774–2121, 2020, [Online]. Available: https://journal-computing.org/index.php/journal-ita/index
[12] S. Rahayu and Y. Yamasari, "Klasifikasi Penyakit Stroke dengan Metode Support Vector Machine (SVM)," *Journal of Informatics and Computer Science*, vol. 05, no. 03, pp. 440–446, 2024.
[13] S. R. Bagaskara and D. H. Bangkalang, "Analisis dan Implementasi Market Basket Analysis (MBA) Menggunakan Algoritma Apriori dengan Dukungan Visualisasi Data," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 4, p. 612, Jul. 2023, doi: 10.30865/json.v4i4.6351.
[14] S. L. Harris and D. Harris, "Architecture," *Digital Design and Computer Architecture*, pp. 298–390, Jan. 2022, doi: 10.1016/B978-0-12-820064-3.00006-4.
[15] A. Harmain, H. Kurniawan, Kusrini, and D. Maulina, "Normalisasi Data Untuk Efisiensi K-Means Pada Pengelompokan Wila-yah Berpotensi Kebakaran Hutan Dan Lahan Berdasarkan Sebaran Titik Panas," *TEKNIMEDIA*, vol. 2, no. 2, pp. 83–89, 2021.
[16] A. Toha, P. Purwono, and W. Gata, "Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 4, no. 1, pp. 12–21, May 2022, doi: 10.12928/biste.v4i1.6079.
[17] E. Sutoyo and M. Asri Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *Jurnal Edukasi dan Penelitian Informatika*, vol. 6, no. 3, pp. 379–385, 2020.
[18] I. S. Ramadhan and A. Salam, "Teknik Random Undersampling untuk Mengatasi Ketidakseimbangan Kelas pada CT Scan Kista Ginjal," *Techno.COM*, vol. 23, no. 1, pp. 20–28, 2024.