Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 3, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i3.8424

Optimasi *Random Forest* untuk Deteksi Dini Penyakit Stroke dengan Data Rekam Medis

Theo Krisna Amarya [1], Aidina Ristyawan [2], Rina Firliana [3]
[1-3] Universitas Nusantara PGRI Kediri, Jl. Ahmad Dahlan No.76, Mojoroto, Kota Kediri, 64112 Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-02-06
Revised 2025-08-21
Accepted 2025-08-26

Corresponding Author:

Aidina Ristyawan

Email: aidinaristi@unpkediri.ac.id

Abstract *– A stroke is a medical condition that occurs when blood flow to the brain is blocked, causing damage to brain tissue. Stroke is the second leading cause of death and disability worldwide, affecting people of all ages and influenced by various risk factors, such as unhealthy lifestyles, high blood pressure, high blood sugar levels, and other risks. It is crucial to detect strokes in patients as soon as possible to prevent them. This study proposes optimizing the performance of the Random Forest algorithm as an early detection model for stroke by utilizing a hybrid sampling method called SMOTE-Tomek and conducting several experiments on the parameter settings of the Random Forest algorithm. The results of this study show an improvement compared to previous research, which had an accuracy of 94% with a standard deviation of 2%. In this study, an accuracy of 96% was achieved with a standard deviation of 0% and an ROC curve (AUC) value of 0.96 or 96%. The algorithm with 96% accuracy in the discussion is the Random Forest algorithm as an estimator of AdaBoost.*

Keywords: *Adaboost; Hybrid Sampling; Random Forest; Smote-Tomek; Stroke Predict.*

Abstrak *– Stroke adalah kondisi medis yang terjadi ketika aliran darah ke otak terhambat, sehingga menyebabkan kerusakan pada jaringan otak. Stroke merupakan penyebab kematian dan kecacatan terbesar kedua di dunia, penyakit ini dapat menyerang segala usia dan dipengaruhi oleh berbagai aspek risiko, seperti gaya hidup tidak sehat, tekanan darah tinggi, kadar gula darah yang tinggi, dan risiko lainnya. Sangat penting untuk mendeteksi stroke pada pasien sesegera mungkin untuk mencegahnya. Penelitian ini mengusulkan optimalisasi kinerja algoritma Random Forest sebagai model deteksi dini penyakit stroke dengan memanfaatkan metode hybrid sampling yang disebut SMOTETomek dan juga melakukan beberapa percobaan terhadap pengaturan parameter algoritma Random Forest. Hasil dari penelitian ini menunjukkan peningkatan dibandingkan dengan penelitian sebelumnya yang memiliki akurasi sebesar 94% dengan standar deviasi sebesar 2%, Pada penelitian ini berhasil mencapai akurasi sebesar 96% dengan standar deviasi sebesar 0% dengan nilai kurva ROC (AUC) sebesar 0.96 atau 96%. Algoritma yang memiliki akurasi 96% pada pembahasan tersebut adalah Algoritma Random Forest sebagai estimator dari AdaBoost.*

Kata Kunci*: Adaboost, Hybrid Sampling, Random Forest, Smote-Tomek, Stroke Predict*

I. PENDAHULUAN

Stroke adalah kondisi medis yang terjadi ketika aliran darah ke otak terhambat, sehingga mengakibatkan kerusakan jaringan otak [1]. Stroke merupakan salah satu penyebab utama kematian dan kecacatan nomor dua di dunia. Penyakit ini dapat menyerang semua kelompok usia, mulai dari anak-anak hingga lansia. Faktor risiko utama yang mempengaruhi terjadinya stroke meliputi pola hidup tidak sehat, seperti kurang olahraga dan konsumsi makanan tidak bergizi, tekanan darah tinggi, serta kadar gula darah yang tidak terkontrol. Selain itu, faktor seperti merokok, obesitas, dan riwayat keluarga juga berkontribusi terhadap peningkatan risiko stroke.

Sangat penting untuk mendeteksi stroke pada pasien sesegera mungkin untuk mencegahnya. Teknologi *machine learning* [2] telah banyak digunakan di berbagai bidang, termasuk bidang kesehatan. Klasifikasi adalah teknik yang sering digunakan untuk diagnosis dan prognosis penyakit. Klasifikasi menggunakan sejumlah algoritma untuk memprediksi risiko stroke berdasarkan data pasien. Dengan menemukan algoritma terbaik yang dapat memberikan hasil yang paling akurat dan konsisten, mengingat banyaknya algoritma yang tersedia dan dataset yang besar, prediksi penyakit stroke juga dapat dilakukan menggunkan data *electronic health record* yang pernah dilakukan oleh [3] dengan menerapkan beberapa algortima klasifikasi.

Algoritma seperti *Random Forest*, *K-Nearest Neighbors* (KNN), *Naive Bayes*, *Decision Trees*, *Support Vector Machines* (SVM), *Neural Networks*, dan *Logistic Regretion* sering digunakan untuk klasifikasi [4][5]. Namun, kinerja setiap algoritma dapat berbeda tergantung pada karakteristik dataset dan teknik preprocessing yang digunakan, seperti penyeimbangan data dan validasi silang. Distribusi data yang tidak merata, seperti ketidakseimbangan data pada dataset, dapat menyebabkan terjadinya bias atau penyimpangan sistematis pada model, yang dapat mempengaruhi hasil klasifikasi dan mengurangi akurasi algoritma [6]. Untuk membuat model klasifikasi menjadi lebih akurat dalam memprediksi kasus stroke, penerapan metode penyeimbangan data dapat

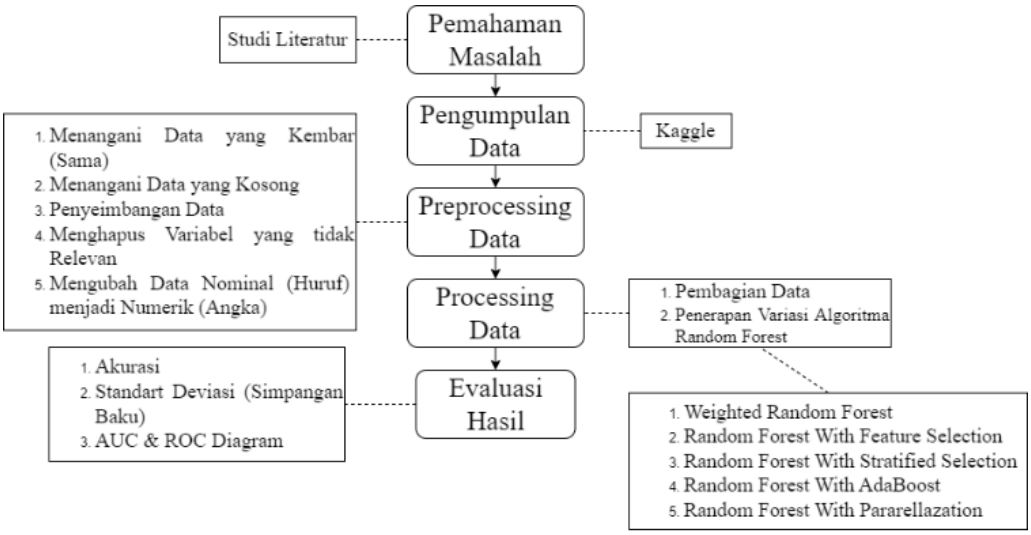
digunakan untuk membantu meningkatkan kinerja algoritma klasifikasi. Selain itu, teknik validasi silang juga digunakan untuk menilai kinerja model.

Penelitian sebelumnya melakukan perbandingan Algoritma Klasifikasi berdasarkan komposisi label, menghasilkan beberapa kinerja algortima klasifikasi seperti algotirma KNN (*K-Neares Neighbour*) yang menghasilkan kinerja sebesar 89%, algortima *Naive Bayes* menghasilkan kinerja sebesar 79%, algoritma *Decission Tree* sebesar 91%, algortima SVM (*Support Vector Machine*)sebesar 77%, algoritma *Neural Network* sebesar 83%, algoritma *Logistic Regression* sebesar 81%, dan algoritma *Random Forest* dengan tingkat akurasi 94% dengan standar deviasi 2% [7]. Pada penelitian lain yang menggunakan dataset sama seperti yang dilakukan oleh[8] membahas prediksi penyakit stroke menggunakan algortima *Logistik Regression* menghasilkan kinerja sebesar 86%,dan penelitian lain yang pernah dilakukan oleh[9] membahas klasifikasi penyakir stroke menggunakan algoritma SVM (*Support Vector Machine*) menghasilkan kinerja sebesar 85.45% dengan perbandingan data 80:20 dan sebesar 85.24% dengan perbandingan data 70:30.

Penelitian[10] mengaplikasikan algoritma *Random Forest* sebagai algoritma dalam memprediksi stroke, kemudian penelitian [11] membuktikan bahwa algoritma *Random Forest* adalah algoritma yang cocok dalam memprediksi penyakit stroke dengan hasil akurasi sebesar 94%, namun masih memiliki kekurangan pada bagian *standart deviasi*. Hal ini yang membuat penelitian ini berfokus dalam mengoptimalkan algoritma *Random Forest* agar menghasilkan standar deviasi seminimal mungkin dan mengembangkan model prediksi berdasarkan algoritma *Random Forest* sebagai sarana deteksi dini penyakit stroke dan untuk mengetahui potensi penyakit stroke secara dini, dengan menggunakan data rekam medis.

II. METODE

Studi ini mengenakan prosedur eksperimen, prosedur eksperimen menggambarkan pendekatan studi kuantitatif yang kerap digunakan guna menguji hipotesis kausal serta menguasai hubungan sebab akibat antar variabel. Tahapan yang digunakan bisa dilihat pada Gambar 1 di bawah ini.



Gambar 1. Alur Penelitian

Pada gambar 1, dapat dijelaskan sebagai berikut:

A. Pemahaman Masalah

Langkah pertama yang harus dilakukan adalah memahami bahwa stroke merupakan penyebab kematian terbesar di Indonesia dan bahkan nomor dua di dunia[12]. Untuk membantu dalam pengklasifikasian penyakit ini, diperlukan analisis yang akurat dengan menggunakan algoritma klasifikasi yang tepat. Dari hasil penelitian sebelumnya, algoritma *Random Forest* akan menjadi fokus utama dalam penelitian ini karena kekuatannya dalam menangani data yang kompleks.

B. Pengumpulan Data

Dataset yang digunakan pada penelitian ini bersumber dari website Kaggle [3], [13]–[15] nama dataset yang digunakan adalah "*stroke_dataset.csv*" [16]. Dengan dataset yang mempunyai 12 kolom yang terdiri dari 11 kolom variabel independen serta 1 kolom variabel dependen. Variabel independen pada dataset tersebut merupakan id, jenis kelamin, umur, tekanan darah tinggi, riwayat penyakit jantung, status perkawinan, tipe pekerjaan, tipe tempat tinggal, rata- rata kandungan gula darah, berat tubuh sempurna, perokok statis, serta

atribut dependennya merupakan penyakit stroke yang sebagian besar masih berformat kategorik(abjad) serta numerik.

C. Pemrosesan Awal Data

Setelah proses pengumpulan data selesai dilakukan, langkah selanjutnya adalah melakukan pengecekan apakah pada dataset tersebut terdapat data yang sama (*duplicated*) jika ada maka akan dihapus, selanjutnya melakukan pengecekan data yang hilang (*missing value*) jika terdapat data yang hilang maka akan diisi dengan rata-rata dari atribut tersebut. Selanjutnya mengecek data apakah data yang digunakan seimbang atau tidak. Karena dataset stroke ini tidak seimbang (misal jumlah pasien yang terkena stroke lebih sedikit dibandingkan dengan pasien yang tidak terkena stroke), maka diperlukan teknik penyeimbangan data.

Untuk mengatasi ketidakseimbangan data tersebut, digunakan teknik penyeimbangan *hybrid*. Teknik ini menggabungkan metode *oversampling* (menambah jumlah data minoritas) dan *undersampling* (mengurangi jumlah data mayoritas). Setelah proses penyeimbangan, data menjadi lebih representatif dan siap untuk diproses. Langkah selanjutnya adalah menghilangkan kolom ID dan kemudian melakukan transformasi data, karena algoritma *Random Forest* hanya dapat memproses data numerik, maka atribut yang berupa huruf atau abjad harus dikonversi menjadi angka melalui teknik transformasi data, seperti label encoding. Transformasi ini bertujuan agar data siap untuk diproses oleh model klasifikasi.

D. Pengolahan Data

Membagi data menjadi data training dan data testing. Perbandingan yang digunakan adalah 80% untuk data training dan 20% untuk data testing. Setelah data siap, varian algoritma Random Forest diterapkan satu per satu, varian algoritma yang digunakan adalah Weighted *Random Forest*, *Random Forest With* Feature Selection, *Random Forest With* Stratified Selection, *Random Forest With* AdaBoost, dan *Pararellizied Random Forest*. Setiap varian algoritma akan diuji coba untuk melihat akurasi dan standar deviasi dari hasil klasifikasi. Hasil dari setiap algoritma akan dicatat untuk dianalisis lebih lanjut.

E. Evaluasi Hasil

Berdasarkan hasil penerapan masing-masing algoritma akan dilakukan evaluasi akurasi, standar deviasi dan melihat plot kurva pembelajaran AUC (Area Under Curve) dan ROC (*Receiver Operating Characteristics*). Algoritma *Random Forest* yang menunjukkan akurasi terbaik dan standar deviasi terkecil akan dipilih sebagai algoritma terbaik untuk klasifikasi penyakit stroke. Dari hasil analisis dan evaluasi tersebut akan diambil kesimpulan mengenai algoritma *Random Forest* terbaik yang paling efektif dalam mengklasifikasikan penyakit stroke. Algoritma ini akan menjadi dasar untuk diimplementasikan lebih lanjut dalam sistem deteksi penyakit stroke.

Penelitian ini menggunakan beberapa metode tambahan untuk meningkatkan performa algoritma *Random Forest* seperti *Random Forest* Stratified Selection [17], *Random Forest* Feature Selection [18], Parallelization *Random Forest* [19], Weighted *Random Forest* [20], dan *Random Forest* dengan Adaboost [21].

III. HASIL DAN PEMBAHASAN

Langkah pertama adalah membaca dataset yang didapatkan melalui website kaggle dengan menggunakan library pandas yang tersedia pada aplikasi anaconda, setelah dataset terbaca, terdapat 12 fitur dimana kolom "stroke" merupakan variabel dependen (y) dan sisi kolom merupakan variabel independen (X), fitur dataset tersebut dapat dilihat pada Tabel 1.

TABEL 1
ATRIBUT DATASET STROKE

Kode	Variabel	Deskripsi
A	Id	unique identifier
B	Gender	"Male", "Female" or "Other"
C	Age	age of the patient
D	Hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
E	Heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
F	Ever_married	"No" or "Yes"
G	Work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
H	Residence_type	"Rural" or "Urban"
I	Avg_glucose_level	average glucose level in blood
J	Bmi	body mass index
K	Smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown" *Note: "Unknown" in smoking_status means that the information is unavailable for this patient
L	Stroke	1 if the patient had a stroke or 0 if not

Langkah kedua adalah mengecek dataset seperti data yang hilang (kosong), data yang terduplikasi, kemudian menanganinya. Setelah dilakukan pengecekan pada dataset, ternyata terdapat fitur "bmi" yang memiliki data kosong. Jumlah data yang hilang pada dataset dapat dilihat pada Tabel 2.

TABEL 2
MISSING VALUE PADA DATASET STROKE

Feature	Number of Missing Value
Id	0
Gender	0
Age	0
Hypertension	0
Heart_disease	0
Ever_married	0
Work_type	0
Residence_type	0
Avg_glucose_level	0
Bmi	201
Smoking_status	0
Stroke	0

Sampel nilai yang hilang dapat dilihat pada Gambar 2.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

Gambar 2. Sampel Mising Value pada kolom Bmi

Dari 5110 total record dataset, kolom "bmi" memiliki total 201 nilai yang hilang (kosong). Kemudian penulis mengatasi missing value tersebut dengan menggunakan nilai rata-rata dari fitur "bmi", dan juga menghapus "Id", hasil dari langkah ini dapat dilihat pada Tabel 3.

TABEL 3
HASIL IMPITASI MISSING VALUE

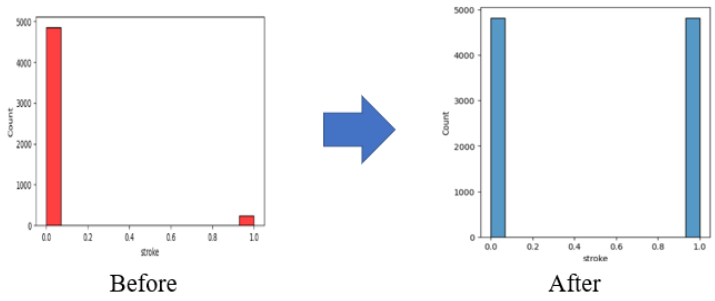
B	C	…	I	J	K	L
Female	61	…	202.21	28.8	never smoked	1
Male	80	…	105.92	32.5	never smoked	1
Female	49	…	171.23	34.4	smokes	1
Female	79	…	174.12	24	never smoked	1
…	…	…	…	…	…	...

Langkah selanjutnya adalah mentransformasikan fitur yang bernilai non-numerik (huruf) menjadi numerik dengan menggunakan teknik Labelencoder, hasilnya ditunjukkan pada Tabel 4. Langkah ini penting dilakukan agar dataset dapat dilatih menggunakan algoritma *Random Forest*.

TABEL 4
HASIL TRANSFORMASI DATA

B	C	D	E	F	G	H	…	K	L
1	67	0	1	1	2	1	…	1	1
0	61	0	0	1	3	0	…	2	1
1	80	0	1	1	2	0	…	2	1
0	49	0	0	1	2	0	…	3	1
0	79	1	0	1	3	1	…	2	1

Setelah melakukan transformasi data, peneliti menemukan bahwa dataset stroke mengalami ketidakseimbangan, ketidakseimbangan data dapat menyebabkan penurunan performa algoritma *Random Forest*, sehingga perlu dilakukan metode penyeimbangan data dengan menggunakan teknik *hybrid* (*SMOTETomek*). Hasil sebelum dan sesudah penyeimbangan data dapat dilihat pada Gambar 3.



Gambar 3. Hasil Balancing Data

Langkah selanjutnya adalah membagi data menjadi dua bagian, 80% untuk data training dan 20% untuk data testing. Kemudian latih data training dengan beberapa skema parameter pada algoritma *Random Forest* dan prediksi label menggunakan data testing. Skema-skema tersebut ditunjukkan pada Tabel 5. Evaluasi performa model menggunakan cross-validation dengan jumlah fold sebanyak 5.

TABEL 5
SKEMA PARAMETER RANDOM FOREST

Experiment Name	Parameter(s) and Value (s)
Weighted *Random Forest* (WRF)	n_estimators=100, random_state=42, class_weight='balanced'
Random Forest with Feature Selection (RFWFS)	SelectKBest(f_classif, k=5)),n_estimators=100, random_state=42,
Random Forest With Stratified Sampling (RFWSS)	Cross-validation method : StratifiedKFold(n_splits=5, shuffle=True, random_state=42), n_estimators=100, random_state=42
AdaBoost *Random Forest* (ABRF)	AdaBoostClassifier(estimator=RandomForestClassifier(n_estimators=50, random_state=42), n_estimators=100, algorithm="SAMME", random_state=42)
Random Forest with Parallelization (RFWP)	n_estimators=100, n_jobs=-1, random_state=42

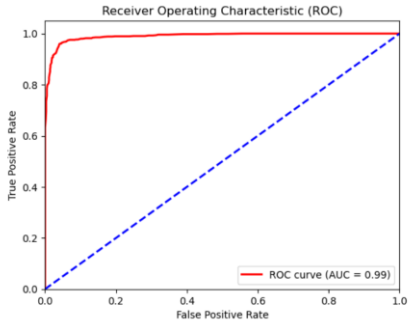
Hasil dari percobaan yang dilakukan adalah: laporan klasifikasi WRF ditunjukkan pada Tabel 6.

TABEL 6
HASIL KLASIFIKASI WRF

Class	Precision	Recall	F1-Score	Support
0	0.97	0.93	0.95	983
1	0.93	0.97	0.95	943
Accuracy			0.95	1926

Berdasarkan Tabel 6, model WRF memperoleh akurasi 95% dengan standar deviasi 0.00. Nilai precision kelas 0 mencapai 0.97, sedangkan recall 0.93, dan untuk kelas 1 precision 0.93 dengan recall 0.97. Hal ini menunjukkan keseimbangan prediksi antar kelas meskipun terdapat sedikit perbedaan pada recall. Konsistensi model ini dipengaruhi oleh penerapan pembobotan kelas (*class_weight='balanced'*) yang mampu menangani ketidakseimbangan data secara efektif [20] *Weighted Random Forest* memberikan bobot lebih besar pada kelas minoritas sehingga mengurangi kesalahan prediksi terhadap kelas dengan jumlah sampel lebih sedikit.

Nilai akurasi cross-validation yang dihasilkan oleh model WRH adalah 0.9543094496365524 dengan Standar Deviasi 0.00.



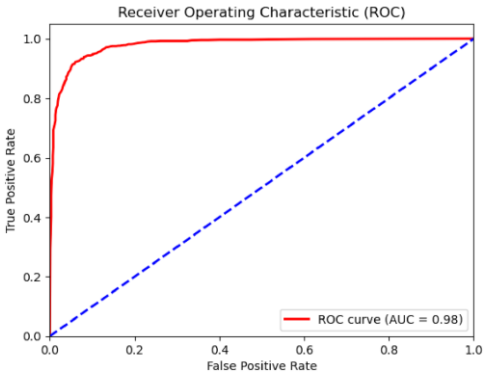
Gambar 4. Grafik ROC dan AUC WRF

Dari Gambar 4 terlihat bahwa kurva ROC menunjukkan model WRF dapat mengklasifikasikan data dengan hampir sempurna dengan nilai AUC sebesar 0.99. Hasil klasifikasi dari RFWFS dapat dilihat pada Tabel 7.

TABEL 7
HASIL KLASIFIKASI RFWFS

Class	Precision	Recall	F1-Score	Support
0	0.94	0.90	0.92	983
1	0.90	0.94	0.92	943
Accuracy			0.92	1926

Hasil pada Tabel 7 menunjukkan akurasi 92% dengan standar deviasi 0.00. Seleksi fitur menggunakan *SelectKBest* dengan lima fitur terbaik memberikan efisiensi model meskipun sedikit menurunkan akurasi dibandingkan WRF. Strategi ini sejalan dengan pen1elitian [18] yang menjelaskan bahwa seleksi fitur dapat mengurangi kompleksitas model, mempercepat proses pelatihan, dan mengurangi risiko overfitting tanpa mengorbankan terlalu banyak akurasi. Meskipun jumlah fitur berkurang, model tetap dapat mempelajari pola yang relevan. Nilai akurasi *cross-validation* yang dihasilkan oleh model RFWFS adalah 0.9236760124610592 dengan Standar Deviasi 0.00.



Gambar 5. Grafik ROC dan AUC RFWFS

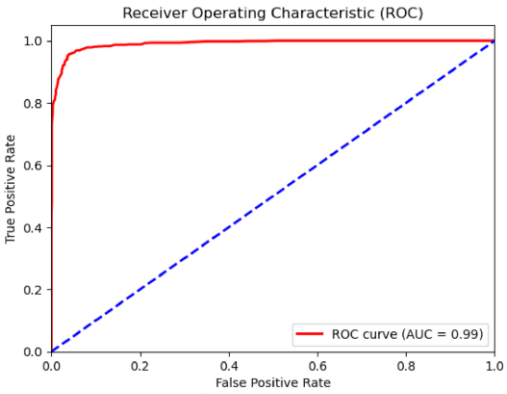
Dari Gambar 5 terlihat bahwa kurva ROC menunjukkan bahwa model RFWFS dapat mengklasifikasikan data dengan hampir sempurna meskipun lebih kecil dari model sebelumnya dengan nilai AUC sebesar 0,98. *Laporan klasifikasi RFWSS ditunjukkan pada Tabel 8.*

TABEL 8
HASIL KLASIFIAKAN RFWSS

Class	Precision	Recall	F1-Score	Support
0	0.94	0.90	0.92	983
1	0.90	0.94	0.92	943
Accuracy			0.92	1926

Berdasarkan Tabel 8, model RFWSS mencatat akurasi 95% dengan standar deviasi 0.00. Precision kelas 0 mencapai 0.97 dan recall 0.93, sedangkan kelas 1 memiliki precision 0.93 dan recall 0.97. Performa stabil ini sesuai dengan teori [17] yang menyatakan bahwa stratified sampling menjaga proporsi distribusi kelas pada setiap fold saat validasi silang, sehingga meningkatkan stabilitas dan konsistensi model pada dataset yang tidak seimbang.

Nilai akurasi cross-validation yang dihasilkan oleh model RFWSS adalah 0.952232606438214 dengan Standar Devias 0.00.



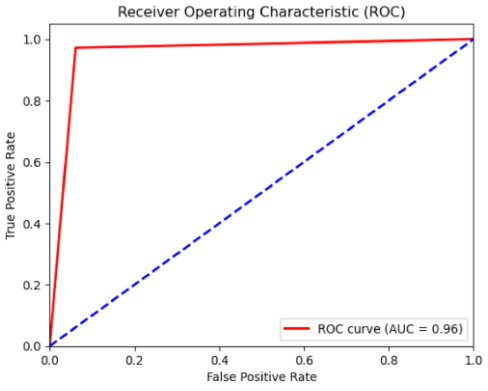
Gambar 6. Grafik ROC dan AUC RFWSS

Dari Gambar 6 dapat dilihat bahwa kurva ROC menunjukkan bahwa model RFWSS dapat mengklasifikasikan data dengan hampir sempurna dengan nilai AUC sebesar 0.99. Laporan klasifikasi dari ABRF ditunjukkan pada Tabel 9.

TABEL 9
HASIL KLASIFIKASI ABRF

Class	Precision	Recall	F1-Score	Support
0	0.97	0.94	0.96	983
1	0.94	0.97	0.96	943
Accuracy			0.96	1926

Hasil terbaik diperoleh pada model ABRF dengan akurasi 96% dan standar deviasi 0.00 (Tabel 9). Model ini memiliki precision dan recall yang tinggi pada kedua kelas ($\geqslant$ 0.94). Peningkatan ini selaras dengan kajian [21] yang menjelaskan bahwa kombinasi AdaBoost dengan Random Forest mampu mengurangi bias, meningkatkan akurasi, dan memperkuat kemampuan generalisasi model. AdaBoost memberi fokus pada sampel yang sulit diklasifikasikan, sementara Random Forest menjaga stabilitas prediksi. Nilai akurasi cross-validation yang dihasilkan oleh model RFWSS adalah 0.9553478712357217 dengan Standar Devias 0.00. Selama beberapa kali percobaan sebelumnya, model ini memiliki performa yang paling baik. Model ini memiliki peningkatan dalam mengenali kelas 1 dan meminimalisir prediksi yang salah baik untuk kelas 0 maupun 1.



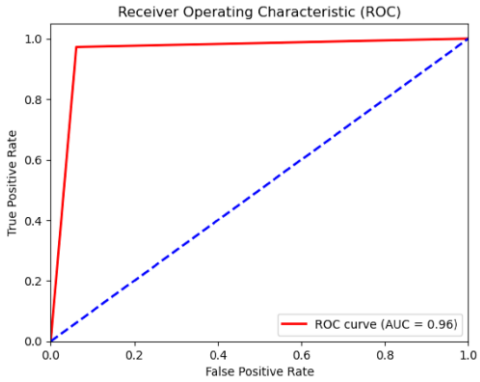
Gambar 7. Grafik ROC dan AUC ABRF

Dari Gambar 7 dapat dilihat bahwa kurva ROC menunjukkan bahwa model ABRF dapat mengklasifikasikan data dengan hampir sempurna dengan nilai AUC sebesar 0,96. *Laporan klasifikasi dari RFWP ditunjukkan pada Tabel 10.*

TABEL 10
HASIL KLASIFIKASI RFWP

Class	Precision	Recall	F1-Score	Support
0	0.97	0.93	0.95	983
1	0.93	0.97	0.95	943
Accuracy			0.95	1926

Pada Tabel 10, model RFWP memperoleh akurasi 95% dengan standar deviasi 0.00. Nilai precision–recall sebanding dengan WRF, namun waktu komputasi lebih efisien. Hal ini sesuai dengan hasil penelitian [19] yang menunjukkan bahwa paralelisasi proses pelatihan Random Forest dapat mempercepat pembentukan pohon tanpa menurunkan akurasi, terutama pada dataset berukuran besar. Nilai akurasi cross-validation yang dihasilkan oleh model RFWSS adalah 0.952232606438214 dengan Standar Devias 0.00.



Gambar 8. ROC dan AUC RFWP

Dari Gambar 8 dapat dilihat bahwa kurva ROC menunjukkan bahwa model RFWP dapat mengklasifikasikan data dengan hampir sempurna dengan nilai AUC sebesar 0.96. Berdasarkan hasil dan pembahasan, perbandingan kinerja dari teknik optimalasi algoritma *Random Forest* dapat dilihat pada table 11 berikut.

TABEL 11
HASIL PERBANDINGAN KINERJA TEKNIK PENGOPTIMALAN

No	Teknik Pengoptimalan	Akurasi	Std. Dev
1	Weigted *Random Forest* (WRF)	0.95	0.0
2	*Random Forest* With Feature Selection (RFWFS)	0.92	0.0
3	*Random Forest* With Stratified Selection (RFWSS)	0.95	0.0
4	Ada Boost *Random Forest* (ABRF)	0.96	0.0
5	*Random Forest* With Pararelazation (RFWP)	0.95	0.0

Berdasarkan tabel 11 nilai akurasi dari teknik pengoptimalan menggunakan *Weighted Random Forest* sebesar 0.95 atau 95% dengan nilai *Standart Deviasi sebesar 0.00 atau 0%*, nilai akurasi dari teknik pengoptimalan menggunakan *Random Forest With Feature Selection* sebesar 0.92 atau 92% dengan nilai *Standart Deviasi sebesar 0.00 atau 0%*, nilai akurasi dari teknik pengoptimalan menggunakan *Random Forest With Stratified Selection* sebesar 0.95 atau 95% dengan nilai *Standart Deviasi sebesar 0.00 atau 0%*, nilai akurasi dari teknik pengoptimalan menggunakan *AdaBoost Random Forest* sebesar 0.96 atau 96% dan nilai *Standart Deviasi* sebesar 0.00 atau 0%, *%*, nilai akurasi dari teknik pengoptimalan menggunakan *Random Forest With Pararelazation* sebesar 0.95 atau 95% dan nilai *Standart Deviasi* sebesar 0.00 atau 0%. Berdasarkan hasil peneltian diatas, penelitian ini masih memiliki beberapa keterbatasan dalam proses *preprocessing*. Penelitian ini belum menerapkan cara dalam mengangani kebocoran data.

IV. SIMPULAN

Penerapan metode data balancing pada dataset stroke dapat meningkatkan performa algoritma *Random Forest*, akurasi pada penelitian sebelumnya sebesar 94% dengan standar deviasi 2%, dengan penerapan teknik Adaboost tersebut mampu meningkatkan akurasi menjadi 96% dengan standar deviasi 0% dengan nilai kurva ROC (AUC) sebesar 0.96 atau 96%. Untuk penelitian selanjutnya, dapat melakukan overfitting pada dataset, melakukan optimasi yang lebih kompleks, dan dapat mencegah kebocoran data.

UCAPAN TERIMAKASIH

Terimakasih penulis ucapkan kepada Universitas Nusantara PGRI Kediri dan Lembaga Penelitian dan Pengabdian Masyarakat Universitas Nusantara PGRI atas terlaksananya penelitian dan publikasi ini.

DAFTAR PUSTAKA

[1] W. Riyadina and E. Rahajeng, "Determinan Penyakit Stroke," *Kesmas Natl. Public Heal. J.*, vol. 7, no. 7, p. 324, Feb. 2013, doi: 10.21109/kesmas.v7i7.31.

[2] A. Ramadhanu, R. Ayu Mahessya, M. Raihan Zaky, M. Isra, S. Informasi, and U. Putra Indonesia YPTK Padang, "PENERAPAN TEKNOLOGI MACHINE LEARNING DENGAN METODE VADER PADA APLIKASI SENTIMEN TAMU DI HOTEL DYMENS," *JOISIE J. Inf. Syst. Informatics Eng.*, vol. 7, no. 1, pp. 165–173, Jun. 2023.

[3] M. S. Pathan, Z. Jianbiao, D. John, A. Nag, and S. Dev, "Identifying Stroke Indicators Using Rough Sets," *IEEE Access*, vol. 8, pp. 210318–210327, Nov. 2020, doi: 10.1109/ACCESS.2020.3039439.

[4] T. K. Amarya, A. C. A. Galuh, R. Achmad, E. Daniati, and A. Ristyawan, "Analisa Perbandingan Algoritma Classification Berdasarkan Komposisi Label," *Pros. SEMNAS INOTEK*, vol. 8, no. 1, pp. 32–40, Aug. 2024, doi: https://doi.org/10.29407/inotek.v8i1.4906.

[5] A. P. Wibawa, M. Guntur, A. Purnama, M. Fathony Akbar, and F. A. Dwiyanto, "Metode-metode Klasifikasi," *Pros. Semin. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 1, 2018.

[6] U. Bradter, J. D. Altringham, W. E. Kunin, T. J. Thom, J. O'Connell, and T. G. Benton, "Variable ranking and selection with random forest for unbalanced data," *Environ. Data Sci.*, vol. 1, pp. 1–23, Nov. 2022, doi: 10.1017/eds.2022.34.

[7] T. Krisna Amarya, A. G. Candra Andy, R. Achmad, E. Daniati, and A. Ristyawan, "Analisa Perbandingan Algoritma Classification Berdasarkan Komposisi Label," *Pros. SEMNAS INOTEK*, vol. 8, pp. 2549–7952, Aug. 2024, Accessed: Sep. 30, 2024. [Online]. Available: https://proceeding.unpkediri.ac.id/index.php/inotek

[8] M. Guhdar, A. Ismail Melhum, and A. Luqman Ibrahim, "Optimizing Accuracy of Stroke Prediction Using Logistic Regression," *J. Technol. Informatics*, vol. 4, no. 2, pp. 41–47, 2023, doi: 10.37802/joti.v4i2.278.

[9] E. Wulandari *et al.*, "Classification Of Stroke Prediction Using The Support Vector Machine (SVM) Method 1,2," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 3, 2024.

[10] W. Wang *et al.*, "A systematic review of machine learning models for predicting outcomes of stroke with structured data," *PLoS One*, vol. 15, no. 6, pp. 1–16, 2020, doi: 10.1371/journal.pone.0234722.

[11] X. Huang *et al.*, "Novel Insights on Establishing Machine Learning-Based Stroke Prediction Models Among Hypertensive Adults," *Front. Cardiovasc. Med.*, vol. 9, no. May, pp. 1–11, 2022, doi: 10.3389/fcvm.2022.901240.

[12] Muvida and F. Amar, "The Features of Comorbidity of Stroke in The Indonesian Population: Findings from The Indonesian Family Life Survey (IFLS-5)," *Magna Neurol.*, vol. 2, no. 2, pp. 42–47, Jul. 2024, doi: 10.20961/magnaneurologica.v2i2.948.

[13] M. Guhdar, A. Ismail Melhum, and A. Luqman Ibrahim, "Optimizing Accuracy of Stroke Prediction Using Logistic Regression," *J. Technol. Informatics*, vol. 4, no. 2, pp. 41–47, Apr. 2023, doi: 10.37802/joti.v4i2.278.

[14] E. Wulandari *et al.*, "Classification Of Stroke Prediction Using The Support Vector Machine (SVM)," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 3, pp. 17–29, Sep. 2024.

[15] T. K. Amarya, A. C. A. G, R. Achmad, E. Daniati, and A. Ristyawan, "Analisa Perbandingan Algoritma Classification Berdasarkan Komposisi Label," *Pros. SEMNAS INOTEK*, vol. 8, no. 1, pp. 32–40, 2024, doi: https://doi.org/10.29407/inotek.v8i1.4906.

[16] Fedesoriano, "Stroke Prediction Dataset," *https://www.kaggle.com/*, 2024. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data (accessed Oct. 28, 2024).

[17] Y. Ye, Q. Wu, J. Zhexue Huang, M. K. Ng, and X. Li, "Stratified sampling for feature subspace selection in random forests for high dimensional data," *Pattern Recognit.*, vol. 46, no. 3, pp. 769–787, Mar. 2013, doi: 10.1016/J.PATCOG.2012.09.005.

[18] H. Fei *et al.*, "Cotton Classification Method at the County Scale Based on Multi-Features and Random Forest Feature Selection Algorithm and Classifier," *Remote Sens.*, vol. 14, no. 4, pp. 1–28, Feb. 2022, doi: 10.3390/rs14040829.

[19] B. H. Sadiq and S. R. Zeebaree, "Parallel Processing Impact on Random Forest Classifier Performance: A CIFAR-10 Dataset Study," *Indones. J. Comput. Sci. Attrib.*, vol. 13, no. 2, pp. 1833–1846, Apr. 2024.

[20] M. Shahhosseini and G. Hu, "Improved Weighted Random Forest for Classification Problems."

[21] A. J. Wyner, M. Olson, J. Bleich, and D. Mease, "Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers," 2017. [Online]. Available: http://jmlr.org/papers/v18/15-240.html.