Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 847~855
e-ISSN: 2548-6861

Analysis of SMOTE and Random Search on Machine Learning Algorithms for Stroke Disease Diagnosis

Ubaid Khoir Julio Dn [1]*, Majid Rahardi [2]*
* Informatika, Universitas Amikom Yogyakarta
ubaidkhoir22@students.amikom.ac.id [1], majid@amikom.ac.id [2]

Article Info

Article history:

Received 2025-12-16
Revised 2025-12-29
Accepted 2026-01-08

Keyword:

CatBoost,
Machine Learning,
Random Search,
SMOTE,
Stroke Prediction.

ABSTRACT

Stroke is a critical medical condition in which false negative predictions may lead to delayed treatment and increased mortality. Therefore, predictive models in the medical domain should prioritize sensitivity (recall) in addition to overall accuracy. This study analyzes the impact of the Synthetic Minority Over-sampling Technique (SMOTE) and Random Search hyperparameter optimization on five machine learning algorithms—Random Forest, XGBoost, Support Vector Machine (SVM), Logistic Regression, and CatBoost—for stroke disease diagnosis. Two experimental scenarios were conducted, namely models trained without SMOTE and models trained with SMOTE applied only to the training data to prevent data leakage. Model performance was evaluated using accuracy, precision, recall, and F1-score, with particular emphasis on recall due to its clinical relevance. In clinical practice, low recall may lead to false negative predictions, where high-risk stroke patients are not identified by the system, potentially resulting in delayed medical intervention. Therefore, recall is emphasized as the primary performance metric in this study. Experimental results demonstrate that SMOTE consistently improves recall across all models, while Random Search further enhances performance. CatBoost achieved the best performance with an accuracy of 96.61%, recall of 97%, and F1-score of 97%. Despite its superior performance, potential overfitting risks are critically discussed. These findings indicate that the proposed approach produces a clinically relevant decision-support model for stroke risk prediction.

I. PENDAHULUAN

Stroke merupakan kondisi medis serius yang terjadi ketika aliran darah menuju otak terhambat, sehingga memicu kerusakan jaringan saraf yang dapat menyebabkan kecacatan permanen maupun kematian[1]. Organisasi Kesehatan Dunia (WHO) mencatat bahwa stroke merupakan penyebab kematian kedua di dunia, sementara sebagian besar penderita mengalami gangguan fungsi jangka panjang. Angka terjadi stroke sekitar 15 juta orang setiap tahun, lebih dari 87% kematian akibat stroke terjadi di negara-negara yang memiliki tingkat pendapatan menengah dan rendah[2]. Pada tahun 2020 berdasarkan profil Kesehatan Indonesia tercatat cukup tinggi penduduk yang mengalami atau menderita stroke yaitu sebesar 1.789.261 [3]. Menurut Riskesdes tahun 2018 penduduk dengan usia 15 tahun ke atas yang terkena stroke sekitar 14,7% dan meningkat dengan bertambahnya usia 75 tahun ke atas yaitu sebesar 50,2%[4]. Dengan keadaan seperti ini perlu ada nya penanganan cepat untuk memprediksi terkena stroke secara efisien dan akurat[5].

Seiring dengan perkembangan teknologi informasi dan komputasi, berbagai bidang mulai memanfaatkan kecanggihan untuk meningkatkan kualitas layanan dan efektivitas kerja termasuk dunia kesehatan. Machine Learning hadir sebagai salah satu metode yang dirancang untuk menganalisis pola data yang kompleks. Teknologi ini tidak hanya membantu dalam proses analisis data, tetapi juga memberikan dukungan penting dalam pengambilan keputusan klinis yang lebih cerdas[6].

Penelitian ini bertujuan untuk memprediksi stroke dengan membandingkan akurasi dari lima algoritma

Machine Learning. Random Forest (RF), XGBoost, Support Vector Machine (SVM), Logistic Regression, dan CatBoost. Masing-masing algoritma ini memiliki pendekatan komputasional yang berbeda, sehingga menghasilkan performa yang beragam terhadap dataset yang digunakan. Melalui penelitian ini dapat diketahui algoritma mana yang paling optimal naik dari segi akurasi, kemampuan generalisasi, maupun kemampuan mengolah data yang kompleks dalam konteks prediksi risiko stroke[7].

Salah satu tantangan dalam pengembangan prediksi penyakit stroke adalah ketidakseimbangan data atau class imbalance. Sebagian besar dataset medis menunjukkan bahwa jumlah pasien yang mengalami stroke jauh lebih sedikit dibandingkan pasien yang tidak mengalami stroke. Kondisi ini menyebabkan model Machine Learning cenderung bias terhadap kelas mayoritas sehingga mengabaikan kelas minoritas. Akibatnya menurunkan kemampuan model dalam mendeteksi kasus stroke secara akurat. Untuk mengatasi ini diperlukan metode penyeimbangan data berbasis oversampling, dengan menggunakan SMOTE (Synthetic Minority Oversampling Technique). Teknik ini dapat bekerja dan menghasilkan data yang seimbang terutama pada aspek sensitivitas dan F1-score, karena model menjadi lebih mampu mendeteksi pasien dengan risiko stroke secara lebih akurat[8].

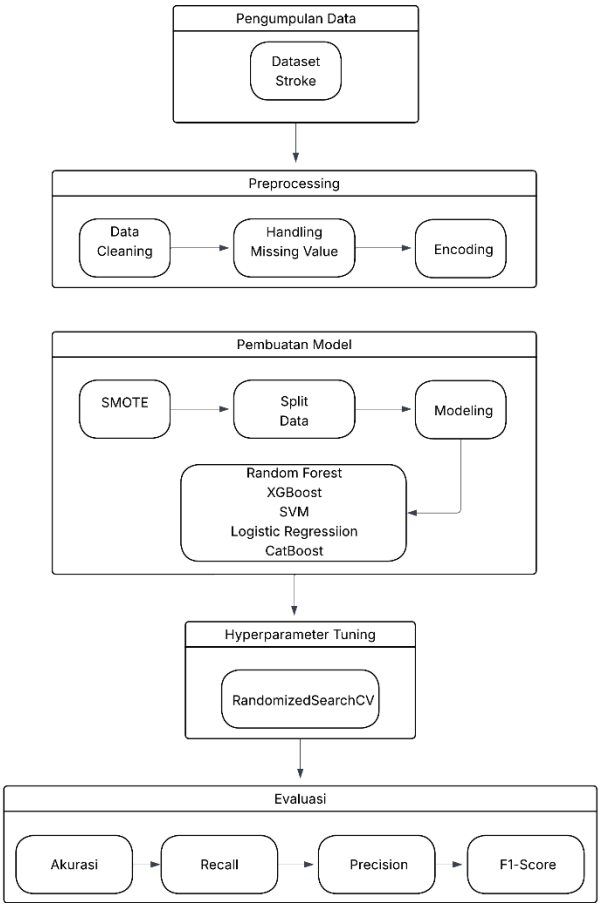
Selain permasalahan ketidakseimbangan data, prediksi stroke yang optimal juga memerlukan proses pemilihan parameter terbaik dengan tujuan memberikan performa yang maksimal. Dalam penelitian ini digunakan Random Search sebagai Teknik hyperparameter tuning karena metode ini lebih efisien dibandingkan Grid Search, terutama untuk jumlah kombinasi parameter sangat besar. Dengan demikian, penggunaan Random Search diharapkan dapat menghasilkan model prediksi stroke yang lebih optimal ketika dipadukan dengan teknik penyeimbangan data seperti SMOTE[9].

Penelitian ini diharapkan memberikan kontribusi nyata terhadap pencegahan penyakit stroke. Dengan membangun model prediksi risiko stroke yang menggabungkan Teknik penyeimbangan data (SMOTE) dan optimasi parameter (Random Search). Selain itu,hasil penelitian dapat menjadi sistem pendukung keputusan dan membantu dalam identifikasi pasien yang berisiko tinggi secara efisien bagi tenaga medis atau layanan kesehatan[10].

II. METODE

Dalam penelitian ini dilakukan beberapa tahapan sistematis yang dirancang untuk membangun model prediksi risiko stroke secara optimal. Seluruh proses ini dibagi menjadi beberapa tahapan, yaitu: (1) Pengumpulan data, (2) Preprocessing data, (3) Pembuatan model, (4) Hyperparameter tuning, (5) Evaluasi model.

Algoritma *Machine Learning* yang digunakan dalam analisis prediksi penyakit stroke ini antara lain: Random Forest (RF), XGBoost, Support Vector Machine (SVM), Logistic Regression, dan CatBoost. Seperti pada Gambar 1 berikut ini.



Gambar 1. Alur Penelitian

A. *Pengumpulan Data*

Data stroke yang digunakan dalam penelitian ini berasal dari sumber data publik di Kaggle yang terdiri dari 5110 baris data dengan 12 variabel. Dataset ini memuat informasi karakteristik, riwayat kesehatan, serta kebiasaan hidup yang relevan dalam proses prediksi risiko stroke.

B. *Preprocessing*

Tahap preprocessing dilakukan untuk meningkatkan kualitas data dan memastikan bahwa seluruh variabel berada dalam format yang sesuai dengan algoritma Machine Learning. Tahapan ini sangat penting karena kualitas pra-pemrosesan sangat memengaruhi performa model. Proses preprocessing diawali dengan data cleaning, yaitu menghapus fitur-fitur yang tidak relevan atau tidak memiliki pengaruh terhadap proses klasifikasi. Selanjutnya, dilakukan penanganan missing value, mengingat data medis sering mengandung nilai yang hilang, khususnya pada variabel numerik seperti Body

Mass Index (BMI). Pada penelitian ini digunakan metode imputasi median karena lebih robust terhadap keberadaan outlier dibandingkan nilai rata-rata (mean), sehingga distribusi data tetap terjaga dan tidak menimbulkan bias pada proses pelatihan model. Setelah itu, seluruh variabel kategorikal dikonversi ke dalam bentuk numerik menggunakan teknik label encoding, yang mengubah setiap kategori menjadi representasi angka agar dapat diproses oleh algoritma machine learning pada tahap pemodelan.

C. Pembuatan Model

Setelah data berada dalam kondisi yang bersih dan siap, selanjutnya adalah pembuatan model prediksi. Dalam penelitian ini, proses pembangunan model dilakukan dengan dua skenario berbeda untuk melihat pengaruh penyeimbang data terhadap performa klasifikasi. Skenario pertama adalah pelatihan model menggunakan data asli tanpa penerapan SMOTE dan pelatihan skenario kedua dilakukan setelah data diseimbangkan menggunakan metode SMOTE. Proses pelatihan model dilakukan dengan cara sebagai berikut:

- SMOTE (Synthetic Minority Oversampling Technique)
  Permasalahan dalam dataset stroke adalah ketidakseimbangan kelas, di mana jumlah yang terkena stroke lebih sedikit dibandingkan jumlah yang tidak terkena stroke. Maka untuk mengatasinya digunakan Teknik SMOTE yang bekerja dengan membuat data sintetis. Teknik ini membantu menghasilkan data kelas menjadi lebih seimbang sehingga model dapat mempelajari dengan baik. Penerapan SMOTE hanya dilakukan pada data latih untuk menghindari terjadinya data leakage dan memastikan evaluasi model tetap objektif.
- Split Data
  Data yang sudah seimbang kemudian dibagi menjadi dua bagian, yaitu 80% untuk pelatihan (training set) dan 20% untuk pengujian (testing tes). Pembagian ini bertujuan untuk menguji kemampuan model dalam melakukan generalisasi terhadap data baru yang tidak pernah dilihat sebelumnya.
- Modeling

Tahap ini adalah inti dari proses pembuatan model prediksi. Setiap algoritma dilatih menggunakan data training untuk mempelajari pola hubungan antara fitur masukan dan kelas target. Algoritma ini dipilih karena masing-masing memiliki karakteristik berbeda dalam memproses data, sehingga memungkinkan perbandingan performa yang komprehensif. Beberapa algoritma Machine Learning yang digunakan, yaitu:

1) Random Forest (RF)
   Random Forest adalah algoritma berbasis ensemble yang membangun banyak pohon keputusan untuk menghasilkan prediksi yang lebih stabil[11].
2) XGBoost
   XGBoost adalah algoritma boosting yang terkenal dengan performanya yang tinggi. XGBoost juga sangat optimal dan efisien karena pohon yang dibangun secara berurutan, dan setiap pohon memperbaiki kesalahan pohon sebelumnya[12].
3) Logistic Regression
   Logistic Regression adalah algoritma statistic yang digunakan untuk memperkirakan probabilitas terjadinya suatu peristiwa dengan dua kemungkinan hasil[13]. Dan jika Logistic Regression dikombinasikan dengan metode penyeimbangan (SMOTE) dapat meningkatkan hasil prediksi[14]
4) Support Vector Machine (SVM)
   Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi[15]. SVM sering dipilih dalam domain kesehatan atau klasifikasi penyakit karena kemampuannya menangani data berdimensi tinggi serta kestabilannya dalam menghasilkan model dengan generalisasi baik[16].
5) CatBoost
   CatBoost adalah algoritma boosting berbasis pohon (gradient boosting) yang dirancang khusus untuk menangani data kategorikal dengan lebih efisien dibandingkan metode boosting lainnya[17]. CatBoost memiliki kemampuan generalisasi yang baik dan sering menunjukkan performanya unggul dibandingkan algoritma lain dalam tugas klasifikasi terutama ketika dataset memiliki variabel campuran.[18].

D. Hyperparameter Tuning

Proses RandomizedSearchCV dilakukan dengan menentukan ruang pencarian (search space) pada parameter-parameter utama setiap algoritma. Pada Random Forest, parameter yang dioptimalkan meliputi n_estimators (100–600), max_depth (3–10), dan max_features (sqrt, log2). Untuk XGBoost, parameter yang dieksplorasi meliputi n_estimators (100–600), learning_rate (0.01–0.3), max_depth (3–8), subsample (0.6–1.0), dan colsample_bytree (0.6–1.0). Logistic Regression dioptimalkan pada parameter C (0.01–10) dan solver. Pada SVM, parameter kernel, C (0.1–10), dan gamma (0.001–1) diatur. Sementara itu, CatBoost dioptimalkan pada parameter iterations (200–500), depth (4–8), dan learning_rate (0.01–0.3). Untuk mengoptimalkan performa di setiap algoritma maka digunakan Teknik hyperparameter tuning. Optimasi ini dilakukan dengan RandomizedSearchCV yaitu metode pencarian acak dalam ruang parameter yang telah ditentukan. Random Search dipilih karena lebih efisien dibandingkan Grid Search, terutama saat ruang parameter sangat besar[19]. Teknik ini juga dilengkapi dengan cross-validation sehingga setiap kombinasi

parameter diuji secara berulang untuk memastikan stabilitas performa model dan menghindari overfitting[20]. RandomizedSearchCV juga memberikan akurasi yang lebih baik serta waktu tuning jauh lebih cepat daripada Grid Search[21]. Tahap ini bertujuan untuk menemukan konfigurasi parameter terbaik.

E. Evaluasi Model

Langkah ini merupakan tahap akhir dari penelitian, yaitu evaluasi performa model untuk mengukur kemampuan model dalam memprediksi risiko stroke. Tahap ini bertujuan untuk memastikan bahwa model tidak hanya bekerja baik pada data pelatihan, tetapi juga mampu melakukan generalisasi pada data yang belum pernah dilihat sebelumnya. Melaui evaluasi tersebut peneliti dapat menentukan apakah model sudah memenuhi kriteria performa, serta membandingkan beberapa algoritma untuk memilih model yang memberikan hasil paling optimal dalam memprediksi risiko stroke[22]. Ada empat metrik evaluasi digunakan untuk memberikan gambaran yang lebih komprehensif, yaitu:

- Akurasi
  Akurasi mengukur seberapa besar proporsi prediksi yang benar dibandingkan seluruh data uji. Metrik ini menunjukkan performa keseluruhan model[23]. Rumusnya:
  $$Accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$
- Recall
  Recall sangat penting dalam konteks medis karena menunjukkan kemampuan model mendeteksi pasien yang benar-benar berisiko stroke (kelas positirf). Recall yang rendah dapat menyebabkan kasus stroke terlewat (false negative)[24].
  Rumusnya:
  $$Recall = \frac{\text{TP}}{\text{TP+FN}}$$
- Precision
  Precision mengukur ketetapan prediksi positif. Metrik ini memastikan bahwa pasien yang diprediksi stroke benar-benar memiliki risiko sehingga mengurangi false positive[25].
  Rumusnya:
  $$Precision = \frac{\text{TP}}{\text{TP+FP}}$$
- F1-Score
  F1-Score merupakan rata-rata harmonis antara precision dan recall. Metrik ini sangat relevan untuk dataset tidak seimbang karena memberikan penilaian yang lebih proporsional antara kedua aspek[26]. Rumusnya:
  $$F1 - Score = 2 \text{ x} \frac{\text{Precision x Recall}}{\text{Precision+Recall}}$$

III. HASIL DAN PEMBAHASAN

Pada bab ini disajikan hasil penelitian yang diperoleh dari seluruh tahapan yang telah dilakukan, mulai dari pengolahan data, penerapan metode penyeimbangan data, implementasi algoritma machine learning, hingga evaluasi kinerja model. Pembahasan difokuskan pada analisis hasil eksperimen untuk menilai performa masing-masing algoritma dalam memprediksi risiko penyakit stroke, baik sebelum maupun sesudah dilakukan *hyperparameter tuning*. Selain itu, pada bab ini juga dilakukan perbandingan hasil penelitian dengan penelitian sebelumnya guna menunjukkan posisi dan kontribusi penelitian ini dalam pengembangan model prediksi stroke berbasis machine learning.

A. Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 5110 baris data dan 12 atribut yang berkaitan dengan faktor risiko stroke. Berikut menunjukkan daftar atribut berserta deskripsinya.

TABEL I
DATASET STROKE

No	Atribut	Deskripsi
1	id	identifikasi unik setiap pasien
2	gender	jenis kelamin pasien Male or Female
3	age	usia pasien dalam satuan tahun
4	hypertension	bernilai 0 apabila pasien tidak memiliki hipertensi, dan 1 apabila pasien memiliki hipertensi
5	heart_disease	bernilai 0 apabila pasien tidak memiliki riwayat penyakit jantung, dan 1 apabila memiliki penyakit jantung
6	ever_married	status pernikahan pasien Yes or No
7	work_type	jenis pekerjaan: Children, govt_job, never_worked, private, atau self-employed
8	residence_type	tipe tempat tinggal: Urban atau Rural
9	avg_glucose_level	rata-rata kadar glukosa dalam darah
10	bmi	indeks masa tubuh pasien
11	smoking_status	status merokok: formerly smoked, never smoked, smokes or unknown
12	stroke	label kelas: bernilai 1 jika pasien pernah mengalami stroke, dan 0 jika tidak

Analisis distribusi kelas menunjukkan bahwa dataset yang digunakan memiliki ketidakseimbangan kelas yang signifikan, dengan 4.861 data termasuk kelas tidak stroke dan hanya 249 data pada kelas stroke. Kondisi ini menegaskan perlunya penerapan teknik penyeimbangan data seperti SMOTE untuk meningkatkan sensitivitas model terhadap kelas minoritas.

B. Preprocessing

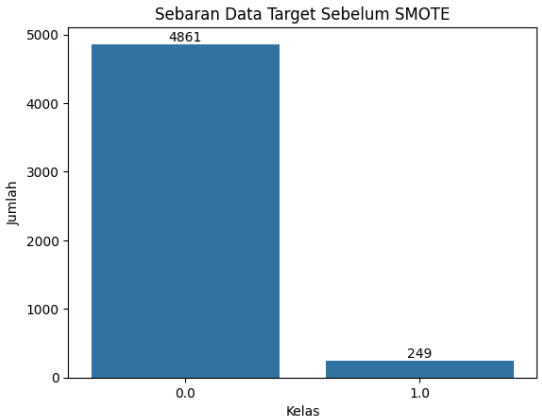
Tahapan preprocessing dilakukan untuk memastikan data sebelum digunakan pada proses pelatihan model. Berikut rangkaian preprocessing yang dilakukan:

1) Data Cleaning
   Penghapusan Kolom Tidak Relevan. Kolom id dihapus karena tidak memiliki kontribusi terhadap proses prediksi dan hanya berfungsi sebagai identitas.
2) Penanganan Missing Values
   Pemeriksaan awal menunjukkan bahwa atribut *bmi* mengandung nilai kosong. Nilai kosong pada kolom BMI diisi menggunakan metode imputasi median untuk menjaga distribusi data dan mengurangi pengaruh outlier.
3) Encoding Data Kategorikal
   Atribut kategorikal kolom gender, ever_married, residence_type, work_type dan smoking_status diubah menjadi bentuk numerik. Proses encoding dilakukan agar fitur kategorikal dapat diolah oleh algoritma klasifikasi.

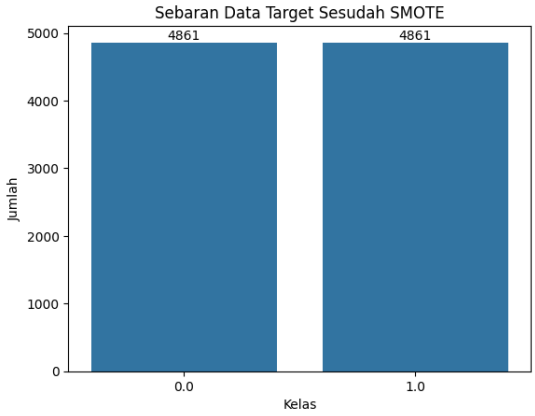
C. Modelling

Setelah keseluruhan tahap preprocessing dilakukan, dataset berada dalam kondisi bersih dan siap untuk proses modelling. Proses modelling meliputi:

1) SMOTE (Synthetic Minority Over-sampling Technique)
   Dataset stroke memiliki ketidakseimbangan kelas yang tinggi, di mana jumlah data pasien yang tidak mengalami stroke jauh lebih besar dibandingkan pasien yang mengalami stroke. Berikut adalah gambar data sebelum dan sesudah di SMOTE.



Gambar 2. Distribusi kelas dataset sebelum penerapan SMOTE.



Gambar 3. Distribusi kelas dataset setelah penerapan SMOTE.

Gambar 2 menunjukkan kondisi data sebelum dilakukan penyeimbangan data menggunakan metode SMOTE. Terlihat jumlah data pada kelas negatif (tidak mengalami stroke) jauh lebih banyak dibandingkan kelas positif (mengalami stroke). Jumlah kelas negatif 4861 dan kelas positif 249. Sedangkan Gambar 3 menunjukkan kondisi data sesudah diterapkan metode SMOTE. Berdasarkan hasil eksperimen, penerapan SMOTE pada data latih menghasilkan jumlah data yang seimbang pada kedua kelas. Dengan kondisi data yang telah seimbang, proses pelatihan model menjadi lebih optimal dalam mempelajari pola risiko stroke.

2) Pembagian Data Train dan Test
   Data dibagi menjadi rasio 8:2, yaitu: 80% data latih (train) dan 20% data uji (test). Pembagian ini memastikan model memiliki data yang cukup untuk belajar sekaligus generalisasi pada data baru.
3) Implementasi Algoritma
   Pada tahap implementasi algoritma, penelitian ini menerapkan lima algoritma machine learning, yaitu Random Forest, XGBoost, Logistic Regression, Support Vector Machine (SVM), dan CatBoost, dengan parameter yang telah ditentukan sebagaimana ditunjukkan pada Tabel II. Algoritma Random Forest diimplementasikan dengan jumlah pohon (*n_estimators*) sebanyak 300, *class_weight* bernilai *balanced* untuk menangani ketidakseimbangan kelas, serta *random_state* 42 guna menjaga konsistensi hasil. XGBoost menggunakan *n_estimators* 300, *learning_rate* 0,1, dan *max_depth* 4, dengan penyesuaian *scale_pos_weight* berdasarkan rasio kelas serta

eval_metric logloss sebagai fungsi evaluasi. Logistic Regression diimplementasikan dengan *max_iter* 500, *class_weight balanced*, dan *solver liblinear* untuk memastikan proses optimasi berjalan optimal. Pada SVM, digunakan *kernel Radial Basis Function (RBF)* dengan *class_weight balanced* serta proses penskalaan data menggunakan **StandardScaler** untuk meningkatkan performa model. Sementara itu, CatBoost diimplementasikan dengan 300 iterasi, *depth* 6, *learning_rate* 0,05, dan *loss_function logloss*, dengan *verbose* dinonaktifkan. Seluruh algoritma diimplementasikan menggunakan parameter tersebut untuk memperoleh performa yang optimal dan dapat dibandingkan secara objektif.

TABEL II
HASIL ARSITEKTUR ALGORITMA

Algoritma	Parameter	Value
Random Forest	n_estimators	300
	class_weight	balanced
	random_state	42
XGBoost	n_estimators	300
	learning_rate	0.1
	max_depth	4
	scale_pos_weight	rasio kelas
	eval_metric	logloss
Logistic Regression	max_iter	500
	class_weight	balanced
	Solver	liblinear
Support Vector Machine	Kernel	rbf
	class_weight	balanced
	scaler	StandardScaler
CatBoost	iterations	300
	depth	6
	learning_rate	0.05
	loss_function	logloss
	verbose	false

D. Hasil Sebelum Tuning

Pada tahap awal evaluasi, seluruh algoritma Machine Learning diuji menggunakan parameter dasar tanpa dilakukan proses hyperparameter tuning. Tahap ini bertujuan untuk mengetahui performa awal (baseline) dari masing-masing model terhadap dataset yang digunakan. Evaluasi ini dilakukan menggunakan beberapa metrik, yaitu akurasi, precision, recall, dan f1-score. Hasil evaluasi ditampilkan pada tabel berikut.

TABEL III
HASIL EVALUASI SEBELUM TUNING

Algoritma	Akurasi	Precision	Recall	F1-Score
Random Forest	0.9568	0.95	0.97	0.96
XGBoost	0.9481	0.95	0.95	0.95
Logistic Regression	0.8200	0.81	0.84	0.82
Support Vector Machine	0.8946	0.81	0.84	0.82
CatBoost	0.9414	0.94	0.95	0.94

Hasil evaluasi sebelum tuning menunjukkan bahwa algoritma berbasis ensemble, khususnya Random Forest, XGBoost, dan CatBoost memberikan performa paling unggul dibandingkan algoritma lainnya.

E. Hasil Sesudah Tuning

Setelah dilakukan proses *hyperparameter tuning* menggunakan metode RandomizedSearhCV, terdapat perubahan konfigurasi parameter model. Hasil tuning menunjukkan bahwa setiap algoritma mengalami penyesuaian parameter utama yang memengaruhi struktur dan kompleksitas model dibandingkan dengan parameter awal yang digunakan pada tahap modelling. Penyesuaian ini bertujuan untuk memperoleh konfigurasi model yang lebih optimal dalam mempelajari pola data, meningkatkan kemampuan generalisasi, serta memaksimalkan performa klasifikasi risiko stroke. Parameter terbaik hasil tuning untuk masing-masing algoritma disajikan pada Tabel IV.

TABEL IV
HASIL ARSITEKTUR ALGORITMA SETELAH TUNING

Algoritma	Parameter	Value
Random Forest	n_estimators	558
	max_depth	5
	max_features	sqrt
	bootstrap	True
	random_state	42
XGBoost	n_estimators	558
	learning_rate	0.1158
	max_depth	5
	subsample	0.7670
	colsample_bytree	0.8238
	gamma	0.0026
	reg_lambda	0.3293
	eval_metric	logloss
Logistic Regression	C	2.4476
	solver	liblinear
	max_iter	500
	class_weight	balanced
Support Vector Machine	kernel	rbf
	C	9.7563
	gamma	0.0818
	class_weight	balanced
	scaler	StandardScaler
CatBoost	iterations	389
	depth	6
	learning_rate	0.1955
	l2_leaf_reg	1.9061
	loss_function	logloss
	verbose	False

Setelah diperoleh parameter terbaik hasil *hyperparameter tuning* pada masing-masing algoritma sebagaimana ditunjukkan pada Tabel IV, tahap selanjutnya adalah melakukan evaluasi kinerja model. Evaluasi ini bertujuan untuk mengetahui pengaruh penyesuaian parameter terhadap performa klasifikasi risiko stroke. Hasil evauasi kinerja algoritma setelah tuning disajikan pada Tabel V.

TABEL V
HASIL EVALUASI SESUDAH TUNING

Algoritma	Akurasi	Precision	Recall	F1-Score
Random Forest	0.9604	0.96	0.96	0.96
XGBoost	0.9640	0.96	0.97	0.96
Logistic Regression	0.8200	0.81	0.84	0.82
Support Vector Machine	0.9105	0.90	0.93	0.91
CatBoost	0.9661	0.97	0.97	0.97

Berdasarkan hasil evaluasi setelah tuning yang disajikan pada Tabel V, dapat dilihat bahwa sebagian besar algoritma mengalami peningkatan performa dibandingkan dengan hasil sebelum tuning. Algoritma CatBoost menunjukkan performa terbaik dengan nilai akurasi sebesar 0.9661, serta nilai precision, recall, dan F1-Score yang masing-masing mencapai 0.97. Hal ini menunjukkan bahwa proses tuning parameter pada CatBoost mampu meningkatkan kemampuan model signifikan dan menghasilkan performa yang stabil pada kedua kelas. Algoritma XGBoost dan algoritma Random Forest juga menunjukkan performa yang sangat baik setelah tuning, dengan nilai akurasi masing-masing sebesar 0.9640 dan 0.9604, serta nilai F1-Score yang tinggi dan seimbang. Untuk algoritma Support Vector Machine (SVM) mengalami peningkatan performa yang cukup jelas setelah tuning, khususnya pada nilai recall kelas positif. Berbeda dengan algoritma lainnya, Logistic Regression tidak menunjukkan peningkatan performa yang signifikan setelah tuning, yang mengindikasikan keterbatasan model linear dalam menangkap pola non-linear pada dataset ini. Secara keseluruhan, hasil ini menegaskan bahwa proses *hyperparameter tuning* berperan penting dalam meningkatkan kinerja model, terutama pada algoritma berbasis ensemble, dan CatBoost ditetapkan sebagai model terbaik dalam penelitian ini.

Meskipun CatBoost menunjukkan performa tertinggi dibandingkan algoritma lainnya, hasil ini perlu diinterpretasikan secara hati-hati. Penerapan teknik SMOTE yang dikombinasikan dengan proses hyperparameter tuning berpotensi meningkatkan risiko overfitting, terutama karena evaluasi model dilakukan pada dataset yang sama tanpa validasi eksternal. Untuk meminimalkan risiko tersebut, proses optimasi parameter pada penelitian ini telah menggunakan cross-validation melalui RandomSearchCV. Namun demikian, penelitian ini lanjutan dengan menggunakan dataset klinis independen sangat diperlukan untuk memastikan kemampuan generalisasi model CatBoost dalam konteks implementasi nyata.

Selain evaluasi performa, aspek interpretabilitas model menjadi hal yang penting dalam aplikasi medis. Meskipun CatBoost merupakan model berbasis ensemble, penelitian-penelitian sebelumnya pada dataset stroke yang sama menunjukkan bahwa variabel usia, kadar glukosa rata-rata, hipertensi, penyakit jantung, dan indeks massa tubuh (BMI) merupakan faktor yang paling berpengaruh dalam prediksi stroke. Faktor-faktor tersebut memiliki relevansi klinis yang kuat dan sejalan dengan pengetahuan medis, sehingga meningkatkan kepercayaan terhadap model yang diusulkan sebagai sistem pendukung keputusan. Analisis ini difokuskan pada metrik recall karena memiliki implikasi klinis yang signifikan dalam diagnosis penyakit stroke.

TABEL VI
PERBANDINGAN KINERJA MODEL SEBELUM DAN SESUDAH SMOTE

Algoritma	Recall Tanpa SMOTE	Recall Dengan SMOTE
Random Forest	0.91	0.96
XGBoost	0.92	0.97
Logistic Regression	0.70	0.84
Support Vector Machine	0.78	0.93
CatBoost	0.94	0.97

Tabel VI menunjukkan bahwa penerapan SMOTE memberikan peningkatan signifikan pada nilai recall seluruh algoritma yang diuji. Peningkatan ini terlihat paling jelas pada algoritma Support Vector Machine dan Logistic Regression yang sebelumnya memiliki sensitivitas rendah terhadap kelas minoritas. Dalam konteks medis, peningkatan recall sangat krusial karena mampu menurunkan risiko false negative, yaitu kondisi ketika pasien yang sebenarnya berisiko stroke tidak terdeteksi oleh sistem, sehingga berpotensi menyebabkan keterlambatan penanganan klinis.

Dalam praktik klinis, model yang diusulkan dapat diintegrasikan sebagai sistem pendukung keputusan dengan menghasilkan nilai probabilitas risiko stroke, bukan sekadar klasifikasi biner. Informasi ini dapat membantu tenaga medis dalam melakukan skrining awal dan menentukan prioritas pasien yang memerlukan pemeriksaan lanjutan, seperti pemeriksaan laboratorium atau pencitraan medis. Model ini tidak dimaksudkan untuk menggantikan keputusan klinis, melainkan sebagai alat bantu untuk meningkatkan deteksi dini risiko stroke.

F. Perbandingan Hasil Penelitian

Beberapa penelitian sebelumnya telah mengeksplorasi penggunaan metode pembelajaran mesin untuk prediksi risiko stroke, dengan berbagai teknik penangan ketidakseimbangan kelas dan optimasi model. Sutcu et al.[27] melakukan analisis komprehensif pada dataset stroke yang sama (5110 observasi) dan menyoroti variabel-variabel penting seperti usia, kadar glukosa rara-rata, BMI, hipertensi, dan penyakit jantung sebagai prediktor utama: penelitian tersebut melaporkan hasil model yang sebanding pada metrik-metrik utama meskipun metologi pra-pemrosesan dan penanganan imbalanced berbeda.

Natasha et al.[28] melaporkan bahwa penerapan SMOTE dikombinasikan dengan algoritma XGBoost mampu meningkatkan performa prediksi stroke, khususnya pada metrik sensitivitas dan F1-score. Temuan ini mendukung pentingnya penanganan class imbalance dalam meningkatkan kemampuan model mendeteksi kasus stroke.

TABEL VII
HASIL PERBANDINGAN PENELITIAN

Peneliti	Model	Akurasi	Keterangan
Sutcu M et al[27]	CatBoost	0.94	Tanpa SMOTE
Natasha et al[28]	XGBoost	0.95	Menggunakan SMOTE
Akinwumi[29]	Random Forest	0.95	Recall relatif rendah
Penelitian Ini	CatBoost	0.97	SMOTE + Random Search, recall & F1 tinggi
	XGBoost	0.96	Performa seimbang
	Random Forest	0.96	Stabil pada semua metrik

Meskipun nilai akurasi yang diperoleh pada penelitian ini sebagaimana ditunjukkan pada Tabel VII relatif sebanding dengan beberapa penelitian sebelumnya, pendekatan yang digunakan dalam penelitian ini menunjukkan keunggulan pada metrik recall dan F1-score, yang sangat krusial dalam konteks medis. Nilai recall yang tinggi mengindikasikan kemampuan model dalam mendeteksi pasien yang benar-benar berisiko stroke secara efektif, sehingga dapat meminimalkan kesalahan false negative yang berpotensi membahayakan pasien. Dibandingkan penelitian sebelumnya yang umumnya hanya menekankan akurasi, penelitian ini menekankan keseimbangan performa model melalui penerapan SMOTE[30] dan RandomizedSearchCV[31], sehingga menghasilkan model yang tidak hanya akurat secara keseluruhan, tetapi juga lebih andal dalam mendeteksi kasus stroke. Dengan demikian, kontribusi utama penelitian ini tidak hanya terletak pada peningkatan sensitivitas dan kestabilan performa model, yang menjadikannya lebih relevan untuk diterapkan sebagai sistem pendukung Keputusan dalam bidang kesehatan.

G. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Dataset yang digunakan berasal dari respositori publik Kaggle sehingga belum tentu sepenuhnya merepresentasikan kondisi populasi klinis nyata dengan karakteristik pasien yang beragam. Selain itu, penelitian ini belum melakukan validasi eksternal menggunakan data rumah sakit atau data klinis independent, sehingga kemampuan generalisasi model masih terbatas. Penerapan teknik SMOTE menghasilkan data sintetis yang mungkin tidak sepenuhnya mencerminkan kondisi pasien sebenarnya. Oleh karena itu, penelitian lanjutan dengan melibatkan dataset klinis nyata serta validasi eksternal sangat diperlukan sebelum model ini dapat diterapkan secara luas dalam praktik klinis.

IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan teknik penyeimbangan data menggunakan SMOTE dan optimasi parameter menggunakan RandomizedSearchCV mampu meningkatkan performa model Machine Learning dalam memprediksi risiko penyakit stroke. Hasil evaluasi menunjukkan bahwa Sebagian besar algoritma mengalami peningkatan performa setelah dilakukan hyperparameter tuning, terutama pada metrik recall dan F1-score yang sangat penting dalam konteks medis.

Algoritma berbasis ensemble, khususnya CatBoost, XGBoost, dan Random Forest menunjukkan performa lebih unggul dibandingkan algoritma lainnya. CatBoost menjadi algoritma terbaik dengan nilai akurasi sebesar 96,61% serta precision, recall, dan F1-score masing-masing 97% yang menunjukkan kestabilan dan kemampuan generalisasi yang sangat baik. Sementara itu, Logistic Regression tidak menunjukkan peningkatan performa yang signifikan setelah tuning, yang mengindikasikan keterbatasan model linear dalam menangkap pola non-linear pada dataset stroke.

Secara keseluruhan, hasil penelitian ini membuktikan bahwa kombinasi Teknik SMOTE dan Random Search sangat efektif dalam menangani permasalahan ketidakseimbangan data dan optimasi parameter pada prediksi penyakit stroke. Model yang dihasilkan diharapkan dapat digunakan sebagai sistem pendukung Keputusan untuk membantu tenaga medis dalam mengidentifikasi pasien yang berisiko stroke secara lebih akurat dan efisien.

DAFTAR PUSTAKA

[1] D. Kuriakose, “Pathophysiology and Treatment of Stroke : Present Status and Future Perspectives,” 2020.
[2] A. Byna and M. Basit, “Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes,” vol. 09, no. November, pp. 407–411, 2020.
[3] K. Gorontalo, “Medic nutricia 2025,” vol. 13, no. 4, pp. 25–31, 2025, doi: 10.5455/mnj.v1i2.644xa.
[4] O. Acces, “Open Acces,” vol. 03, no. 01, pp. 1660–1665, 2021.
[5] M. M. Jakarta, “Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Stroke pada Rumah Sakit Pusat Otak Nasional Prof .,” vol. 26, no. 1, pp. 144–153.
[6] M. Putri, “Prediksi Penyakit Stroke Menggunakan Machine Learning Dengan Algoritma Random Forest,” vol. 9, no. 2, 2024.
[7] K. Sari *et al.*, “Deteksi Dini Stroke Menggunakan Machine Learning,” vol. 4, no. 4, pp. 706–720, 2025, doi: 10.55123/insologi.v4i4.5590.
[8] B. Nemade, V. Bharadi, S. S. Alegavi, and B. Marakarkandy, “Intelligent Systems And Applications In A Comprehensive Review : SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques , Evaluation , and Result Comparisons,” 2023.
[9] M. H. Rizky, M. R. Faisal, I. Budiman, and D. Kartini, “Effect

of Hyperparameter Tuning Using Random Search on Tree-Based Classification Algorithm for Software Defect Prediction," vol. 18, no. 1, pp. 95–106, 2024.
[10] A. Hassan, S. G. Ahmad, and N. Ramzan, "Predictive modelling and identification of key risk factors for stroke using machine learning," *Sci. Rep.*, no. 0123456789, pp. 1–23, 2024, doi: 10.1038/s41598-024-61665-4.
[11] X. Yuan, S. Liu, W. Feng, and G. Dauphin, "Feature Importance Ranking of Random Forest-Based End-to-End Learning Algorithm," pp. 1–20, 2023.
[12] M. R. Kurniawanda, F. Adline, T. Tobing, and P. S. Informatika, "Analysis Sentiment Cyberbullying in Instagram Comments with XGBoost Method," vol. 9, no. 1, 2022.
[13] A. A. Putri *et al.*, "Penerapan Metode Logistic Regression Untuk," vol. 7, no. 1, pp. 95–107.
[14] V. H. Vidiasari, H. Hairani, H. Santoso, and F. M. Amin, "Penerapan Logistic Regression dan SMOTE untuk Memprediksi Atrisi Karyawan pada Imbalanced Data," no. September, pp. 7–14, 2025.
[15] D. R. Nurqotimah, A. N. Khudori, and R. S. Pradini, "Journal Of Applied Computer Science And Technology (JACOST) Implementasi Algoritma Support Vector Machine (SVM) Untuk Klasifikasi Penyakit Stroke," vol. 5, no. 2, pp. 179–185, 2024.
[16] M. A. Ramadhani *et al.*, "Implementasi Algoritma Support Vector Machine (SVM) Untuk Diagnosis Kesehatan Manusia Berbasis Web," vol. 9, pp. 896–902, 2025.
[17] A. B. Mawardi, R. S. Pradini, M. S. Haris, and G. Boosting, "Boosting," vol. 13, no. 3.
[18] A. Informatics and A. Info, "Perbandingan Performa Algoritma XGBoost , CatBoost Dan," vol. 8, no. 1, pp. 268–273, 2025.
[19] M. Sholeh, U. Lestari, and D. Andayati, "Hyperparameter Optimization Using Grid Search and Random Search to Improve the Performance of Prediction Models with Decision Trees," vol. 3, no. 03, pp. 453–464, 2025.
[20] S. Fitria, A. Khansa, N. Ulinnuha, and W. D. Utami, "Grid Search And Random Search Hyperparameter Tuning Optimization In Xgboost Algorithm For Parkinson ' S Disease Classification," vol. 19, no. 3, pp. 1609–1624, 2025.
[21] J. Multidisiplin and D. Sains, "Randomsearchcv Untuk Meningkatkan Akurasi," vol. 1, no. 2, pp. 121–135, 2025.
[22] F. Adha, H. Airi, T. Suprapti, and A. Bahtiar, "Komparasi Metode Klasifikasi Data Mining Untuk Prediksi," vol. 18, pp. 73–79, 2023.
[23] S. Rahayu and S. F. Romdoni, "Bayesian Optimized Pretrained CNNs for Mango Leaf Disease Classification : A Comparative Study," vol. 6, no. 5, pp. 3051–3078, 2025.
[24] N. Abay, D. Istanto, B. Satrio, W. Poetro, U. Islam, and S. Agung, "Implementasi Algoritma Faster R-Cnn Dalam Deteksi," vol. 3, no. 1, pp. 43–59, 2025.
[25] S. Helmiyah, R. Pramestiawan, and R. Lampung, "Analisis Komparatif Algoritma Machine Learning dengan Metrik Akurasi , Presisi , Recall , dan F1-Score pada Dataset Kacang Kering," vol. 6, no. 3, pp. 152–159, 2025.
[26] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced Big Data," *J. Big Data*, 2023, doi: 10.1186/s40537-023-00724-5.
[27] M. Sutcu, D. Jouda, B. Yildiz, and J. Katrib, "Predicting Stroke Risk Using Machine Learning : A Data-Driven Approach to Early Detection and Prevention," vol. 2025, 2025.
[28] F. Natasha, B. Zahari, and K. Ramakrishnan, "Machine Learning-Driven Stroke Prediction Using Independent Dataset," vol. 8, no. May, 2024.
[29] P. O. Akinwumi *et al.*, "Evaluating machine learning models for stroke prediction based on clinical variables," no. September, 2025, doi: 10.3389/fneur.2025.1668420.
[30] G. Samudra *et al.*, "Efektivitas Teknik SMOTE Dalam Meningkatkan Performa Naïve Bayes Deteksi Gangguan Kecemasan Mahasiswa," vol. 12, no. 3, 2025.
[31] A. D. Rachmatsyah *et al.*, "Perbandingan Teknik Optimasi Grid Search dan Randomized Search dalam Meningkatkan Akurasi Metode Klasifikasi SVM Pada Sentimen Ulasan Pengguna Aplikasi JKN Mobile," vol. 8, pp. 13–22, 2025.