

Stroke Risk Classification Using the Ensemble Learning Method of XGBoost and Random Forest

Gullam Almuzadid ^{1*}, Egia Rosi Subhiyakto ^{2**}

* Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia

** Research Center for Intelligent Distributed Surveillance and Security (IDSS), Universitas Dian Nuswantoro, Semarang, 50131, Indonesia

111202113743@mhs.dinus.ac.id¹, egia@dsn.dinus.ac.id²

Article Info

Article history:

Received 2025-04-28

Revised 2025-05-03

Accepted 2025-05-06

Keyword:

*Ensemble Learning,
Random Forest,
SMOTE-ENN,
Stroke Risk Classification,
XGBoost.*

ABSTRACT

Stroke is a leading cause of global death and disability. This study proposes a stroke risk classification model using ensemble learning that combines Random Forest and XGBoost algorithms. A Kaggle dataset with 5110 samples (249 stroke, 4861 non-stroke) presented significant class imbalance. To address this, a comprehensive preprocessing pipeline was implemented, including feature encoding, feature scaling, feature selection using ANOVA F-test, outlier handling with Z-Score and IQR methods, and missing value imputation using MICE. The SMOTE-ENN approach was applied to handle class imbalance, resulting in a more balanced sample distribution. The dataset was split into 80% training and 20% testing data (hold-out test) to ensure objective evaluation. Hyperparameter optimization was performed using Bayesian optimization, while model evaluation employed stratified K-fold cross-validation to prevent overfitting. Validation on the hold-out test set demonstrated exceptional ensemble model performance with an AUC of 0.99, 98% accuracy, 98% precision, and 98% recall. Feature importance analysis identified average glucose level and age as the strongest stroke risk predictors. The proposed approach significantly improved predictive accuracy compared to previous research, demonstrating the effectiveness of ensemble learning and preprocessing methods in developing reliable, high-performing machine learning models for early stroke risk assessment.



This is an open access article under the [CC-BY-SA](#) license.

I. PENDAHULUAN

Salah satu penyebab utama kematian dan disabilitas di seluruh dunia adalah stroke, yang menyebabkan 6,6 juta kematian pada tahun 2022 [1]. Menurut World Stroke Organization pada tahun 2022, satu dari empat orang dewasa di atas usia 25 tahun akan mengalami stroke dalam hidupnya hal ini meningkat 50% dari 17 tahun terakhir. Di seluruh dunia terdapat 12,2 juta kasus stroke baru per tahun atau satu kasus per 3 detik. Stroke merupakan penyakit yang disebabkan tersumbatnya aliran darah ke otak dengan gejala hilangnya kemampuan gerak dan komunikasi serta gangguan kognitif. Penyakit stroke terpecah menjadi 2 jenis diantaranya, ishemic stroke dan hemorrhagic stroke [2]. Menurut World Stroke Organization (WSO) pada tahun 2022 secara global lebih dari 62% dari semua kasus stroke

merupakan ishemic stroke sedangkan sisanya ialah hemorrhagic stroke. Stroke ishemic disebabkan karena penyumbatan aliran darah sementara itu hemorrhagic stroke disebabkan karena pecahnya pembuluh darah [3].

Stroke adalah penyebab utama kematian dan disabilitas di Indonesia, menyumbang 11,2% dari kecacatan total dan 18,5% dari kematian total, menurut Kementerian Kesehatan. Data yang dikumpulkan oleh Survei Kesehatan Indonesia pada tahun 2023 menunjukkan bahwa tingkat prevalensi penyakit stroke di Indonesia adalah 8,3 per 1.000 orang. Pada tahun 2023, stroke akan mencapai Rp 5,2 triliun, menjadi salah satu penyakit katastropik dengan pembiayaan tertinggi ketiga setelah kanker dan penyakit jantung. Perawatan yang tepat serta diagnosis yang akurat memainkan peran penting dalam mengurangi dampak dari penyakit ini. Klasifikasi

stroke merupakan langkah awal untuk memahami pola, faktor risiko, dan karakteristik individu yang lebih rentan terhadap penyakit tersebut, yang dapat mendukung diagnosis yang lebih cepat dan mengurangi biaya yang meningkat dalam upaya penanggulangan penyakit jangka panjang [4]. Oleh karena itu, berkembangnya teknologi informasi dan pengajaran mesin telah membuka peluang baru untuk meningkatkan akurasi dan efektivitas diagnosis stroke..

Machine learning, memberikan berbagai teknik dan algoritma yang dapat digunakan untuk menganalisis data medis yang kompleks dan beragam, serta machine learning dapat mempelajari pola dari data riwayat dan memprediksi berdasarkan hasil analisis tersebut yang sangat bermanfaat dalam mendiagnosis penyakit [5]. Teknik machine learning digunakan untuk menganalisis pola yang terdapat dalam riwayat medis pasien, khususnya dalam mengidentifikasi faktor-faktor risiko yang berkaitan dengan stroke. Beberapa atribut penting, seperti usia, hipertensi dan lainnya yang diketahui berperan signifikan dalam mempengaruhi tingkat kerentanan seseorang terhadap penyakit stroke [6]. Berbagai algoritma machine learning, antara lain Random Forest, Naive Bayes, dan KNN dapat dimanfaatkan untuk mengatasi tantangan dalam proses diagnosis dan klasifikasi penyakit.

Penelitian terdahulu menunjukkan berbagai metode yang efektif untuk klasifikasi penyakit stroke. Riset yang dilaksanakan oleh Agus Fajar Riany dan kolega berhasil mencapai tingkat keberhasilan 92,48% dalam klasifikasi stroke menggunakan metode Naive Bayes, dengan hasil tersebut masuk dalam kategori Good Classification [7]. Sementara itu, penelitian Suhliyyah dan tim pada 2023 menerapkan algoritma Logistic Regression untuk klasifikasi penyakit dengan tingkat akurasi mencapai 94% [8]. Di tahun yang sama, Khairul Akmal beserta rekannya melakukan studi komparatif antara algoritma Naive Bayes KNN untuk klasifikasi stroke, di mana algoritma KNN menunjukkan performa superior dengan akurasi 94,36%, dibandingkan Naive Bayes yang hanya mencapai 90,10% [9]. Masih pada 2023, Ary Prandika Siregar mengimplementasikan algoritma Random Forest dalam klasifikasi stroke dan memperoleh akurasi 95%, dengan pembagian data 80% untuk pelatihan dan 20% untuk pengujian [10].

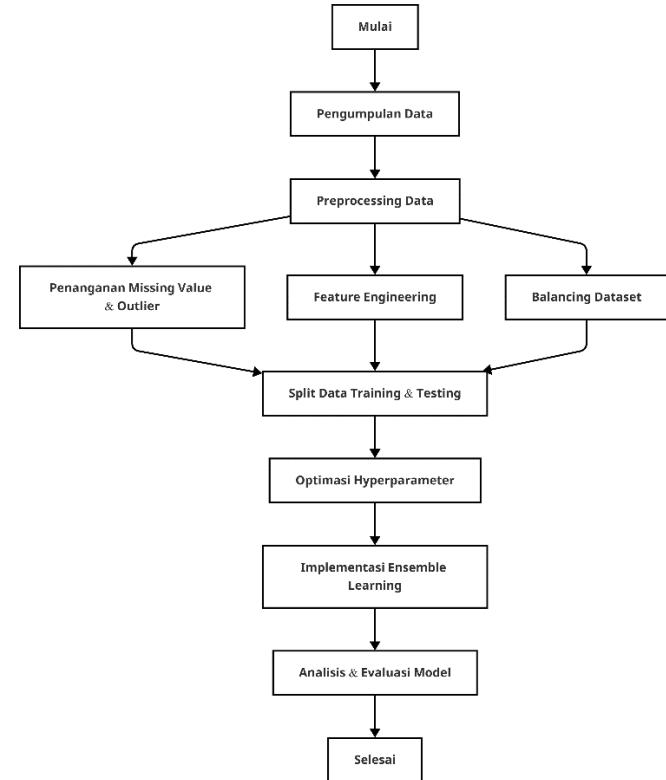
Di antara berbagai algoritma prediktif, penelitian ini akan memanfaatkan algoritma Extreme Gradient Boosting (XGBoost) [11]. Mengingat kompleksitas dan multivariatnya faktor risiko penyakit stroke, Algoritma XGBoost menjadi pilihan yang strategis untuk klasifikasi kondisi ini. XGBoost mampu mengolah data berdimensi tinggi dan menangani interaksi non-linear antar variabel, seperti hipertensi, kadar glukosa, dan BMI. Implementasi XGBoost dapat menghasilkan klasifikasi yang tidak hanya presisi, tetapi juga menyediakan wawasan komprehensif tentang faktor-faktor yang berkontribusi terhadap risiko stroke [12]. Penelitian terbaru oleh Danang Triantoro Murdiansyah pada 2024 dengan mengaplikasikan XGBoost untuk prediksi stroke berhasil mencapai precision 94,3%, recall 96,6%, F-Measure 95,4%, dan akurasi 95,4% [2]. Kemudian pada 2025, Vina

Wulandari dan tim melakukan studi klasifikasi risiko stroke dengan beberapa algoritma, dimana XGBoost mencapai akurasi 95,21% dengan pembagian data pelatihan dan pengujian sebesar 80:20 [3]. Penelitian ini tidak hanya menggunakan XGBoost tetapi juga mengintegrasikan algoritma random forest melalui pendekatan ensemble learning.

Meskipun sejumlah studi terdahulu telah mendemonstrasikan hasil yang menjanjikan, masih terbuka peluang untuk meningkatkan akurasi dan efisiensi model prediksi. Penelitian Danang Triantoro Murdiansyah berhasil mencapai tingkat akurasi 95,4%, namun masih belum melampaui performa algoritma Decision Tree C4.5 yang mencatatkan akurasi 99,07% berdasarkan studi yang dilaksanakan oleh Randi Estian Pambudim dan rekan pada tahun 2022 [13]. Dengan demikian, penelitian ini diarahkan untuk mengkaji ulang algoritma XGBoost dan Random Forest dengan pendekatan yang lebih optimal, khususnya dalam mengatasi ketidakseimbangan data serta meningkatkan akurasi dan seleksi fitur yang relevan dalam konteks klasifikasi stroke.

II. METODE PENELITIAN

Proses pengumpulan data, pra-pemrosesan, balancing, dan implementasi model pembelajaran kelompok disusun secara sistematis sebagai bagian dari metodologi penelitian ini. Setiap langkah dimaksudkan untuk meningkatkan keakuratan prediksi risiko stroke dan menghasilkan model yang dapat diandalkan dan berguna untuk membantu pengambilan keputusan di bidang kesehatan.



Gambar 1. Tahapan Penelitian

A. Tahapan Penelitian

Untuk mencapai hasil yang optimal dalam implementasi klasifikasi penyakit stroke menggunakan ensemble learning, diperlukan metodologi penelitian yang sistematis dan terstruktur. Metodologi ini dirancang dalam bentuk diagram alur yang menggambarkan tahapan penelitian secara berurutan, sehingga setiap proses dapat terorganisir dengan baik dan mendukung tercapainya tujuan penelitian secara efektif.

B. Pengumpulan Data

Dataset yang dimanfaatkan dalam penelitian ini identik dengan yang digunakan oleh peneliti sebelumnya, bersumber dari platform Kaggle (<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>). Dataset tersebut memuat informasi terkait berbagai faktor yang berkorelasi dengan kejadian stroke, mencakup usia, jenis kelamin, hipertensi, riwayat penyakit jantung, status pernikahan, lingkungan pekerjaan, lingkungan tempat tinggal, kadar glukosa rata-rata, massa tubuh (BMI), dan status merokok. Dataset ini juga dilengkapi dengan label target stroke yang mengindikasikan apakah individu tersebut pernah mengalami stroke atau tidak. Total sampel yang dianalisis dalam penelitian ini berjumlah 5110, terdiri dari 4861 sampel tanpa riwayat stroke dan 249 sampel dengan riwayat stroke. Jika dihitung dalam persentase 95% data merupakan kasus tanpa stroke dan 5% data merupakan data kasus yang terkena stroke. Ketidakseimbangan proporsi ini berpotensi menyebabkan bias dalam model, dimana model cenderung lebih akurat memprediksi kasus non-stroke, sehingga mengurangi sensitivitas dalam mendekripsi kasus stroke positif. Konsekuensinya, permasalahan ketidakseimbangan ini perlu diatasi melalui beberapa tahapan, termasuk preprocessing data, agar model dapat mengidentifikasi pola dengan tingkat akurasi yang lebih tinggi.

C. Preprocessing Data

Pra-pemrosesan data merupakan tahapan krusial dalam *machine learning* karena pengabaian tahap ini dapat mengakibatkan data tidak terstruktur, mengandung anomali, atau memiliki ketidaklengkapan yang berdampak negatif pada performa model dan menghasilkan prediksi dengan reliabilitas rendah [14]. Penelitian ini menggunakan dataset dengan distribusi yang tidak seimbang, sehingga implementasi pra-pemrosesan data dapat memitigasi efek overlap antar kelas dan pergeseran dataset dalam klasifikasi dengan ketidakseimbangan [15]. Dalam studi ini, kami mengimplementasikan beragam teknik pra-pemrosesan, mencakup penanganan nilai yang hilang, rekayasa fitur, dan penyeimbangan dataset. Rangkaian proses tersebut dilaksanakan untuk memastikan model mampu menghasilkan output yang akurat dan andal. Dengan demikian, pra-pemrosesan data menjadi landasan yang kokoh untuk analisis lanjutan dan pengembangan model yang efektif.

1) *Penanganan Missing Value dan Outlier*: Judul Fitur Body Mass Index (BMI) dalam dataset memiliki nilai hilang, sehingga digunakan metode Multiple Imputation by Chained Equations (MICE) dengan algoritma iterative imputer untuk mengisinya secara berulang sambil mempelajari pola antar fitur dalam dataset [16]. Proses ini diulang sebanyak 10 kali untuk meningkatkan akurasi estimasi dengan mempertimbangkan korelasi antar fitur. Sementara itu, outlier nilai yang menyimpang jauh dari distribusi umum dapat mengganggu proses pembelajaran model. Oleh karena itu, outlier pada fitur BMI dan avg_glucose_level diidentifikasi menggunakan metode Z-Score dan Interquartile Range (IQR). Z-Score digunakan untuk BMI yang terdistribusi normal, dengan data yang memiliki nilai Z di luar -3 hingga +3 dihapus [17]. Sedangkan untuk avg_glucose_level yang tidak terdistribusi normal, digunakan IQR, di mana data di luar rentang normal dianggap sebagai outlier dan dihapus [18]. Kombinasi metode ini dipilih agar model dapat mempelajari pola secara akurat tanpa terganggu oleh nilai ekstrem.

2) *Feature Engineering*: Dalam meningkatkan kinerja model dalam memprediksi risiko stroke, penelitian ini melakukan pengolahan data yang mencakup pengkodean fitur kategorikal, standarisasi fitur numerik, dan seleksi fitur. Fitur kategorikal seperti gender, dan smoking_status yang bertipe non-numerik dikonversi melalui metode One-Hot Encoding, yang mengubah kategori menjadi kolom biner (0 atau 1) [19]. Metode ini unggul dalam menghasilkan ruang fitur berdimensi tinggi, terutama dengan banyak kategori dalam dataset ini [20], dan untuk menghindari multikolinearitas, opsi drop='first' digunakan [21]. Proses pengkodean dilakukan menggunakan ColumnTransformer dari scikit-learn dengan remainder='passthrough' untuk mempertahankan fitur non-kategorikal. Setelah itu, Standardizing Numerical Features dilakukan untuk menyamakan skala fitur numerik, mengoptimalkan kinerja model, dan menghindari bias terhadap fitur dengan skala lebih besar. Dalam penelitian ini, StandardScaler digunakan untuk mengubah nilai fitur menjadi distribusi rata-rata, dengan fitur yang dinormalisasi mencakup age, avg_glucose_level, dan bmi. Langkah terakhir adalah seleksi fitur menggunakan SelectKBest dan ANOVA F-test untuk memilih 8 fitur terbaik yang paling berpengaruh terhadap prediksi stroke [22], sehingga model dapat mempelajari pola data dengan lebih efisien dan akurat.

3) *Balancing Dataset*: Dalam dataset yang digunakan, terdapat ketidakseimbangan kelas yang signifikan antara pasien yang mengalami stroke dan yang tidak, dengan sekitar 95% sampel adalah kelas negatif (tidak stroke) dan 5% kelas positif (stroke). Untuk mengatasi ini ketidakseimbangan ini, kami menggunakan metode SMOTE-ENN, yang melakukan oversampling pada kelas minoritas dan undersampling pada kelas mayoritas, membantu menyeimbangkan sampel antar kelas, mengurangi noise, dan mempermudah model dalam mempelajari pola secara lebih seimbang dan akurat. SMOTE bekerja dengan menghasilkan sampel sintesis untuk kelas

minoritas dengan mengelompokkan data berdasarkan tetangga terdekat [23]. Sedangkan ENN membersihkan data dengan menghapus sampel kelas mayoritas yang mengganggu pola antar kelas [24]. Kombinasi ini efektif untuk memperbaiki ketidakseimbangan kelas dalam dataset ini, khususnya mengingat bahwa algoritma XGBoost sangat sensitif terhadap ketidakseimbangan data, yang dapat menurunkan recall dan kemampuan model dalam mengidentifikasi pasien berisiko stroke. Dengan penggunaan SMOTE-ENN, diharapkan recall pada kelas positif dapat meningkat tanpa mengurangi akurasi dan presisi model secara keseluruhan.

D. Split Data Untuk Training dan Testing

Setelah proses preprocessing selesai, langkah berikutnya adalah memisahkan data menjadi 20% data uji dan 80% data latih menggunakan metode `train_test_split` dari library scikit-learn. Pemisahan ini dilakukan untuk memastikan bahwa model dapat mempelajari data latih dan menggunakan informasi ini pada data uji. Distribusi kelas tetap seimbang pada data uji dan latih dengan penggunaan parameter `stratify`. Stratify memungkinkan model mempelajari pola dengan adil, menghindari kecerundungan prediksi yang tidak diinginkan. Hanya setelah pelatihan selesai, data uji digunakan untuk mengevaluasi kinerja model. Hal itu dilakukan untuk mencegah data bocor, yang dapat menyebabkan hasil evaluasi tidak sah.

E. Optimasi Hyperparameter

Optimasi hyperparameter bertujuan meningkatkan kinerja model dengan menemukan kombinasi parameter terbaik untuk prediksi yang paling akurat. Dalam penelitian ini, hyperparameter yang dioptimasi meliputi `learning rate`, `max depth`, `n_estimators`, `gamma`, `subsample`, `colsample_bytree`, serta `scale_pos_weight` yang menangani ketidakseimbangan kelas dengan memberi bobot lebih besar pada kelas minoritas (stroke). Kami menggunakan *Bayesian Optimization*, metode berbasis probabilitas, yang memanfaatkan *Gaussian Process* untuk memprediksi kombinasi berikutnya berdasarkan hasil sebelumnya, berbeda dengan *grid* atau *random search*. Proses optimasi dilakukan menggunakan *BayesSearchCV* dari library *scikit-optimize* dengan ruang pencarian parameter yang telah ditentukan. Metode ini bekerja secara iteratif melalui eksplorasi dan eksloitasi. Selain itu, digunakan *StratifiedKFold* untuk menjaga distribusi kelas tetap seimbang dalam setiap fold, sehingga pembelajaran lebih konsisten dan mengurangi risiko *overfitting*.

F. Implementasi Ensemble Learning

Setelah proses optimasi hyperparametere selesai, Bayesian Optimization digunakan untuk menginisialisasi model dengan parameter terbaik yang telah di peroleh. Fungsi `XGBClassifier` dari library XGBoost digunakan untuk memulai model XGBoost dengan konfigurasi parameter terbaik tersebut. Setelah model diinisialisasi, proses pelatihan dimulai dengan menggunakan fungsi `fit()` dengan data pelatihan `X_train` dan `y_train` sebagai input. Selama pelatihan,

XGBoost akan membangun pohon keputusan secara bertahap. Setiap pohon akan berusaha memperbaiki kesalahan dari pohon sebelumnya dengan menggunakan metode peningkatan gradient untuk meminimalkan loss function. Proses ini diulang hingga jumlah maksimum pohon yang ditentukan oleh `n_estimators` tercapai atau hingga peningkatan kinerja model menjadi tidak signifikan. Untuk meningkatkan kemampuan prediksi, model Random Forest juga diinisialisasi sebagai bagian dari pendekatan ensemble bersama dengan model XGBoost. Ensemble Learning adalah teknik kombinasi algoritma Machine Learning yang bertujuan untuk memperoleh hasil akurasi prediksi yang lebih tinggi dibandingkan dengan algoritma yang digunakan secara tunggal.

Setelah proses pelatihan selesai, kami menggunakan *StratifiedKFold* kembali dilakukan untuk mengevaluasi performa model secara keseluruhan. *StratifiedKFold* memastikan bahwa pembagian data uji dan data latih pada setiap fold tetap mempertahankan distribusi kelas yang sama, sehingga hasil evaluasi menjadi lebih akurat dan konsisten. Proses validasi ini membantu untuk mengukur bagaimana model akan bekerja pada data yang belum pernah dilihat sebelumnya. Model Random Forest diinisialisasi menggunakan parameter `n_estimators` sebanyak 300 pohon dengan `random_state` untuk memastikan hasil yang konsisten. Setelah itu, kedua model digabungkan ke dalam *Voting Classifier* menggunakan metode *Voting* yang dikenal sebagai "soft". Metode ini menggabungkan hasil prediksi probabilistik dari kedua model utama dan membuat prediksi akhir berdasarkan probabilitas tertinggi. Tujuan penggabungan XGBoost dan Random Forest dalam *Voting Classifier* ini adalah untuk memanfaatkan keunggulan masing-masing model sehingga model dapat mempelajari pola dari kedua model dasar untuk membuat prediksi akhir yang lebih stabil dan akurat.

G. Analisis dan Evaluasi Model

Evaluasi model dilaksanakan untuk mengukur tingkat kehandalan model dalam memprediksi risiko stroke berdasarkan data pengujian. Untuk menghindari overfitting, model dilatih menggunakan strategi validasi silang *Stratified K-Fold* ($k=5$) serta hold-out test set sebesar 20%. Selain itu, dilakukan Bayesian Optimization untuk menyesuaikan hyperparameter model XGBoost, termasuk parameter `scale_pos_weight` yang disesuaikan dengan distribusi kelas. Evaluasi performa dilakukan pada data uji yang terpisah dari data pelatihan, sehingga hasil yang diperoleh merefleksikan kemampuan generalisasi model secara lebih objektif. Evaluasi awal memanfaatkan *classification report* yang menyajikan metrik-metrik esensial seperti precision, recall, F1-score dan akurasi. Model juga dianalisis menggunakan Receiver Operating Characteristic (ROC) Curve, yang memvisualisasikan relasi antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai threshold prediksi. Di samping itu, kinerja model juga dikaji dengan Confusion

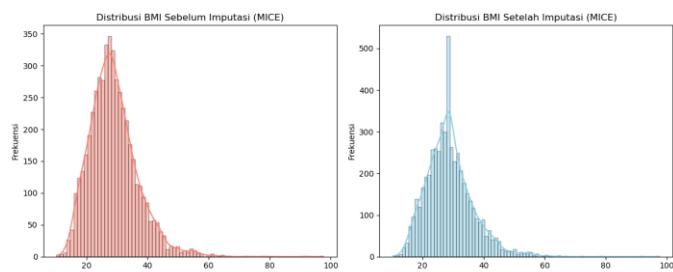
Matrix, yang memperlihatkan kuantitas prediksi akurat dan keliru.

III. HASIL DAN PEMBAHASAN

Menyajikan hasil penggunaan model dan juga membahas seberapa baik model mengklasifikasikan risiko stroke. Kinerjanya dinilai menggunakan berbagai metrik dan dibandingkan dengan penelitian sebelumnya untuk menunjukkan seberapa baik metode dalam penelitian ini.

A. Hasil Preprocessing

Pada subbab ini, disajikan hasil yang diperoleh dari penerapan model serta evaluasi kinerjanya dalam mengklasifikasikan risiko stroke. Selain itu, juga ditampilkan hasil dari setiap tahapan preprocessing yang dilakukan, sehingga analisis yang diberikan mencakup gambaran menyeluruh mengenai pengaruh preprocessing terhadap performa model, serta efektivitas dan kemampuan generalisasi model yang dikembangkan.



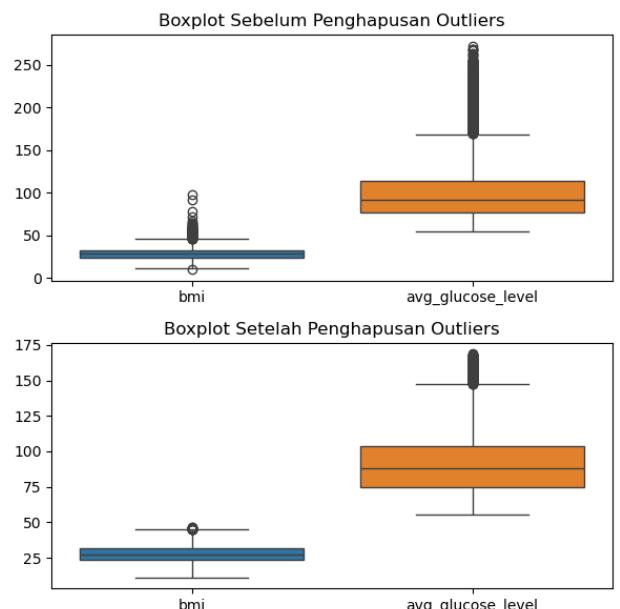
Gambar 2. Distribusi nilai BMI sebelum dan sesudah MICE

TABEL I
DESKRIPSI STATISTIK SETELAH IMPUTASI

Deskripsi Statistik BMI Setelah Imputasi (MICE):	
	BMI
count	5110,00
mean	28,89
std	7,69
min	10,30
25%	23,80
50%	28,40
75%	32,80
max	97,60

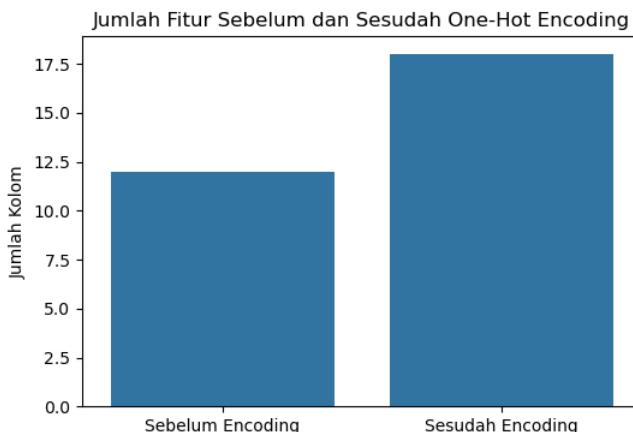
Gambar 2 menunjukkan distribusi nilai Body Mass Index (BMI) sebelum dan sesudah dilakukan proses imputasi menggunakan metode Multiple Imputation by Chained Equations (MICE). Pada distribusi sebelum imputasi, terlihat bahwa sebagian data BMI tidak tersedia, yang tercermin dari jumlah missing values sebanyak 201 data, hal ini mengakibatkan distribusi data yang kurang lengkap dan dapat memengaruhi kualitas analisis yang dilakukan. Setelah dilakukan imputasi menggunakan metode MICE, jumlah missing values berhasil dikurangi menjadi nol, artinya seluruh data yang hilang telah berhasil diisi, dengan hasil deskripsi

statistik setelah imputasi yang terdapat pada Tabel I menunjukkan bahwa nilai *count* pada variabel BMI mencapai 5110, yang sesuai dengan jumlah total data dalam dataset. Distribusi setelah imputasi menunjukkan pola yang lebih halus dan merata, meskipun tetap mempertahankan bentuk distribusi aslinya, sehingga hasil ini menunjukkan bahwa metode MICE efektif dalam menangani masalah data hilang tanpa mengubah karakteristik utama dari data.



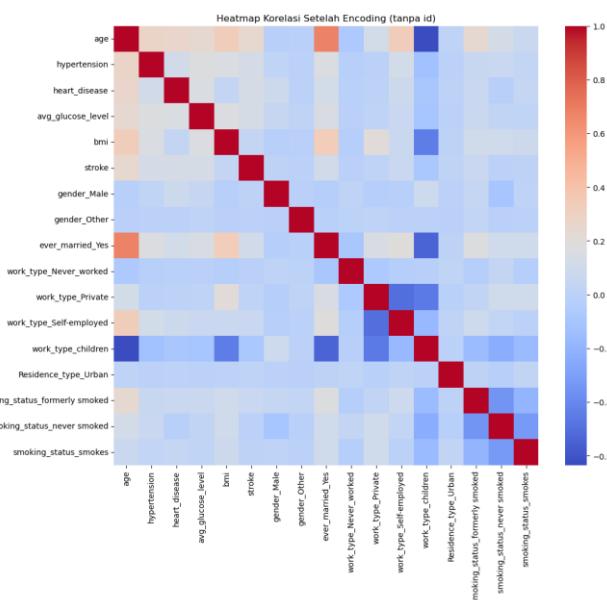
Gambar 3. Boxplot distribusi sebelum dan sesudah penghapusan outlier

Gambar 3 menunjukkan distribusi variabel BMI dan Average Glucose Level sebelum dan sesudah dilakukan penghapusan outlier. Sebelum penghapusan outlier, terlihat bahwa kedua variabel memiliki banyak titik ekstrem yang muncul di luar batas normal, khususnya pada nilai BMI di atas 50 dan average glucose level di atas 150 hingga mendekati 300. Keberadaan outlier ini dapat memengaruhi performa model machine learning dengan menyebabkan bias pada proses pembelajaran dan menurunkan akurasi prediksi. Setelah dilakukan penghapusan outlier menggunakan metode yang diimplementasikan dalam coding, distribusi data terlihat lebih bersih dan terpusat. Nilai BMI berkisar antara 15 hingga 45, sedangkan average glucose level berada pada rentang 50 hingga 150. Berdasarkan hasil ini, dapat disimpulkan bahwa tahap deteksi dan penghapusan outlier merupakan langkah penting untuk meningkatkan kualitas data dan performa model klasifikasi risiko stroke dengan metode ensemble XGBoost dan Random Forest.



Gambar 4. Jumlah fitur sebelum dan sesudah One-Hot Encoding

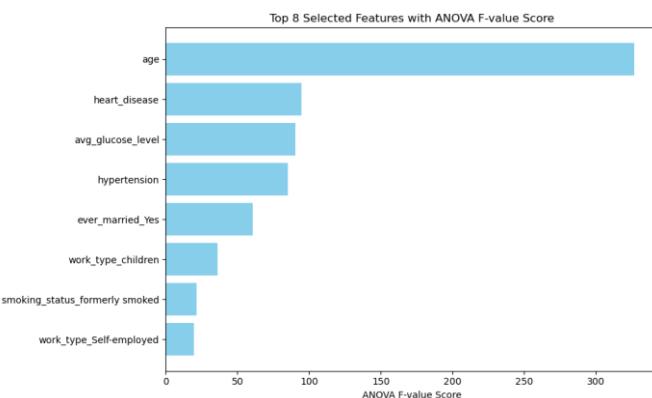
Gambar 4 menunjukkan perbandingan jumlah fitur sebelum dan sesudah diterapkannya teknik One-Hot Encoding pada data kategorikal. Sebelum dilakukan encoding, jumlah fitur dalam dataset adalah sebanyak 12 kolom, yang terdiri dari fitur numerik dan beberapa fitur kategorikal seperti gender, ever_married, work_type, Residence_type, dan smoking_status. Setelah diterapkannya One-Hot Encoding, jumlah fitur meningkat menjadi 18 kolom. Peningkatan ini terjadi karena setiap fitur kategorikal dikonversi menjadi sejumlah kolom biner (dummy variables) berdasarkan jumlah kategori unik yang dimilikinya. Proses ini dilakukan agar model pembelajaran mesin dapat memahami informasi kategorikal dalam bentuk numerik tanpa memperkenalkan urutan atau skala yang salah secara implisit.



Gambar 5. Heatmap korelasi antar fitur

Visualisasi heatmap korelasi pada Gambar 5 menunjukkan hubungan linier antar fitur setelah dilakukan encoding. Secara umum, tidak ditemukan korelasi yang sangat tinggi antar fitur, yang berarti multikolinearitas dalam dataset cukup

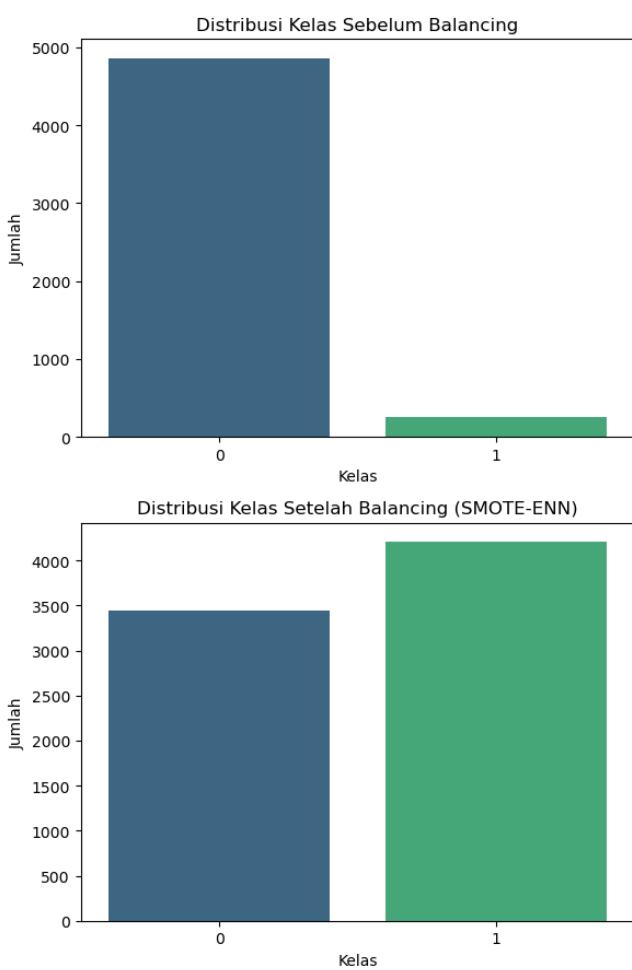
rendah. Beberapa hubungan korelasi yang menonjol antara lain adalah korelasi positif antara age dan ever_married_Yes, yang mencerminkan bahwa semakin bertambah usia seseorang, semakin besar kemungkinan mereka sudah menikah hal ini sesuai dengan pola umum dalam kehidupan sosial. Selain itu, terdapat korelasi negatif antara age dan work_type_children, yang menggambarkan bahwa semakin muda usia individu, semakin besar kemungkinan mereka termasuk dalam kategori anak-anak yang belum memasuki dunia kerja. Contoh lain adalah korelasi positif antara hypertension dan heart_disease, yang menonjol karena secara medis individu dengan tekanan darah tinggi memang memiliki risiko lebih besar terkena penyakit jantung. Adapun variabel target stroke menunjukkan korelasi yang lemah terhadap fitur lainnya, yang menandakan bahwa risiko stroke kemungkinan dipengaruhi oleh kombinasi kompleks beberapa faktor, bukan hanya bergantung pada satu variabel saja.



Gambar 6. Seleksi Fitur dengan ANOVA F-Test

Seleksi fitur dilakukan menggunakan metode ANOVA F-value untuk mengidentifikasi fitur-fitur yang paling berkontribusi terhadap prediksi risiko stroke. Hasil visualisasi pada Gambar 6 menunjukkan delapan fitur teratas yang dipilih berdasarkan skor F-value tertinggi. Fitur age memiliki skor dominan sebesar 326.92, menandakan bahwa usia merupakan faktor paling signifikan dalam memengaruhi risiko stroke. Disusul oleh heart_disease (94.70), avg_glucose_level (90.50), dan hypertension (84.95), yang semuanya merupakan faktor klinis yang relevan secara medis dalam kasus stroke.

Fitur ever_married_Yes dengan skor 60.67 menunjukkan bahwa status perkawinan juga memiliki korelasi terhadap risiko, kemungkinan karena berkaitan dengan umur atau stabilitas gaya hidup. Sementara itu, kategori pekerjaan work_type_children (36.18) dan work_type_Self-employed (19.82) serta status merokok smoking_status_formerly smoked (21.38) juga dipilih karena kontribusi informatifnya terhadap prediksi meskipun dengan skor yang lebih rendah. Fitur-fitur ini kemudian dipertahankan dalam tahap-tahap klasifikasi selanjutnya karena telah terbukti relevan secara statistik berdasarkan hasil seleksi fitur ini.



Gambar 7. Perbandingan Dlistribusi Kelas

Gambar 7 menunjukkan distribusi awal kelas pada dataset menunjukkan ketimpangan yang signifikan. Dari total 5110 data, sebanyak 4861 data (sekitar 95,13%) berasal dari kelas 0 (tidak terkena stroke), sedangkan hanya 249 data (sekitar 4,87%) berasal dari kelas 1 (terkena stroke). Ketidakseimbangan ini berisiko menyebabkan model lebih bias terhadap kelas mayoritas, sehingga performa dalam mendeteksi kasus stroke dapat menurun.

Untuk mengatasi hal tersebut, dilakukan balancing data menggunakan teknik SMOTE-ENN (*Synthetic Minority Over-sampling Technique – Edited Nearest Neighbours*). Setelah proses balancing, distribusi kelas berubah menjadi lebih seimbang dengan 3449 data pada kelas 0 (sekitar 45,06%) dan 4206 data pada kelas 1 (sekitar 54,94%) dari total 7655 data. Proporsi yang lebih seimbang ini diharapkan dapat meningkatkan performa model dalam mengenali pola dari kedua kelas secara adil, khususnya dalam mengidentifikasi kasus stroke yang sebelumnya kurang terwakili.

B. Evaluasi Model

Pada penelitian ini kami lakukan evaluasi menyeluruh terhadap performa model klasifikasi yang dikembangkan dalam mendeteksi risiko stroke, menggunakan berbagai metrik seperti precision, recall, f1-score, akurasi, serta analisis tambahan berupa classification report, ROC curve, confusion matrix, dan feature importance. Evaluasi akhir dilakukan menggunakan test set (20% dari total data) yang tidak digunakan dalam proses pelatihan, sehingga dapat memberikan gambaran yang lebih objektif mengenai kekuatan dan kelemahan model dalam memprediksi risiko stroke.

TABEL II
CLASSIFICATION REPORT ENSEMBLE

	precison	recall	f1-score	support
0	0,9827	0,9870	0,9848	690
1	0,9893	0,9857	0,9875	841
accuracy			0,9863	1531
macro avg	0,9860	0,9863	0,9862	1531
weighted avg	0,9863	0,9863	0,9863	1531

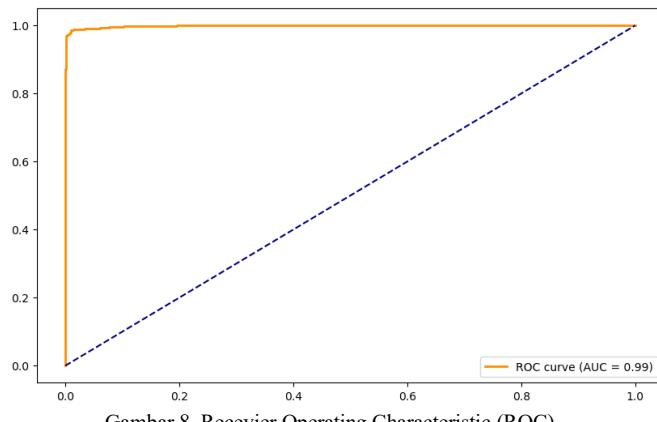
Hasil classification report mengindikasikan bahwa model klasifikasi menunjukkan kinerja yang sangat baik dalam membedakan kedua kategori. Evaluasi ini dilakukan menggunakan data uji (test set) yang tidak digunakan dalam proses pelatihan maupun tuning, sehingga mencerminkan kemampuan generalisasi model terhadap data baru (unseen data). Untuk kategori negatif (tidak stroke), model memperoleh nilai precision sebesar 0,9827, recall 0,9870, dan f1-score 0,9848, dengan jumlah sampel (support) sebanyak 690. Sementara itu, untuk kategori positif (stroke), model mencapai precision 0,9893, recall 0,9857, dan f1-score 0,9875, dengan support sebanyak 841. Akurasi keseluruhan model mencapai 0,9863 atau sekitar 98,6%, menunjukkan bahwa hampir seluruh prediksi dilakukan dengan tepat. Selain itu, nilai macro average dan weighted average yang juga berada di kisaran 0,9860 - 0,9863 menunjukkan performa yang konsisten di seluruh kelas, memperkuat bukti bahwa model memiliki kemampuan klasifikasi yang seimbang dan andal.

TABEL III
PERBANDINGAN HASIL INDIVIDUAL DENGAN ENSEMBLE

Model	precision		recall		f1-score		Akuras i (%)
	0	1	0	1	0	1	
XGBoost	0,98 27	0,99 16	0,98 99	0,98 57	0,98 63	0,98 87	0,9876
Random Forest	0,98 24	0,97 64	0,97 10	0,98 57	0,97 67	0,98 11	0,9791
Ensemble (XGBoost + Random Forest)	0,98 27	0,98 93	0,98 70	0,98 57	0,98 48	0,98 75	0,9863

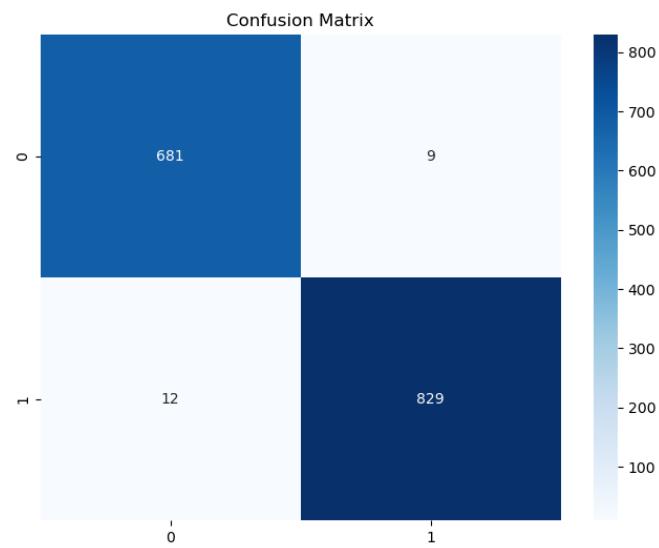
Tabel III hasil evaluasi, model XGBoost menunjukkan performa yang sangat baik dengan precision 0,9827 untuk kelas 0 (tidak stroke) dan 0,9916 untuk kelas 1 (stroke), serta f1-score rata-rata sebesar 0,9887 dan akurasi sebesar 98,76%. Model Random Forest juga memberikan hasil yang baik, meskipun sedikit lebih rendah, dengan precision 0,9824 untuk kelas 0 dan 0,9764 untuk kelas 1, serta f1-score rata-rata 0,9811 dan akurasi 97,91%. Sementara itu, model Ensemble yang dibangun menggunakan metode Voting Classifier dengan menggabungkan XGBoost dan Random Forest menunjukkan performa kompetitif, dengan precision 0,9827 untuk kelas 0 dan 0,9893 untuk kelas 1, serta f1-score rata-rata 0,9875 dan akurasi 98,63%.

Hasil ini menunjukkan bahwa penggunaan Voting Classifier sebagai metode ensemble dapat menggabungkan kekuatan kedua model untuk menghasilkan prediksi yang lebih stabil dan akurat, meskipun peningkatannya tidak selalu signifikan dibandingkan dengan model XGBoost tunggal. Keunggulan metode ensemble ini terlihat dari keseimbangan performa antar kelas dan konsistensi skor evaluasi, di mana precision, recall, dan f1-score pada kedua kelas saling mendekati, menunjukkan model mampu mendeteksi baik kasus stroke maupun non-stroke secara seimbang. Selain itu, meskipun akurasi ensemble sedikit di bawah XGBoost, Voting Classifier tetap memberikan keuntungan dari segi robustness karena dapat mengurangi bias atau kelemahan model tunggal, meningkatkan generalisasi terhadap data baru, serta membuat prediksi yang lebih tahan terhadap fluktuasi data.



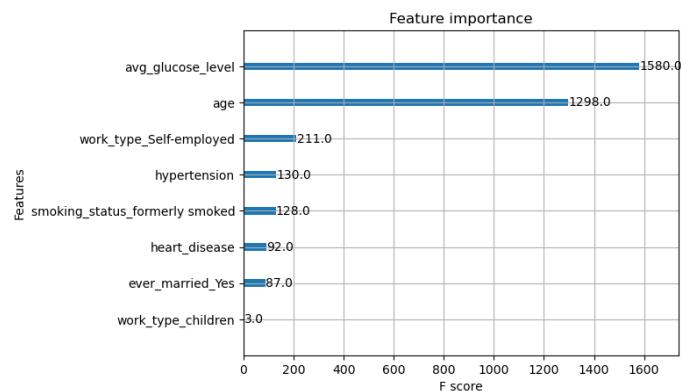
Gambar 8. Recevier Operating Characteristic (ROC)

Melalui analisis ROC curve yang menggambarkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai ambang batas probabilitas, diperoleh nilai Area Under Curve (AUC) sebesar 0,99. Nilai AUC yang sangat tinggi ini menunjukkan bahwa model memiliki kapabilitas klasifikasi yang sangat baik dalam membedakan antara pasien dengan riwayat stroke dan pasien tanpa riwayat stroke. Kurva ROC yang terletak jauh di atas garis diagonal menandakan bahwa model tidak melakukan klasifikasi secara acak, melainkan memiliki kemampuan diskriminatif yang kuat.



Gambar 9. Confusion Matrix

Berdasarkan hasil evaluasi performa model melalui confusion matrix yang tertera pada gambar 9, diketahui bahwa model mampu mengklasifikasikan data dengan tingkat akurasi yang sangat tinggi. Model berhasil mengidentifikasi 681 data sebagai *true negative* dan 829 data sebagai *true positive*, serta hanya menghasilkan 9 kesalahan klasifikasi pada kelas negatif (*false positive*) dan 12 pada kelas positif (*false negative*). Tingkat sensitivitas (recall) untuk kelas positif mencapai 98,57%, yang menunjukkan bahwa model hampir sepenuhnya mampu mengenali kasus stroke. Di sisi lain, spesifitas model terhadap kelas negatif juga sangat baik, yaitu sebesar 98,69%, yang mencerminkan kemampuannya dalam menghindari kesalahan deteksi terhadap kasus non-stroke. Keseimbangan yang tinggi antara sensitivitas dan spesifitas ini menunjukkan bahwa model yang dikembangkan memiliki performa prediktif yang kuat, sehingga dinilai efektif untuk diterapkan dalam sistem pendukung keputusan medis dalam mendeteksi risiko stroke secara otomatis.



Gambar 10. Hasil Visualisasi Feature Importance

Berdasarkan hasil visualisasi *feature importance* dari model XGBoost, diperoleh informasi bahwa fitur yang memiliki kontribusi terbesar dalam proses prediksi risiko stroke adalah avg_glucose_level dengan skor F tertinggi sebesar 1580, diikuti oleh age dengan skor sebesar 1298. Kedua fitur ini menunjukkan korelasi yang sangat kuat terhadap kemungkinan terjadinya stroke dan menjadi indikator utama dalam klasifikasi model. Fitur lain yang juga berpengaruh namun dengan kontribusi lebih rendah meliputi work_type_Self-employed (211), hypertension (130), dan smoking_status_formerly smoked (128). Adapun fitur heart_disease, ever_married_Yes, dan work_type_children memberikan pengaruh yang relatif kecil dengan skor masing-masing sebesar 92, 87, dan 3. Hasil ini mengindikasikan bahwa faktor fisiologis dan riwayat medis individu, seperti kadar glukosa dan usia, memiliki peran dominan dalam membentuk keputusan model, sedangkan karakteristik demografis dan status pekerjaan memberikan kontribusi yang lebih minor terhadap hasil prediksi.

C. Perbandingan Dengan Penelitian Sebelumnya

Dilakukan perbandingan terhadap beberapa penelitian terdahulu yang juga memanfaatkan algoritma XGBoost dan Random Forest dalam mengklasifikasi risiko stroke untuk menilai keunggulan dari pendekatan penelitian ini. Perbandingan dilakukan berdasarkan model yang digunakan, teknik yang diterapkan dalam proses optimasi atau pemrosesan data, serta nilai akurasi yang dihasilkan.

TABEL IV
PERBANDINGAN HASIL PENELITIAN SEBELUMNYA

Peneliti	Model Yang Digunakan	Metode Yang Digunakan	Akurasi (%)
Danang Triantoro Murdiansyah (2024) [2]	XGBoost	Bayesian Optimization + SMOTE	95,4%
Susi Handayani (2024) [4]	XGBoost	SMOTE	96,94%
Ary Prandika Siregar (2023) [10]	Random Forest	-	95,0%
Penelitian ini	XGBoost + Random Forest	Bayesian Optimization	98,63%

Tabel IV, dapat diambil kesimpulan bahwa model yang dikembangkan dalam studi ini mencapai tingkat akurasi tertinggi dibandingkan dengan penelitian-penelitian terdahulu. Dengan tingkat akurasi 98% maka model yang diimplementasikan mampu melakukan klasifikasi dengan tingkat presisi yang sangat tinggi. Walaupun model-model pada penelitian sebelumnya juga menunjukkan keberhasilan, hasil dari penelitian ini memperlihatkan peningkatan akurasi yang substansial. Hasil dari penelitian ini mengonfirmasi bahwa model yang diajukan berpotensi untuk diaplikasikan

untuk membantu kebutuhan klinis, khususnya untuk memfasilitasi deteksi lebih awal penyakit stroke.

IV. KESIMPULAN

Berdasarkan hasil analisis penelitian ini menunjukkan keberhasilan dalam membangun model klasifikasi risiko stroke yang akurat dan andal dengan menggabungkan algoritma XGBoost dan Random Forest dalam pendekatan ensemble learning. Model dapat mengatasi masalah utama dengan data medis yang tidak seimbang dan kompleks menggunakan metode seperti imputasi missing value menggunakan MICE, penanganan outlier, encoding fitur kategorikal, normalisasi fitur numerik, seleksi fitur relevan dengan ANOVA F-test, dan penyeimbangan data menggunakan metode SMOTE-ENN. Hasil evaluasi menunjukkan kinerja yang sangat baik dengan akurasi mencapai 98%, precision dan recall mencapai 98%, dan nilai AUC sangat baik sebesar 0,99. Pencapaian ini menunjukkan kemampuan klasifikasi yang unggul dan kestabilan model untuk menemukan pola data yang berkaitan dengan risiko stroke. Metode penelitian ini terbukti lebih efektif dan akurat dibandingkan dengan penelitian sebelumnya. Oleh karena itu, model yang dibuat tidak hanya memenuhi tujuan penelitian, tetapi juga membantu dalam pengembangan sistem pendukung keputusan medis yang dapat diandalkan berbasis machine learning untuk deteksi dini risiko stroke. Salah satu keterbatasan penelitian ini adalah bahwa penelitian ini hanya menggunakan dataset publik Kaggle tanpa melakukan validasi terhadap dataset lain. Di penelitian selanjutnya, sangat disarankan untuk melakukan evaluasi terhadap dataset lain dari rumah sakit lokal atau regional untuk menguji generalisasi dan kekuatan model pada berbagai populasi.

DAFTAR PUSTAKA

- [1] World Stroke Organization, "WSO Global Stroke Fact Sheet 2022," 2022. Accessed: Feb. 27, 2025. [Online]. Available: https://www.world-stroke.org/assets/downloads/WSO_Global_Stroke_Fact_Sheet.pdf
- [2] D. T. Murdiansyah, "Prediksi Stroke Menggunakan Extreme Gradient Boosting," *JIKO (Jurnal Informatika dan Komputer)*, vol. 8, no. 2, p. 419, Sep. 2024, doi: 10.26798/jiko.v8i2.1295.
- [3] Vi. Wulandari, Mustakim, R. Novita, and N. E. Rozanda, "Implementation of Machine Learning Algorithm for Stroke Risk Classification by Applying Sequential Forward Selection," 2025.
- [4] S. Handayani, Fajrizal, Taslim, D. Toresa, and Syahril, "Peningkatan Performa Model Gradient Boosting dalam Klasifikasi Stroke Melalui Optimasi Grid Search," *JURNAL FASILKOM*, vol. 14, no. 3, pp. 722–728, Dec. 2024, doi: 10.37859/jf.v14i3.7893.
- [5] M. Issaiy, D. Zarei, S. Kolahi, and D. S. Liebeskind, "Machine learning and deep learning algorithms in stroke medicine: a systematic review of hemorrhagic transformation prediction models," *J Neurol*, vol. 272, no. 1, p. 37, Jan. 2025, doi: 10.1007/s00415-024-12810-6.
- [6] Y. Aulia, A. Andriyansyah, S. Suharjito, and S. W. Nensi, "Analisis Prediksi Stroke dengan Membandingkan Tiga Metode Klasifikasi Decision Tree, Naive Bayes, dan Random Forest," *Jurnal Ilmu Komputer dan Informatika*, vol. 3, no. 2, pp. 89–98, Jan. 2024, doi: 10.54082/jiki.90.
- [7] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve

- [8] Bayes," *Jurnal SAINTEKOM*, vol. 13, no. 1, pp. 42–54, Mar. 2023, doi: 10.33020/saintekom.v13i1.352.
- [9] S. Suhliyyah, H. Hikmayanti Handayani, and K. Ahmad Baihaqi, "Implementasi Algoritma Logistic Regression Untuk Klasifikasi Penyakit Stroke," *Syntax : Jurnal Informatika*, vol. 12, no. 01, pp. 15–23, May 2023, doi: 10.35706/syji.v12i01.8329.
- [10] K. Akmal, A. Faqih, and F. Dikananda, "Perbandingan Metode Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Penyakit Stroke," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 470–477, Mar. 2023, doi: 10.36040/jati.v7i1.6367.
- [11] Ary Prandika Siregar, Dwi Priyadi Purba, Jojor Putri Pasaribu, and Khairul Reza Bakara, "Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke," *Jurnal Penelitian Rumpun Ilmu Teknik*, vol. 2, no. 4, pp. 155–164, Nov. 2023, doi: 10.55606/juprit.v2i4.3039.
- [12] N. Nuraeni, "Klasifikasi Data Mining Untuk Prediksi Penyakit Kardiovaskular," *Jurnal TEKINKOM*, vol. 7, no. 1, 2024, doi: 10.37600/tekinkom.v7i1.1276.
- [13] R. Chen *et al.*, "A study on predicting the length of hospital stay for Chinese patients with ischemic stroke based on the XGBoost algorithm," *BMC Med Inform Decis Mak*, vol. 23, no. 1, p. 49, Mar. 2023, doi: 10.1186/s12911-023-02140-4.
- [14] R. Estian Pambudi, Sriyanto, and Firmansyah, "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C.45," *Jurnal TEKNIKA*, vol. x, No.x, pp. 1–5, Aug. 2022, doi: <https://doi.org/10.5281/zenodo.7535865>.
- [15] M. Hasanudin, S. Dwiasnati, and W. Gunawan, "Pelatihan DataScience pada Pra-Pemrosesan Data untuk Siswa SMK Media Informatika - Jakarta," *Jurnal Pengabdian Pada Masyarakat*, vol. 9, no. 4, pp. 882–888, Nov. 2024, doi: 10.30653/jppm.v9i4.921.
- [16] B. Nugroho and A. Denih, "Perbandingan Kinerja Metode Pra-Pemrosesan Dalam Pengklasifikasian Otomatis Dokumen Paten," *Komputasi: Jurnal Ilmiah Ilmu Komputer dan Matematika*, vol. 17, no. 2, pp. 381–387, Jul. 2020, doi: 10.33751/komputasi.v17i2.2148.
- [17] I. M. Karo Karo and H. Hendriyana, "Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score," *Jurnal Teknologi Terpadu*, vol. 8, no. 2, pp. 94–99, Dec. 2022, doi: 10.54914/jtt.v8i2.564.
- [18] A. Alabrah, "An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method," *Sensors*, vol. 23, no. 9, p. 4406, Apr. 2023, doi: 10.3390/s23094406.
- [19] Z. R. Fadilah and A. W. Wijayanto, "Perbandingan Metode Klasterisasi Data Berlite Campuran: One-Hot-Encoding, Gower Distance, dan K-Prototype Berdasarkan Akurasi (Studi Kasus: Chronic Kidney Disease Dataset)," *Journal of Applied Informatics and Computing*, vol. 7, no. 1, pp. 57–67, Jul. 2023, doi: 10.30871/jaic.v7i1.5857.
- [20] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technologia : Jurnal Ilmiah*, vol. 15, no. 1, p. 93, Jan. 2024, doi: 10.31602/jti.v15i1.13457.
- [21] P. Sanyal and S. K. Dalui, "Computational fluid dynamics and artificial neural network-based analysis and forecasting of wind effects on obliquely parallel multiple building models using categorical variable encoding," *The Structural Design of Tall and Special Buildings*, vol. 33, no. 8, Jun. 2024, doi: 10.1002/tal.2105.
- [22] S. Abdumalikov, J. Kim, and Y. Yoon, "Performance Analysis and Improvement of Machine Learning with Various Feature Selection Methods for EEG-Based Emotion Classification," *Applied Sciences*, vol. 14, no. 22, p. 10511, Nov. 2024, doi: 10.3390/app142210511.
- [23] S. F. N. Halim and U. Azmi, "Analisis Perbandingan Klasifikasi dan Penerapan Teknik SMOTE Dalam Imbalanced Data Pada Credit Card Default," *Jurnal Sains dan Seni ITS*, vol. 12, no. 2, May 2023, doi: 10.12962/j23373520.v12i2.111833.
- [24] I Gede Harsemadi, I Komang Dharmendra, and I Made Pasek Pradnyana Wijaya, "Klasifikasi Emosi Pada Tweet Berbahasa Indonesia Menggunakan Teknik Sampling ENN," *Jurnal Teknologi Informasi dan Komputer*, vol. 9, no. 5, Oct. 2023, doi: 10.36002/jutik.v9i5.2646.