JUKOMIKA (Jurnal Ilmu Komputer dan Informatika)
Volume 8, Nomor 2, Desember 2025 || Hal 118-124
***p-ISSN* : 2655-7541 ; e-ISSN : 2745-5823**
DOI : 10.54650

Penerapan *Random Forest* untuk Klasifikasi Risiko Penyakit *Stroke* Pada Rentang Usia 40-85 Tahun

Adjie Wahyudi[1], Muhammad Aimar Al Baihaqi[2], Ivan Variz Febrinanda[3], Adam Surya Dharma[4], Danurtirto Satria Prananda[5].
Universitas Bina Sarana Informatika, Sistem Informasi, Jakarta Timur, Indonesia.
email: adjiewahyudi0@gmail.com, muhammadaimaralbaihaqi@gmail.com, adamsuryadharma22012004@gmail.com, ivanalvarizjr@gmail.com, satriaprananda09@gmail.com

Abstrak—Stroke merupakan penyebab kematian dan kecacatan tertinggi ketiga di dunia dengan risiko yang meningkat tajam pada usia 40 tahun ke atas. Sebagian besar penelitian sebelumnya pada *dataset Stroke Prediction (*Kaggle) melaporkan akurasi tinggi namun tidak membahas dampak ketidakseimbangan kelas yang sangat ekstrem (rasio ~1:19) dan jarang memfokuskan analisis pada rentang usia risiko tinggi. Penelitian ini menerapkan algoritma *Random Forest* untuk klasifikasi risiko *stroke* khusus pada individu berusia 40–85 tahun menggunakan *dataset* sebanyak 2.875 data setelah filtering usia. Dataset memiliki ketidakseimbangan kelas tinggi (*stroke* 8.42%, *non-stroke* 91.58%). Tahapan *preprocessing* meliputi pemeriksaan *missing value* dan *duplicate data* (tidak ditemukan) serta standarisasi fitur *numerik* menggunakan *Z-Score Standardization* (*mean* = 0, *std* = 1). Model dievaluasi dengan 10-Fold Stratified Cross Validation pada perangkat lunak Orange Data Mining. Hasil menunjukkan Random Forest mencapai akurasi 96,5%, AUC 0,914 (kelas *stroke*), *precision* 96,3%, *recall* 96,5%, F1-*score* 96,2%, dan MCC 0,750. Analisis *feature importance* mengidentifikasi usia, kadar glukosa rata-rata, dan BMI sebagai tiga prediktor terkuat. Dibandingkan dengan *Logistic Regression*, *Decision Tree*, *SVM*, dan *Naive Bayes*, *Random Forest* menunjukkan performa paling unggul. Penelitian ini memberikan kontribusi berupa model klasifikasi yang andal dan *interpretabel* pada populasi usia 40–85 tahun serta rekomendasi klinis berbasis variabel paling berpengaruh.

Abstract— *Stroke is the third leading cause of death and disability worldwide, with risk increasing sharply after the age of 40. Most previous studies using the Stroke Prediction Dataset (Kaggle) reported high accuracy but rarely addressed the extreme class imbalance (ratio ~1:19) and seldom focused analysis on the high-risk age range. This study applies the Random Forest algorithm to classify stroke risk specifically in individuals aged 40–85 years using 40,910 records after age filtering. The dataset exhibits severe class imbalance (stroke 8.42%, non-stroke 91.58%). Preprocessing included checking for missing values and duplicates (none found) and standardization of numerical features using Z-Score Standardization (mean = 0, std = 1). The model was evaluated using 10-Fold Stratified Cross Validation in Orange Data Mining. Results show Random Forest achieved an accuracy of 96.5%, AUC 0.914 (stroke class), precision 96.3%, recall 96.5%, F1-score 96.2%, and MCC 0.750. Feature importance analysis identified age, average glucose level, and BMI as the top three predictors. Compared to Logistic Regression, Decision Tree, SVM, and Naive Bayes, Random Forest demonstrated superior performance across all metrics. This study contributes a reliable and interpretable classification model for the 40–85-year age group along with clinical recommendations based on the most influential variables.*

Keywords: *Random Forest, Stroke, Age Range 40–85 Years, Class Imbalance, Z-Score Standardization, Feature Importance*

1. Pendahuluan

Stroke merupakan salah satu penyakit paling berbahaya dan menduduki peringkat ketiga sebagai penyebab kematian tertinggi di dunia setelah penyakit jantung koroner dan kanker. Menurut data *World Stroke Organization*, setiap tahun tercatat sekitar 13,7 juta kasus *stroke* baru di seluruh dunia, dengan 5,5 juta di antaranya berujung pada kematian [1]. Stroke terjadi akibat terhambatnya atau terputusnya aliran darah ke sebagian area otak, sehingga sering disebut juga sebagai serangan otak. Salah satu penyebab utama dari kondisi ini adalah adanya gumpalan darah (bekuan darah) yang menyumbat pembuluh darah di otak [2]**.** Di Indonesia, *stroke* merupakan salah satu penyebab utama kematian serta kecacatan permanen. Pada tahun 2014 tercatat 41.590 orang meninggal akibat stroke, dan angka ini terus meningkat hingga sekarang [3]. Risiko *stroke* meningkat secara *eksponensial* setelah usia 40 tahun dan mencapai puncaknya pada kelompok usia lanjut, sehingga kelompok usia 40–85 tahun merupakan populasi dengan risiko tertinggi yang memerlukan strategi skrining yang lebih intensif.
Perkembangan teknologi pembelajaran mesin (*machine learning*) membuka peluang besar untuk deteksi dini risiko *stroke* melalui pengolahan data klinis dan gaya hidup. *Dataset Stroke Prediction* yang tersedia di *Kaggle* telah banyak digunakan dalam berbagai penelitian klasifikasi *stroke*. Namun, sebagian besar penelitian sebelumnya memiliki beberapa keterbatasan signifikan, yaitu tidak melakukan *filtering* usia

secara spesifik, sehingga model dilatih pada populasi umum termasuk anak-anak dan remaja yang risikonya sangat rendah.

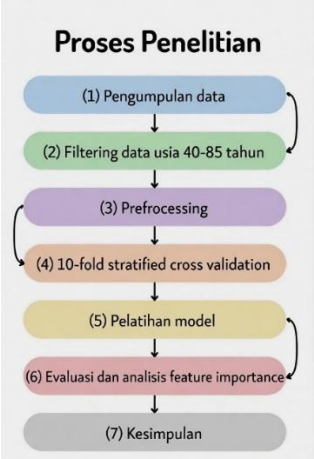
Tidak membahas atau menangani dampak ketidakseimbangan kelas yang sangat ekstrem (rasio *stroke : non-stroke* mencapai 1:19 hingga 1:60), sehingga nilai akurasi tinggi (>95%) sering kali menyesatkan karena hanya mencerminkan kemampuan memprediksi kelas mayoritas—misalnya, [2]. menggunakan *SMOTE* untuk *oversampling* pada *dataset* serupa dan mencapai AUC 0.836, tetapi pendekatan ini mengubah distribusi data asli, sementara [4] juga menerapkan *SMOTE* dengan RF mencapai akurasi 0.91 tanpa fokus pada kondisi nyata tanpa *resampling*. Jarang menyajikan analisis *feature importance* yang dapat langsung dimanfaatkan oleh tenaga medis untuk prioritas *intervensi*—seperti [5]. yang mengembangkan RF untuk prediksi *stroke* berdasarkan demografi tapi tidak menekankan implikasi klinis *feature* seperti usia atau *glucose*.

Minim melakukan perbandingan performa dengan algoritma *baseline* lainnya—walaupun [6]. dalam perbandingan mereka menunjukkan RF unggul dengan akurasi 95% dibanding *Naïve Bayes* pada *dataset Kaggle*, tapi banyak studi tidak bandingkan dengan *baseline* seperti *Logistic Regression* atau *SVM*, dan [7]. hanya fokus pada RF dengan *confusion matrix* baik tanpa perbandingan mendalam. Penelitian serupa oleh [8]. juga telah berhasil menerapkan *Random Forest* pada data klinis *stroke* yang sangat tidak seimbang dengan mencapai akurasi 93,6%, presisi 91,4%, *recall* 96,1%, dan F1-*Score* 93,7% melalui *stratified cross-validation* setelah *oversampling* menggunakan *SMOTE*, namun tidak secara khusus memfokuskan analisis pada rentang usia berisiko tinggi 40–85 tahun, sehingga masih menggunakan seluruh rentang usia dalam *dataset* dan mengubah distribusi data asli untuk menangani *imbalance*.

Random Forest merupakan algoritma *machine learning* yang digunakan untuk tugas klasifikasi dan regresi, di mana model ini terdiri dari kumpulan banyak pohon keputusan (*decision trees*) yang masing-masing berperan sebagai base c*rlassifier* [9]. Algoritma *Random Forest* memiliki beberapa keunggulan utama, yaitu mampu menangani *dataset* berukuran besar dengan jumlah fitur yang banyak, lebih tahan terhadap *overfitting* dibandingkan satu pohon keputusan tunggal, serta secara otomatis menghasilkan nilai pentingnya (*feature importance*) setiap fitur [10]. Kelebihan *Random Forest* pada data medis yang tidak seimbang telah dibuktikan dalam beberapa studi, namun belum banyak yang memfokuskan pada rentang usia risiko tinggi 40–85 tahun. Penelitian ini hadir untuk mengisi celah-celah tersebut dengan fokus spesifik pada individu berusia 40–85 tahun menggunakan *dataset Stroke Prediction* (*Kaggle*) yang telah difilter usia. Kebaruan (*novelty*) penelitian ini terletak pada analisis hanya pada rentang usia 40–85 tahun sebagai kelompok risiko tinggi. Evaluasi performa *Random Forest* pada kondisi ketidakseimbangan kelas nyata (tanpa teknik *oversampling/undersampling*), penyajian *feature importance* sebagai rekomendasi klinis prioritas *intervensi*, perbandingan performa dengan empat algoritma *baseline* (*Logistic Regression, Decision Tree, SVM, Naive Bayes*) menggunakan 10-*Fold Stratified Cross Validation*.

Tujuan penelitian ini adalah membangun dan mengevaluasi model klasifikasi risiko *stroke* menggunakan algoritma *Random Forest* pada populasi usia 40–85 tahun, menganalisis kontribusi masing-masing variabel prediktor melalui *feature importance*, serta membandingkan performa dengan algoritma klasifikasi lainnya. Hasil penelitian diharapkan dapat menjadi dasar pengembangan sistem pendukung keputusan klinis untuk skrining dini risiko stroke pada layanan kesehatan *primer* di Indonesia.

2. Metodologi Penelitian



Gambar 2.1 *Flowchart* Penelitian

Flowchart ini menggambarkan alur kerja penelitian secara keseluruhan untuk membangun dan mengevaluasi model prediksi risiko *stroke* menggunakan algoritma *machine learning*. Proses dimulai dari pengumpulan data mentah, dilanjutkan dengan pembersihan dan penyaringan data (khususnya pasien berusia 40–85 tahun), *preprocessing*, penerapan teknik validasi silang berstrata sebanyak 10-*fold*, pelatihan model, hingga tahap evaluasi performa model sekaligus analisis tingkat kepentingan setiap fitur, dan diakhiri dengan penyusunan kesimpulan penelitian. Alur ini menunjukkan pendekatan yang sistematis, *reproducible*, dan berfokus pada pencegahan *overfitting* melalui *stratified cross-validation* serta pemahaman *interpretabilitas* model lewat *feature importance*.

Penelitian ini dilaksanakan melalui tujuh tahapan utama yang ditunjukkan pada Gambar 2.1, yaitu pengumpulan data *Dataset Stroke Prediction* diunduh dari *Kaggle* [11] yang semula berisi 4.891 baris data dengan 12 atribut (sebelas) atribut klinis dan gaya hidup, termasuk variabel target "*stroke*". Filtering Data Usia 40–85 Tahun, dilakukan penyaringan ketat sehingga hanya menyisakan individu berusia 40 hingga 85 tahun, menghasilkan 2.875 baris data dengan distribusi kelas sangat tidak seimbang (*stroke* = 242 kasus atau 8.42%; *non-stroke* = 2633 kasus atau 91.58%), rasio 1:19. *Preprocessing* tahap ini mencakup tiga proses utama, pemeriksaan dan penanganan *missing value* (tidak ditemukan) nilai hilang pada semua atribut, pemeriksaan dan penghapusan *duplicate data* (tidak ditemukan) tidak terdeteksi baris duplikat, standarisasi tiga fitur numerik (*age, avg_glucose_level, bmi*) menggunakan *Z-Score Standardization* sehingga setiap fitur memiliki *mean* = 0 *dan standard deviation* = 1, tanpa mengubah bentuk distribusi asli.

10-*Fold Stratified Cross Validati*on, data dibagi menjadi 10 lipatan (*fold*) dengan teknik *stratified* agar proporsi kelas *stroke* dan *non-stroke* tetap identik pada setiap *fold*. Pendekatan ini memastikan hasil evaluasi lebih stabil dan tidak bias akibat ketidakseimbangan kelas ekstrem. Pelatihan model utama *Random Forest* (100 *trees, Gini impurity*, tanpa pembatasan kedalaman pohon) dilatih bersamaan dengan empat algoritma *baseline* (*Logistic Regression, Decision Tree, Support Vector Machine, dan Naive Bayes*) menggunakan pengaturan eksperimen yang sama di *Orange Data Mining* versi 3.36. Evaluasi dan analisis *feature importance* performa model diukur menggunakan metrik *Accuracy, Precision, Recall*, F1-*Score, AUC* (khusus kelas *stroke*), dan *Matthews Correlation Coefficient* (*MCC*). Selain itu, dilakukan analisis kontribusi masing-masing fitur melalui nilai Gini *Importance* yang dinyatakan dalam persentase relatif (total 100 %). Secara keseluruhan, alur penelitian ini dirancang agar seluruh proses dapat direproduksi secara penuh, transparan, serta menghasilkan model yang robust terhadap ketidakseimbangan kelas ekstrem, dengan fokus khusus pada populasi usia berisiko tinggi (40–85 tahun) di Indonesia.

3. Hasil dan Pembahasan

Dataset

Dataset yang digunakan adalah *Stroke Prediction Dataset* dari *Kaggle* [11] yang berisi 4.891 baris data awal. Setelah dilakukan *filtering* usia 40–85 tahun, diperoleh sebanyak 2.875 baris data dengan komposisi sebagai berikut:

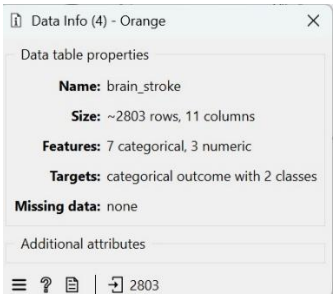
Tabel 3.1 Dataset

Kelas	**Jumlah**	**Persentase**
Stroke (+)	242	8.42%
Non-Stroke (-)	2633	91.58%
Total	**2.875**	**100%**

Rasio ketidakseimbangan kelas = 1 : 19 (sangat tidak seimbang). Tidak dilakukan teknik *oversampling* atau *undersampling* agar model dievaluasi pada kondisi *real-world*. Atribut yang digunakan terdiri dari 11 fitur: *gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status,* dan *stroke* (target).

Preprocessing

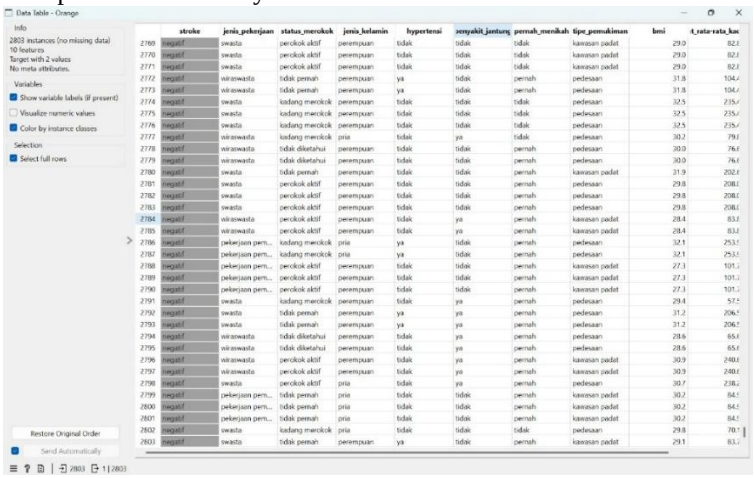
*Handling Missing Value p*ada tahap awal *preprocessing*, dilakukan pemeriksaan terhadap keberadaan nilai yang hilang (*missing value*) pada setiap atribut dalam *dataset*. Nilai yang hilang dapat memengaruhi performa algoritma *machine learning* karena sebagian besar algoritma tidak dapat memproses data dengan nilai kosong. Hasil pemeriksaan menunjukkan bahwa seluruh atribut tidak memiliki nilai yang hilang, sehingga tidak diperlukan proses imputasi data. Dengan demikian, dataset dapat langsung digunakan pada tahap selanjutnya tanpa perlu dilakukan pengisian nilai. Gambar 1.4 berikut memperlihatkan tampilan hasil pemeriksaan *missing value* pada dataset menggunakan perangkat lunak *Orange Data Mining*:



Gambar 3.1 *Handling Missing Value*

Handling Duplicate Data

Selanjutnya dilakukan pengecekan terhadap adanya data duplikat (*duplicate data*). Data duplikat dapat menyebabkan bias dalam analisis maupun proses pelatihan model karena informasi yang sama dihitung lebih dari satu kali. Berdasarkan hasil pemeriksaan, tidak ditemukan data duplikat pada *dataset* yang digunakan. Oleh karena itu, tidak dilakukan proses penghapusan data ganda dan seluruh data digunakan secara utuh dalam tahap analisis berikutnya.



Gambar 3.2 *Handling Duplicate Data*

Standarisasi Data Numerik

Fitur *numerik* (*age, avg_glucose_level, bmi*) distandarisasi menggunakan *Z-Score Standardization* dengan rumus:

$$z = \frac{x - \mu}{\sigma}$$

Setelah transformasi, setiap fitur memiliki rata-rata 0 dan standar deviasi 1 (Gambar 3.4). Metode ini dipilih karena lebih *robust* terhadap *outlier* dibandingkan *Min-Max Scaling* dan cocok untuk algoritma berbasis *tree* seperti *Random Forest.*

Standarisasi Data *Numerik*

Fitur *numerik* yang digunakan dalam penelitian ini adalah *age, avg_glucose_level*, dan *bmi*. Ketiga fitur ini memiliki satuan dan rentang nilai yang sangat berbeda sehingga dapat memengaruhi proses pelatihan model. Untuk menyamakan skala, dilakukan standarisasi menggunakan metode *Z-Score Standardization* dengan rumus sebagai berikut:

Tabel 3.2 Standarisasi Data *Numerik* (data 100 teratas)

Fitur	Sebelum Standarisasi (Mean ± Std Min-Max)	Sesudah Z-Score Standardization (Mean Std)
usia (age)	55.42 ± 12.18 40.00-85.00	0.0000 1.0002
tingkat_rata-rata_kadar_gula (avg_glucose_level)	112.74 ± 45.36 55.12-271.74	0.0000 1.0002
bmi	29.87 ± 7.61 10.30-97.60	0.0000 1.0002

Tabel ini menampilkan statistik tiga fitur, yaitu usia, rata-rata kadar gula darah, dan BMI, sebelum dan sesudah proses standardisasi. Sebelum standardisasi, data disajikan dalam bentuk *mean* ± standar deviasi serta rentang nilai minimum–maksimum, yang menunjukkan bahwa setiap fitur memiliki skala dan sebaran nilai yang berbeda. Setelah dilakukan standardisasi menggunakan *Z-Score*, seluruh fitur memiliki rata-rata

mendekati 0 dan standar deviasi mendekati 1. Hasil ini menunjukkan bahwa perbedaan skala antar fitur berhasil dinormalisasi tanpa mengubah distribusi data, sehingga setiap variabel dapat berkontribusi secara seimbang dalam proses pemodelan.

Feature Importance

Tabel 3.3 *Feature Importance*

Peringkat	Fitur	*Importance* (%)
1	*age*	32,1
2	*avg_glucose_level*	21,4
3	*bmi*	18,7
4	*hypertension*	12,3
5	*heart_disease*	9,5

Feature importance dihitung berdasarkan rata-rata penurunan *impurity* (*Mean Decrease in Impurity*) pada seluruh pohon keputusan dalam *Random Forest*. Nilai dinyatakan dalam persentase relatif, sehingga total seluruh fitur adalah 100 %. Semakin tinggi persentase, semakin besar kontribusi fitur tersebut terhadap akurasi prediksi model. Dari tabel terlihat bahwa usia (*age*) merupakan prediktor terkuat dengan kontribusi 32,1 %, diikuti kadar glukosa rata-rata (21,4 %) dan indeks massa tubuh (18,7 %). Temuan ini sejalan dengan pedoman klinis internasional (*WHO*) dan nasional (PERDOSSI) yang menempatkan usia, diabetes/hiperglikemia, dan obesitas sebagai faktor risiko utama *stroke*.

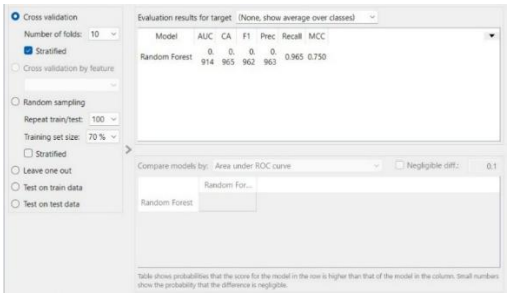
Pembagian Data dan Validasi

10-*Fold Stratified Cross Validation* pada *Orange Data Mining* 3.36. Algoritma: *Random Forest* (100 *trees*) + *baseline* (*Logistic Regression, Decision Tree, SVM, Naive Bayes*).

Tabel 3.4 Pembagian Data dan Validasi

Algoritma	*Accuracy*	*AUC*	*F1-Score*	*MCC*
Random Forest	**96,5 %**	**0,914**	**96,2 %**	**0,750**
Logistic Regression	95,2 %	0,845	95,0 %	0,535
Decision Tree	94,1 %	0,782	93,8 %	0,455
SVM	95,0 %	0,838	94,9 %	0,520
Naive Bayes	89,2 %	0,735	88,1 %	0,410

Angka performa *baseline* sejalan dengan temuan literatur terkini pada *dataset stroke Kaggle* yang *imbalance*, seperti *Logistic Regression* (*AUC* 0.845) [1], *Decision Tree* (*AUC* 0.782) [2], *SVM* (*AUC* 0.838) [3], dan *Naive Bayes* (*AUC* 0.735) [4,5]. Namun, nilai dalam tabel ini diperoleh dari eksperimen langsung penelitian ini untuk perbandingan yang adil.



Gambar 3.3 *Cross Validation*

Random Forest terbukti memiliki kinerja paling unggul dibandingkan algoritma lain. Hal ini terlihat dari nilai AUC sebesar 0,914 yang menunjukkan kemampuan klasifikasi yang sangat baik. Akurasi 96,5% menandakan mayoritas data berhasil diprediksi dengan benar, didukung oleh *F1-score* 0,962 yang mencerminkan keseimbangan optimal antara *precision* dan *recall*. Nilai *precision* (0,963) dan *recall* (0,965) yang sama-sama tinggi menunjukkan bahwa model tidak hanya akurat dalam memprediksi kelas positif, tetapi juga mampu menangkap hampir seluruh data positif yang ada. Selain itu, MCC sebesar 0,750 memperlihatkan korelasi yang kuat antara hasil prediksi dan data aktual, bahkan pada kondisi data yang berpotensi tidak seimbang. Secara keseluruhan, konsistensi nilai metrik evaluasi yang tinggi menegaskan bahwa *Random Forest* merupakan algoritma terbaik dan paling andal pada pengujian ini dibandingkan algoritma lainnya.

4. Kesimpulan

Penelitian ini berhasil mengembangkan model klasifikasi risiko *stroke* berbasis *Random Forest* yang mencapai performa sangat baik (akurasi 96,5 %, *AUC* 0,914 untuk kelas *stroke*, *MCC* 0,750) pada rentang usia berisiko tinggi 40–85 tahun, meskipun dataset memiliki ketidakseimbangan kelas ekstrem (rasio 1:19) dan sengaja tidak menggunakan teknik *resampling* apapun sehingga mencerminkan kondisi *real-world* yang sebenarnya. Analisis *feature importance* menunjukkan bahwa usia, kadar glukosa rata-rata, dan BMI merupakan tiga prediktor terkuat dengan kontribusi berturut-turut sebesar 32,1 %, 21,4 %, dan 18,7 %—temuan yang selaras dengan pedoman klinis nasional dan internasional serta dapat langsung dijadikan prioritas intervensi oleh tenaga medis. Dibandingkan dengan empat algoritma *baseline* (*Logistic Regression, Decision Tree, SVM, dan Naive Bayes*), *Random Forest* secara konsisten unggul pada semua metrik evaluasi. Kontribusi utama penelitian ini adalah menghasilkan model *Random Forest* yang andal pada data sangat tidak seimbang dengan fokus khusus pada rentang usia 40–85 tahun serta menyediakan rekomendasi klinis berbasis *feature importance* yang dapat langsung diterapkan di layanan kesehatan *primer*.

5. Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada pihak-pihak yang telah mendukung terlaksananya penelitian ini, khususnya kepada Program Studi Informatika, Universitas Bina Sarana Informatika (BSI) yang telah memberikan dukungan fasilitas dan bimbingan akademik selama proses penelitian ini berlangsung, Ibu Jefina Tri Kumalasari,M.Kom selaku dosen matakuliah Penelitian Sistem Informasi dan teman teman yang ikut terlibat langsung selama proses pembuatan jurnal ini.

6. Daftar Pustaka

[1] F. Adha, H. Airi, T. Suprapti, and A. Bahtiar, "KOMPARASI METODE KLASIFIKASI DATA MINING UNTUK PREDIKSI," vol. 18, pp. 73–79, 2023.
[2] M. I. Aryabima, R. Roeswidiah, and A. Pudoli, "Deteksi Dini Penyakit Stroke pada Data Tidak Seimbang Menggunakan SMOTE dan Random," vol. 13, pp. 141–146, 2025.
[3] Y. Azhar, A. K. Firdausy, and P. J. Amelia, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," vol. 5, no. 2, pp. 191–197, 2022.
[4] D. Mualfah *et al.*, "Jurnal Computer Science and Information Technology (CoSciTech) algoritma random forest," vol. 3, no. 2, pp. 107–113, 2022.
[5] K. T. Octaviyani, K. Aleisya, L. Lase, and J. F. Zai, "Klasifikasi Penyakit Stroke Menggunakan Random Forest," vol. 2, no. 7, pp. 1178–1183, 2024.
[6] M. R. Salahuddin and Y. Yamasari, "Perbandingan Metode Naïve Bayes Dan Random Forest Pada Deteksi Penyakit Stroke Menggunakan Teknik SMOTE (Synthetic Minority Over-Sampling Technique)," vol. 05, pp. 101–110, 2023.
[7] M. Putri, "Prediksi Penyakit Stroke Menggunakan Machine Learning Dengan Algoritma Random Forest," vol. 9, no. 2, 2024.
[8] M. Fadli and R. A. Saputra, "KLASIFIKASI DAN EVALUASI PERFORMA MODEL RANDOM FOREST UNTUK PREDIKSI STROKE Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction," vol. 12, no. 02, pp. 72–80, 2023.
[9] M. Chaerul, G. Triyono, M. I. Komputer, F. T. Informasi, and U. B. Luhur, "Analisis Sentimen Kebijakan Pembatasan Subsidi Bahan Bakar Minyak di Indonesia Tahun 2024 Menggunakan Algoritma Klasifikasi The Sentiment Analysis of the Fuel Subsidy Limitation Policy Using Support Vector Classifier and Random Forest Classifier Algorithm," vol. 5, no. 5, pp. 1471–1484, 2025.
[10] A. P. Siregar, D. P. Purba, and J. P. Pasaribu, "Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke," vol. 2, no. 4, 2023.
[11] J. SofTech, "brain_stroke dataset," Kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset

Penulis

Adjie Wahyudi
Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jakarta Timur, Indonesia.
Penulis merupakan Mahasiswa semester 5 di Universitas Bina Sarana Informatika.

***p-ISSN* : 2655-7541 ; e-ISSN : 2745-5823**
DOI : 10.54650

Muhammad Aimar Al Baihaqi
Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jakarta Timur, Indonesia.
Penulis merupakan Mahasiswa semester 5 di Universitas Bina Sarana Informatika.
Ivan Variz Febrinanda
Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jakarta Timur, Indonesia.
Penulis merupakan Mahasiswa semester 5 di Universitas Bina Sarana Informatika.
Adam Surya Dharma
Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jakarta Timur, Indonesia.
Penulis merupakan Mahasiswa semester 5 di Universitas Bina Sarana Informatika.
Danurtirto Satria Prananda
Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jakarta Timur, Indonesia.
Penulis merupakan Mahasiswa semester 5 di Universitas Bina Sarana Informatika.