STRING (Satuan Tulisan Riset dan Inovasi Teknologi) p-ISSN: 2527 - 9661
Vol. 11 No. 1 Agustus 2026 e-ISSN: 2549 - 2837

PERBANDINGAN LOGISTIC REGRESSION DAN RANDOM FOREST UNTUK PREDIKSI RISIKO STROKE

Aldrey Diriyah[1*], Affan Alfarabi[2], Fadhlurrohman Arif Mukhlis[3], Putri Nuriya Salsabila[4], I Gde Eka Dirgayussa[5], Nurul Maulidiyah[6]
[1,2,3,4,5,6] Teknik Biomedis, Institut Teknologi Sumatera
aldrey.122430054@student.itera.ac.id [1], affan.122430077@student.itera.ac.id [2],
fadhlurrohman.122430144@student.itera.ac.id [3], putri.122430080@student.itera.ac.id [4] ,
i.dirgayussa@bm.itera.ac.id [5], nurul.maulidiyah@bm.itera.ac.id [6]

Submitted April 4, 2026; Revised June 22, 2026; Accepted July 23, 2026

Abstrak

Stroke merupakan salah satu penyebab utama kematian global yang memerlukan upaya deteksi dini melalui teknologi kecerdasan buatan. Penelitian ini bertujuan untuk melakukan analisis perbandingan antara model Logistic Regression dan Random Forest dalam memprediksi risiko stroke berdasarkan data kesehatan dan gaya hidup. Data yang digunakan bersumber dari *Stroke Prediction Dataset* yang melalui tahap pra-pemrosesan, standarisasi, dan penanganan ketidakseimbangan kelas. Evaluasi performa model dilakukan menggunakan metrik Akurasi, Presisi, *Recall*, dan ROC-AUC. Hasil penelitian menunjukkan bahwa Random Forest mengungguli Logistic Regression secara signifikan dengan nilai AUC sebesar 0,990 berbanding 0,892. Random Forest terbukti lebih efektif untuk skrining medis karena mampu mencapai *Recall* sebesar 96,29%, jauh lebih tinggi dibandingkan dengan Logistic Regression yang hanya mencapai 83,71%. Selain itu, Random Forest lebih unggul dalam menangkap pola data non-linear pada variabel kompleks seperti BMI. Disimpulkan bahwa Random Forest merupakan metode yang lebih direkomendasikan untuk sistem deteksi dini stroke karena kemampuannya meminimalkan kasus positif yang tidak terdeteksi (*false negative*).

Kata Kunci : Data Kesehatan, Logistic Regression, *Machine Learning*, Prediksi Stroke, Random Forest.

Abstract

Stroke is a leading cause of global mortality that requires early detection through artificial intelligence technologies. This study aims to conduct a comparative analysis between Logistic Regression *and Random Forest models in predicting stroke risk based on health and lifestyle data. The study utilized the Stroke Prediction Dataset, which underwent preprocessing, standardization, and class imbalance handling. Model performance was evaluated using Accuracy, Precision, Recall, and ROC-AUC metrics. The results demonstrate that* Random Forest *significantly outperformed* Logistic Regression *with an AUC value of 0.990 compared to 0.892.* Random Forest *proved to be more effective for medical screening, achieving a Recall of 96.29%, which is substantially higher than* Logistic Regression *at 83.71%. Furthermore,* Random Forest *demonstrated superior capability in capturing non-linear data patterns in complex variables such as BMI. It is concluded that* Random Forest *is the recommended method for stroke early detection systems due to its ability to minimize undetected positive cases (false negatives).*

Keywords : *Health Data,* Logistic Regression*, Machine Learning,* Random Forest*, Stroke Prediction.*

1. PENDAHULUAN

Stroke merupakan salah satu penyakit penyumbang angka kematian utama sekaligus penyebab disabilitas fisik di banyak negara [1]. World Stroke Organization (2025) mencatat 12 juta insiden baru dan 7,3 juta korban jiwa akibat kondisi ini setiap tahunnya. Sebagian besar pasien yang selamat sering kali mengalami disfungsi tubuh seperti kelemahan anggota gerak, kesulitan bicara, hingga penurunan kemampuan kognitif yang berdampak panjang pada kualitas hidup [2].

Di Indonesia, stroke menempati urutan kedua sebagai penyakit dengan biaya katastropik tertinggi dalam program JKN (Jaminan Kesehatan Nasional) dengan mencapai 3.899.305 insiden. Peningkatan ini terjadi seiring dengan tingginya prevalensi hipertensi, diabetes, kebiasaan merokok, serta kurangnya aktivitas fisik pada masyarakat modern [3]. Oleh sebab itu, upaya deteksi dini menjadi hal penting agar tindakan preventif dapat segera dilakukan sebelum gejala klinis muncul.

Perkembangan teknologi kecerdasan buatan membuka peluang pemanfaatan rekam medis untuk memprediksi tingkat bahaya suatu penyakit [4]. *Machine learning* mampu mengolah informasi dalam skala masif guna menemukan pola tersembunyi, sehingga pendekatannya banyak diimplementasikan pada riset klinis [5].

Dua algoritma yang paling sering diuji adalah Logistic Regression dan Random Forest. Logistic Regression kerap dipilih karena *output*-nya mudah dipahami serta mampu menunjukkan bobot pengaruh setiap faktor terhadap potensi stroke[6]. Walaupun begitu, model linear ini cenderung kurang akurat saat menghadapi dataset dengan hubungan antarvariabel yang kompleks. Sebaliknya, Random Forest bersifat lebih fleksibel dalam memetakan pola non-linear, sehingga kerap menghasilkan prediksi berakurasi lebih tinggi [7].

Berbagai studi terdahulu menegaskan urgensi perbandingan kedua algoritma tersebut. Penelitian terkait telah mengembangkan sistem prediksi menggunakan beberapa metode *machine learning* dan melaporkan bahwa Random Forest mencatatkan performa terbaik pada dataset publik, terutama yang memuat metrik gaya hidup serta profil medis dasar [8], [9].

Temuan serupa juga ditunjukkan studi lain yang mengevaluasi Decision Tree, Random Forest, beserta Logistic Regression. Hasil riset tersebut mengonfirmasi keunggulan akurasi model *ensemble*, sementara pendekatan Logistic Regression tetap dipertahankan pada konteks klinis karena kemudahan interpretasinya bagi tenaga medis[10] .

Eksperimen dengan data kesehatan nasional menunjukkan bahwa pemodelan Random Forest dan XGBoost secara signifikan melampaui kinerja regresi konvensional. Namun, algoritma linear tetap disarankan sebagai fondasi utama untuk aplikasi medis yang membutuhkan tingkat transparansi tinggi [11].

Rangkuman literatur di atas mengindikasikan bahwa tidak ada satupun arsitektur yang selalu mendominasi pada semua situasi. Evaluasi komparatif antara Logistic Regression dan Random Forest tetap relevan untuk dikaji karena kapabilitas prediksi sangat bergantung pada karakteristik dataset.

Sebagai upaya untuk menguji performa kedua model tersebut pada kasus penyakit kritis, kombinasi variabel klinis dan perilaku pasien dalam *Stroke Prediction Dataset* dari Kaggle memungkinkan dilakukannya evaluasi yang objektif terhadap kedua algoritma tersebut [12], [13]. Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membandingkan efektivitas Logistic Regression dengan Random Forest dalam memprediksi risiko stroke. Luaran yang dihasilkan diharapkan mampu memberikan rekomendasi algoritma paling ideal sebagai sistem pendukung keputusan medis, khususnya dalam program pencegahan dini.

2. METODE PENELITIAN

Metode penelitian ini menggunakan pendekatan kuantitatif dengan menerapkan algoritma machine learning untuk memprediksi risiko stroke berdasarkan data. Penelitian ini menggunakan *Stroke*

Prediction Dataset yang bersumber dari repositori publik Kaggle (diunggah oleh fedesoriano pada tahun 2021). Data klinis sekunder ini mencakup parameter fisiologis dan gaya hidup dari 5.110 partisipan, yang awalnya dihimpun dari catatan medis pasien untuk memprediksi probabilitas terjadinya stroke berdasarkan faktor risiko tertentu [13].

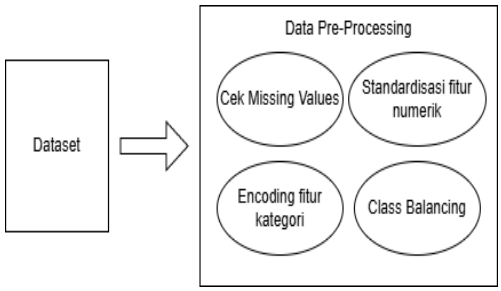
Data ini memuat sejumlah variabel klinis dan perilaku, seperti usia, tekanan darah, kadar glukosa, jenis kelamin, riwayat merokok, hipertensi, penyakit jantung, serta indeks massa tubuh. Untuk memberikan gambaran awal mengenai struktur variabel, disajikan Tabel 1 yang memuat deskripsi setiap fitur dalam dataset.

Tabel 1. Deskripsi Variabel Dataset *Stroke*

Variabel	Tipe	Deskripsi
gender	Kategorikal	Laki-laki/Perempuan
age	Numerik	Usia responden
hypertension	Biner	0/1
heart_disease	Biner	0/1
ever_married	Kategorikal	Yes/No
work_type	Kategorikal	Jenis pekerjaan
average_glucose_level	Numerik	Kadar glukosa
bmi	Numerik	Indeks massa tubuh
stroke	Biner	Label

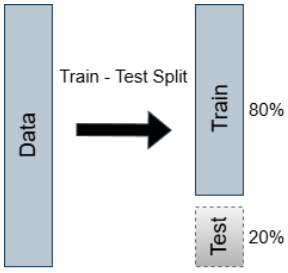
Tahap awal penelitian dimulai dengan proses pra-pemrosesan data. Pada tahap ini dilakukan pemeriksaan nilai hilang, identifikasi anomali, serta penanganan variabel kategorikal menggunakan label *encoding*. Untuk variabel numerik, dilakukan standarisasi agar semua fitur berada pada skala yang seragam sehingga dapat mengurangi bias model pada fitur tertentu [14].

Proses ini juga mencakup penanganan ketidakseimbangan kelas, mengingat jumlah sampel dengan label “stroke” sangat sedikit dibandingkan sampel “non-stroke”. Pada Logistic Regression, digunakan teknik *class weighting*, sedangkan pada Random Forest diterapkan *balanced bootstrap sampling* untuk meningkatkan sensitivitas model terhadap kelas minoritas [15]. Penjelasan proses pra-pemrosesan data divisualisasikan pada gambar 1 berikut.



Gambar 1. Diagram Alir Pra-Pemrosesan Data

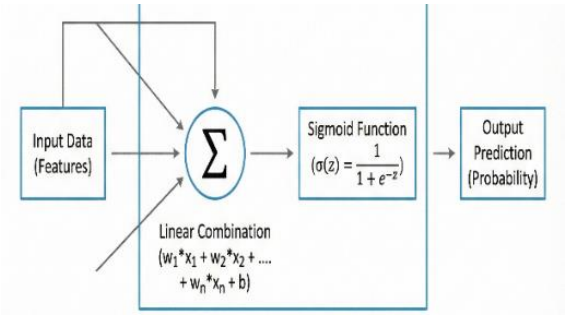
Setelah data dibersihkan dan diproses, dataset dibagi menjadi *training set* dan *testing set* dengan rasio 80:20 untuk memastikan model diuji menggunakan data yang tidak pernah dilihat sebelumnya [15]. Ilustrasi pembagian data ini ditampilkan pada Gambar 2 sebagai representasi mekanisme *train–test split*, yang merupakan praktik standar dalam penelitian *machine learning*.



Gambar 2. Diagram *Test-Split*

Tahap berikutnya adalah pembangunan model menggunakan Logistic Regression dan Random Forest. Logistic Regression digunakan sebagai model *baseline* karena interpretabilitasnya yang tinggi, model ini mampu menunjukkan pengaruh setiap variabel terhadap probabilitas terjadinya stroke melalui koefisien regresi [16].

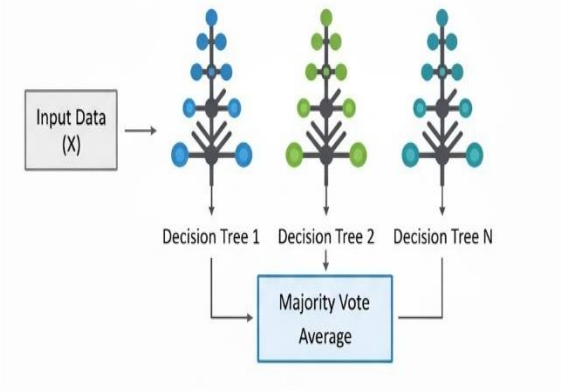
Optimasi Logistic Regression dilakukan menggunakan *solver liblinear* dengan regularisasi L2 dan nilai kekuatan regularisasi (C = 1.0) untuk mengurangi *overfitting* [5]. Selain itu, model juga mengaplikasikan penyeimbangan kelas (class_weight = 'balanced') untuk menangani ketidakseimbangan jumlah sampel pada dataset [10]. Arsitektur sederhana model Logistic Regression digambarkan pada Gambar 3.



Gambar 3. Diagram Arsitektur Model Logistic Regression

Sementara itu, Random Forest dipilih sebagai model *ensemble* karena kemampuannya menangkap pola non-linear dan variabilitas antarindividu yang umum ditemui pada data kesehatan [17]. Model dilatih menggunakan 150 pohon keputusan (n_estimators = 150), kedalaman maksimum pohon tidak dibatasi (max_depth = None), batas minimum sampel untuk pemisahan *node* (min_samples_split = 2), dan jumlah fitur maksimum yang dipertimbangkan pada setiap pemisahan (max_features = 'sqrt') [8].

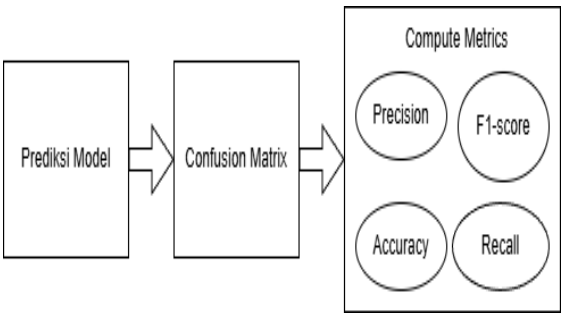
Model ini juga mengaplikasikan penyeimbangan kelas (class_weight = 'balanced') serta penetapan nilai *seed* acak (random_state = 42) guna memastikan replikabilitas model dan meningkatkan generalisasi [16]. Diagram kerja Random Forest ditampilkan pada Gambar 4 sebagai ilustrasi mekanisme *majority voting*.



Gambar 4. Diagram Arsitektur Random Forest

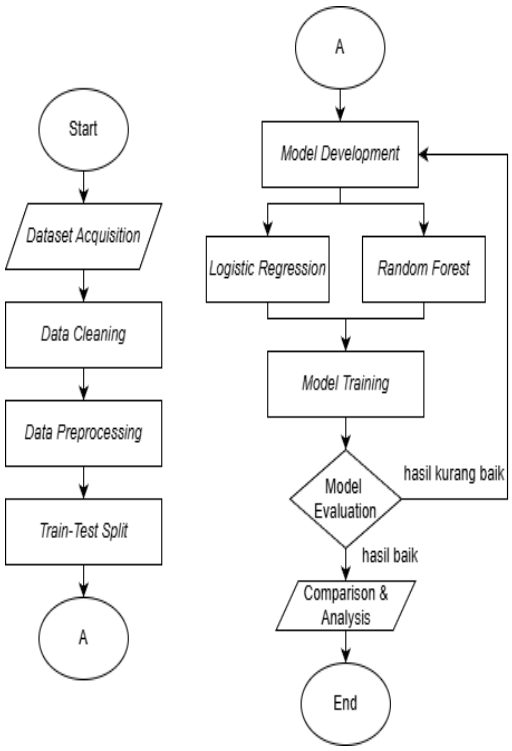
Evaluasi performa kedua model dilakukan menggunakan metrik akurasi, presisi, recall, dan *F1-score*. Pemilihan berbagai metrik evaluasi dilakukan karena kasus ini melibatkan ketidakseimbangan kelas, sehingga akurasi tidak cukup mewakili kualitas prediksi. *Recall* menjadi metrik utama yang diperhatikan, mengingat fokus penelitian ini adalah mendeteksi kasus stroke secara benar.

Selain itu, *Confusion Matrix* digunakan untuk menggambarkan proporsi prediksi benar dan salah, baik pada kelas stroke maupun non-stroke. Tahapan evaluasi lengkap ditampilkan pada Gambar 5 sebagai bagian dari alur penilaian model.



Gambar 5. Diagram Evaluasi Model (*Confusion Matrix* & Metrics)

Seluruh tahapan pemodelan dilakukan menggunakan bahasa pemrograman Python dan *library scikit-learn*, yang merupakan standar analisis data. Untuk memudahkan pemahaman, keseluruhan alur penelitian mulai dari akuisisi data hingga evaluasi model dirangkum dalam Gambar 6 berupa *flowchart* utama penelitian



Gambar 6. Diagram Alir Penelitian

3. HASIL DAN PEMBAHASAN

Bagian ini memaparkan hasil eksperimen komputasi yang telah dilakukan untuk membandingkan performa model *Logistic Regression* dan *Random Forest* dalam memprediksi risiko stroke. Pengujian dilakukan menggunakan *testing set* (20% dari total data) yang telah dipisahkan pada tahap prapemrosesan untuk menjamin objektivitas evaluasi. Fokus utama dari pemaparan hasil ini adalah untuk melihat sejauh mana kedua algoritma mampu mengenali pola pada data kesehatan dan gaya hidup, serta bagaimana keduanya menangani tantangan ketidakseimbangan kelas yang terdapat pada *dataset*.

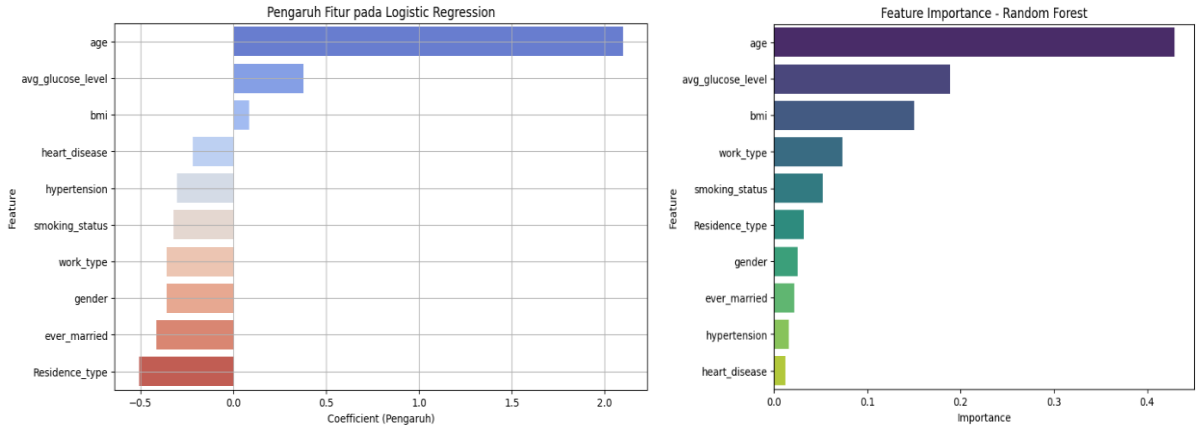
Sebelum mengevaluasi kinerja klasifikasi, kedua model telah melalui tahapan konfigurasi *hyperparameter* secara definitif guna memastikan hasil yang adil dan dapat direplikasi. Pengaturan eksplisit seperti penyesuaian parameter pada Logistic Regression, serta optimasi “max_depth”, “min_samples_split”, dan “max_features” pada *Random Forest* telah diterapkan.

Lebih lanjut, integritas penelitian dijaga ketat dengan menghilangkan atribut pengidentifikasi unik (kolom id) dari *dataset* guna mencegah terjadinya kebocoran data [18]. Langkah preventif ini memastikan bahwa performa akhir yang dihasilkan oleh kedua model murni berasal dari pembelajaran terhadap variabel prediktif klinis.

Analisis diawali dengan meninjau kontribusi setiap variabel terhadap prediksi model melalui nilai koefisien pada Logistic Regression dan *feature importance* pada Random Forest. Evaluasi ini bertujuan untuk mengidentifikasi faktor risiko dominan, seperti usia dan kadar glukosa, serta memahami perbedaan cara kedua model dalam menangkap hubungan antarvariabel, baik yang bersifat linear maupun non-linear. Interpretasi terhadap fitur-fitur ini penting untuk memastikan bahwa model tidak hanya akurat secara statistik, tetapi juga relevan dan dapat dijelaskan dalam konteks medis.

Selanjutnya, kinerja model dievaluasi secara menyeluruh menggunakan *Confusion Matrix* dan kurva ROC-AUC. Mengingat sistem ini ditujukan untuk deteksi dini, analisis akan menitikberatkan pada metrik *recall* (sensitivitas) guna meminimalkan kasus stroke yang tidak terdeteksi (*false negative*). Capaian metrik ini akan menjadi dasar penentuan model terbaik yang aman digunakan sebagai sistem pendukung keputusan.

p-ISSN: 2527 - 9661
e-ISSN: 2549 - 2837



Gambar 7. Grafik Pengaruh Fitur Logistic Regression dan Grafik Pengaruh Fitur Random Forest

Tabel 2. Pengaruh Fitur Logistic Regression

No	*Feature*	*Coefficient*
1	age	2.098491
2	avg_glucose_level	0.378211
3	bmi	0.086637
4	heart_disease	-0.216771
5	hypertension	-0.302377
6	smoking_status	-0.319280
7	work_type	-0.357363
8	gender	-0.359241
9	ever_married	-0.413617
10	Residence_type	- 0.507609

Tabel 3. Pengaruh Fitur Random Forest

No	*Feature*	*Importance*
1	age	0.429243
2	avg_glucose_level	0.188557
3	bmi	0.150357
4	work_type	0.073521
5	smoking_status	0.052271
6	Residence_type	0.031774
7	gender	0.025500
8	ever_married	0.021324
9	hypertension	0.015632
10	heart_disease	0.011821

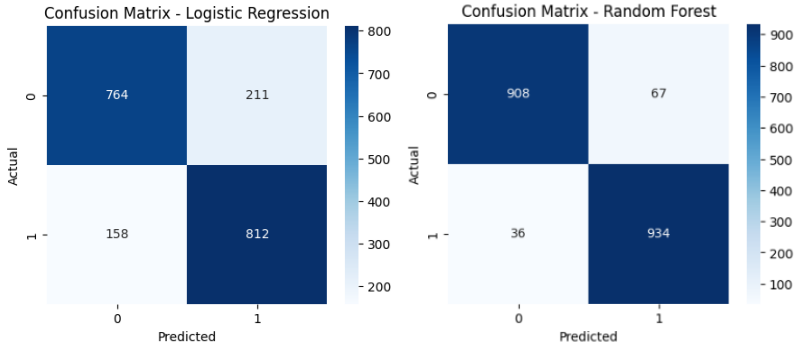
Berdasarkan Gambar 7 dan Tabel yang menyertainya, terlihat bahwa kedua model secara konsisten menempatkan *age* (usia) sebagai faktor paling dominan dalam prediksi risiko stroke. Pada Logistic Regression, usia memiliki koefisien positif terbesar yaitu 2,098, yang berarti peningkatan usia secara kuat menaikkan probabilitas prediksi stroke. Temuan ini diperkuat oleh Random Forest, di mana usia memiliki nilai *importance* tertinggi sebesar 0,429, menunjukkan bahwa fitur ini paling sering dan paling efektif digunakan model dalam memisahkan kelas stroke dan non-stroke.

Selain usia, kedua model juga menegaskan pentingnya kadar glukosa rata-rata (*avg_glucose_level*) dan Indeks Massa Tubuh (*BMI*) dengan variabel fitur *bmi* sebagai prediktor utama. Pada Logistic Regression, *avg_glucose_level* memiliki koefisien positif yang cukup menonjol (0,378), sedangkan *bmi* menunjukkan pengaruh positif yang lebih kecil (0,086). Namun, pada Random Forest, *avg_glucose_level* (0,188) dan *bmi* (0,150) secara berurutan menempati posisi kedua dan ketiga sebagai fitur paling penting. Hal ini mengindikasikan bahwa efek BMI kemungkinan bersifat lebih kompleks dan non-linear, sehingga lebih mudah ditangkap oleh model *ensemble* seperti Random Forest dibandingkan dengan model linear.

Sementara itu, variabel yang berkaitan dengan gaya hidup dan kondisi sosial seperti tipe pekerjaan (*work_type*), status merokok (*smoking_status*), tipe tempat tinggal (*Residence_type*), jenis kelamin

(*gender*), dan status pernikahan (*ever_married*) terlihat memiliki pengaruh dan tingkat kepentingan yang lebih kecil pada kedua model. Meskipun demikian, variabel-variabel tersebut tetap berkontribusi sebagai faktor pendukung yang memperkaya pola prediksi.

Secara keseluruhan, perbandingan kedua gambar menunjukkan bahwa Logistic Regression lebih menonjol dalam memberikan arah pengaruh (positif/negatif) sehingga mudah diinterpretasikan secara klinis, sedangkan Random Forest lebih kuat dalam menangkap pola non-linear dan interaksi antar variabel. Dengan dihilangkannya fitur pengidentifikasi yang tidak relevan, model kini secara murni mendeteksi kontribusi fitur kesehatan dan perilaku dalam memprediksi risiko stroke.



Gambar 8. *Confusion Matrix* Logistic Regression dan *Confusion Matrix* Random Forest

Tabel 4. *Confusion Matrix* Logistic Regression

Kelas	*Precision*	*Recall*	*F1-Score*	*Support*
0	0.83	0.78	0.81	975
1	0.79	0.84	0.81	970
Accuracy	—	—	0.81	1945
Macro Avg	0.81	0.81	0.81	1945
Weighted Avg	0.81	0.81	0.81	1945

Tabel 5. *Confusion Matrix* Random Forest

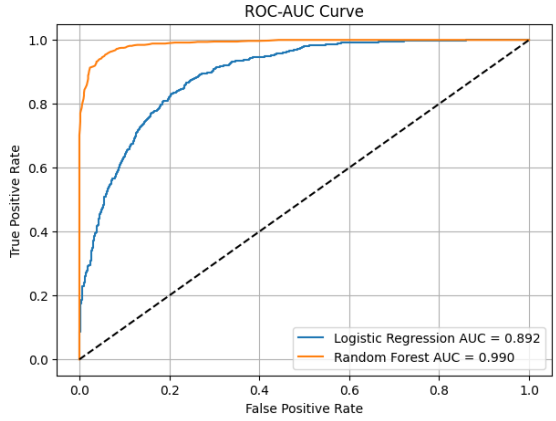
Kelas	*Precision*	*Recall*	*F1-Score*	*Support*
0	0.96	0.93	0.95	975
1	0.93	0.96	0.95	970
Accuracy	–	–	0.95	1945
Macro Avg	0.95	0.95	0.95	1945
Weighted avg	0.95	0.95	0.95	1945

Berdasarkan *Confusion Matrix* pada Gambar 8 dan Tabel yang menyertainya, terlihat perbedaan performa yang sangat jelas antara Logistic Regression dan Random Forest dalam mendeteksi kasus stroke. Pada Logistic Regression, dari total 970 kasus stroke (kelas 1), model berhasil memprediksi benar 812 kasus (TP) namun masih melewatkan 158 kasus (FN), sehingga *recall* atau sensitivitasnya berada pada angka 84%. Selain itu, dari 975 kasus non-stroke (kelas 0), model memprediksi benar 764 kasus (TN) tetapi menghasilkan 211 prediksi yang salah (FP), sehingga spesifisitasnya mencapai 78,36%. Kondisi ini menunjukkan bahwa Logistic Regression masih menghasilkan kesalahan klasifikasi yang cukup besar, baik pada sisi *false negative* (stroke tidak terdeteksi) maupun *false positive* (non-stroke dianggap stroke).

Sebaliknya, pada Random Forest terjadi peningkatan kemampuan secara signifikan pada dua aspek sekaligus: menangkap kasus stroke dan menekan *false positive*. Dari 970 kasus stroke, Random Forest memprediksi

benar 934 kasus (TP) dan hanya 36 kasus yang terlewat (FN), sehingga berhasil mencapai *recall* sebesar 96,29%. Dari 975 kasus non-stroke, model memprediksi benar 908 kasus (TN) dan hanya 67 kasus yang salah prediksi (FP), menghasilkan spesifisitas yang tinggi sebesar 93,13%.

Jika dibandingkan secara langsung, model Random Forest berhasil menurunkan *False Negative* (FN) dari 158 menjadi 36 (turun drastis sebesar 77,22%) dan menurunkan *False Positive* (FP) dari 211 menjadi 67 (turun sebesar 68,25%). Artinya, Random Forest jauh lebih unggul dan efisien untuk konteks skrining medis karena mampu menekan jumlah pasien stroke yang lolos dari deteksi secara maksimal, sekaligus meminimalkan alarm palsu pada pasien yang sebenarnya sehat.



Gambar 9. Grafik Kurva ROC-AUC *True Positive Rate* antara Logistic Regression dengan Random Forest

Berdasarkan Gambar 9 kurva ROC-AUC, terlihat bahwa kurva Random Forest (oranye) berada jauh lebih dekat ke sudut kiri-atas dibandingkan Logistic Regression (biru). Hal ini menunjukkan kemampuan Random Forest dalam mencapai *True Positive Rate* (TPR) yang tinggi dengan tetap mempertahankan *False Positive Rate* (FPR) yang sangat rendah. Nilai AUC Random Forest sebesar 0,990 mengindikasikan kemampuan identifikasi yang sangat kuat dalam membedakan kelas stroke dan non-stroke pada berbagai ambang keputusan (*threshold*).

Sementara itu, Logistic Regression memiliki AUC sebesar 0,892. Meskipun masih tergolong baik, kurvanya berada secara signifikan di bawah Random Forest, terutama pada area FPR rendah. Artinya, jika sistem pendukung keputusan medis diprioritaskan untuk menekan alarm palsu (*false positive*), Logistic Regression akan kehilangan lebih banyak sensitivitas dibandingkan Random Forest. Dengan demikian, hasil kurva ROC-AUC ini memperkuat temuan pada *Confusion Matrix* bahwa Random Forest jauh lebih unggul, stabil, dan dapat diandalkan untuk tugas prediksi risiko stroke berbasis murni pada data kesehatan dan gaya hidup.

4. SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma Random Forest secara signifikan menunjukkan akurasi prediktif yang lebih tinggi daripada Logistic Regression dalam memprediksi risiko stroke. Model Random Forest mencatatkan performa klasifikasi yang sangat baik dengan nilai AUC sebesar 0,990 dan *recall* mencapai 96,29%, jauh melampaui capaian Logistic Regression yang berada pada angka AUC 0,892 dan *recall* 83,71%. Tingginya *recall* membuat Random Forest efektif untuk skrining medis karena mampu mereduksi *false negative* sebesar 77,22% dan *false positive* sebesar 68,25%. Dari segi ekstraksi fitur, kedua algoritma konsisten mengidentifikasi usia dan kadar glukosa rata-rata sebagai faktor risiko paling dominan. Meskipun Logistic Regression memiliki keunggulan dalam hal interpretabilitas arah pengaruh variabel secara linear, Random Forest terbukti lebih baik dalam memetakan pola non-linear pada variabel yang lebih kompleks seperti BMI. Oleh karena itu, penelitian ini merekomendasikan Random Forest sebagai arsitektur yang paling ideal untuk

diimplementasikan pada sistem pendukung keputusan klinis yang memprioritaskan akurasi tinggi dan minimalisasi kesalahan diagnosis.

DAFTAR PUSTAKA

[1] C. Bushnell, W. N. Kernan, and A. Sharrief, "2024 Guideline for The Primary Prevention of Stroke: A Guideline from the American Heart Association/American Stroke Association," Dec. 01, 2024, *American Academy of Neurology*. doi: 10.1161/STR.0000000000000475.

[2] V. L. Feigin, M. Brainin, J. Pandian, and P. Lindsay, "World Stroke Organization: Global Stroke Fact Sheet 2025," *International Journal of Stroke*, vol. 20, no. 2, pp. 132–144, Feb. 2025, doi: 10.1177/17474930241308142.

[3] N. Venketasubramanian, F. L. Yudiarto, and D. Tugasworo, "Stroke Burden and Stroke Services in Indonesia," *Cerebrovasc. Dis. Extra*, vol. 12, no. 1, pp. 53–57, Mar. 2022, doi: 10.1159/000524161.

[4] A. Wahyudi, M. Al Baihaqi, I. Febrinanda, A. Dharma, and D. Prananda, "Penerapan Random Forest untuk Klasifikasi Risiko Penyakit Stroke Pada Rentang Usia 40-85 Tahun," *Jurnal Ilmu Komputer dan Informatika*, vol. 8, no. 2, pp. 2745–5823, Dec. 2025, doi: 10.54650.

[5] S. Dev, H. Wang, C. Nwosu, and N. Jain, "A Predictive Analytics Approach for Stroke Prediction Using Machine Learning and Neural Networks," *Healthcare Analytics*, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100032.

[6] C. Lou, M. Zhang, J. Li, and D. Xu, "Comparison of Logistic Regression and Machine Learning Methods for Predicting Early Neurological Deterioration After Thrombolysis in Patients with Mild Stroke," *Front. Neurol.*, vol. 17, Mar. 2026, doi: 10.3389/fneur.2026.1703890.

[7] M. Rafiq, A. Nazeer, and A. Gilani, "Application of Machine Learning Techniques for Predicting Stroke Disease," *VAWKUM Transactions on Computer Sciences*, vol. 12, no. 2, pp. 123–136, Nov. 2024, doi: 10.21015/vtcs.v12i2.1906.

[8] K. Moulaei, L. Afshari, R. Moulaei, and B. Sabet, "Explainable Artificial Intelligence for Stroke Prediction Through Comparison of Deep Learning and Machine Learning Models," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-82931-5.

[9] A. Byna, M. Modi Lakulu, and I. Y. Panessai, "Machine Learning-Based Stroke Prediction: A Critical Analysis," *IJASEIT*, vol. 14, no. 5, 2024.

[10] J. Huang and W. Liu, "Comparison of Machine Learning Models for Predicting Stroke Risk in Hypertensive Patients Lasso Regression Model, Random Forest Model, Boruta Algorithm Model, and Boruta Algorithm Combined with Lasso Regression Model," *Medicine (United States)*, vol. 104, no. 22, May 2025, doi: 10.1097/MD.0000000000042690.

[11] L. Sitompul, A. Nababan, M. Manihuruk, W. Ponsen, and Supriyandi, "Comparison of Xgboost, Random Forest and Logistic Regression Algorithms in Stroke Disease Classification," *Sinkron*, vol. 9, no. 2, pp. 957–968, Jun. 2025, doi: 10.33395/sinkron.v9i2.14794.

[12] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *Sensors*, vol.

22, no. 13, Jul. 2022, doi: 10.3390/s22134670.

[13] Fedesoriano, "Stroke Prediction Dataset," Kaggle. Accessed: Jun. 22, 2026. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[14] X. He, S. Wang, Y. Li, and J. Wang, "The Association of Domain-specific Physical Activity and Sedentary Activity with Stroke: A Prospective Cohort Study," *Shanghai Academy of Artificial Intelligence for Science*, Dec. 2023.

[15] S. Sundaram, Pavithra, and Poojasree, "Stroke Prediction Using Machine Learning," *IARJSET*, vol. 9, no. 6, Jun. 2022, doi: 10.17148/iarjset.2022.9620.

[16] I. Ovyawan and S. Violina, "Model Prediksi Risiko Stroke Menggunakan Machine Learning Stroke Risk Prediction Model Using Machine Learning," *Journal of Information Technology and Computer Science (INTECOMS)*, vol. 7, no. 4, Jun. 2024.

[17] A. Papadopoulou, D. Harding, G. Slabaugh, E. Marouli, and P. Deloukas, "Prediction of Atrial Fibrillation and Stroke using Machine Learning Models in UK Biobank," *Heliyon*, vol. 10, no. 7, Apr. 2024, doi: 10.1016/j.heliyon.2024.e28034.

[18] W. Deng, C. Li, Z. Yan, and Y. Yuan, "Machine Learning-Based Stroke Prediction," in *Science and Technology Publications*, INSTICC, Sep. 2024, pp. 754–761. doi: 10.5220/0012972300004508.