JACIS : Journal Automation Computer Information System Volume: **6**, Nomor **1**, **2026**, *Hlm*: 127-137

Prediksi Risiko Stroke Berdasarkan Faktor Klinis Menggunakan *Random Forest* Dengan Optimasi *Threshold* dan SHAP

Chaerul Hidayat [1]*, Agung Nugroho [2], Asep Suprianto [3]

*[1] Universitas Pelita Bangsa, chaerulhidayat@mhs.pelitabangsa.ac.id
[2] Universitas Pelita Bangsa, agung@pelitabangsa.ac.id
[3] Universitas Pelita Bangsa, asep.supriyanto@pelitabangsa.ac.id

Abstrak: Stroke merupakan salah satu penyebab utama kematian dan kecacatan di berbagai negara, sehingga diperlukan metode prediksi risiko yang akurat berbasis data klinis. Namun, penelitian sebelumnya umumnya masih menghadapi permasalahan ketidakseimbangan data serta kurang memberikan interpretasi terhadap faktor klinis yang berpengaruh. Penelitian ini mengembangkan model prediksi risiko stroke dengan mengatasi ketidakseimbangan data serta meningkatkan interpretabilitas model. Dataset yang digunakan terdiri dari 5.110 data pasien dengan distribusi kelas yang tidak seimbang. Untuk mengatasi permasalahan tersebut, diterapkan metode SMOTEENN dan algoritma Balanced Random Forest, serta analisis menggunakan pendekatan Explainable Artificial Intelligence (XAI). Hasil evaluasi menunjukkan bahwa model memiliki kinerja yang cukup baik dengan nilai *accuracy* sebesar 79,26% dan ROC-AUC sebesar 82,01%. Namun, nilai *precision* yang relatif rendah (15,68%) menunjukkan masih terdapat prediksi positif yang tidak tepat, sebagai konsekuensi dari peningkatan *recall* sebesar 74% dalam mendeteksi kasus stroke sebagai kelas minoritas. Analisis SHAP menunjukkan bahwa usia, kadar glukosa, hipertensi, dan indeks massa tubuh merupakan faktor utama dalam prediksi risiko stroke. Penelitian ini memberikan kontribusi dalam meningkatkan deteksi kasus stroke serta menyediakan interpretasi model yang lebih transparan.
Keyword: Stroke; Random Forest; SMOTEENN; SHAP; Prediksi Risiko Stroke.

DOI: https://doi.org/10.47134/jacis.v6i1.178
*Correspondensi: Chaerul Hidayat
Email: chaerulhidayat@mhs.pelitabangsa.ac.id

Receive: 15 Maret 2026
Accepted: 01 April 2026
Published: 9 April 2026

Abstrak: Stroke is one of the leading causes of death and disability in many countries, making accurate risk prediction methods based on clinical data essential. However, previous studies have generally faced issues with data imbalance and have failed to provide sufficient interpretation of the influencing clinical factors. This study develops a stroke risk prediction model by addressing data imbalance and improving model interpretability. The dataset consists of 5,110 patient records with an imbalanced class distribution. To address these issues, the SMOTEENN method and the Balanced *Random Forest* algorithm were applied, along with analysis using the Explainable Artificial Intelligence (XAI) approach. Evaluation results show that the model performs reasonably well, with an *accuracy* of 79.26% and an ROC-AUC of 82.01%. However, the relatively low *precision* (15.68%) indicates that there are still inaccurate positive predictions, a consequence of the increased *recall* of 74% in detecting stroke cases as the minority class. SHAP analysis indicates that age, glucose levels, hypertension, and body mass index are the primary factors in stroke risk prediction. This study contributes to improving stroke case detection and provides a more transparent model interpretation.

Keywords: Stroke; Random Forest; SMOTEENN; SHAP; Stroke Risk Prediction.

PENDAHULUAN

Stroke masih menjadi salah satu penyebab utama kematian dan kecacatan di seluruh dunia. Lebih dari 12 juta kasus stroke terjadi setiap tahun dengan sekitar 6,6 juta kematian secara global [1]. Peningkatan jumlah ini sangat berkaitan dengan perubahan gaya hidup dan bertambahnya orang yang memiliki faktor risiko. Di Indonesia, Riskesdas 2018 mencatat prevalensi stroke sebesar 10,9 per 1.000 orang [2]. Penyebaran kasus yang hampir merata di berbagai provinsi menunjukkan bahwa faktor risiko stroke memiliki karakteristik yang kompleks dan bervariasi, sehingga diperlukan pendekatan berbasis machine learning yang mampu menangkap pola hubungan kompleks antar variabel klinis dalam memprediksi risiko stroke.

Deteksi dini terhadap kemungkinan terjadinya stroke sangat krusial untuk menekan tingkat kematian dan kecacatan akibat kondisi tersebut. Metode konvensional seperti *Framingham Stroke Risk Score* (FSRS) kerap digunakan, namun metode tersebut memiliki beberapa kelemahan. Hubungan antar variabel klinis dianggap linier, padahal dalam praktiknya sering kali lebih kompleks. Beberapa riset juga menunjukkan bahwa akurasi FSRS menurun ketika diterapkan pada populasi Asia yang memiliki variasi klinis dan demografi yang lebih beragam [3]. Dalam situasi ini, *machine learning* digunakan karena kemampuannya dalam menangkap pola hubungan kompleks antar variabel klinis yang melibatkan banyak faktor.

Beberapa studi telah memanfaatkan algoritma *Random Forest* (RF) untuk prediksi risiko stroke dan menunjukkan kinerja yang cukup baik. Namun, sebagian besar studi tersebut masih berfokus pada peningkatan akurasi tanpa memperhatikan permasalahan ketidakseimbangan data serta kemampuan model dalam mendeteksi kelas minoritas. Studi pada[4] menunjukkan bahwa *Random Forest* mampu mencapai akurasi yang lebih tinggi dibandingkan *Decision Tree*. Temuan serupa juga dilaporkan pada[5], yang menganalisis perbandingan *antara Random Forest, XGBoost,* dan *SVM*, di mana *Random Forest* menghasilkan klasifikasi yang paling stabil. Selain itu, [6] menekankan efektivitas *Random Forest* dalam mengklasifikasikan diagnosis stroke berdasarkan data klinis. Meskipun demikian, sebagian besar studi tersebut masih terfokus pada perbandingan kinerja algoritma dan belum membahas faktor klinis yang paling berpengaruh dalam prediksi stroke. Selain itu, mayoritas penelitian sebelumnya masih mengandalkan dataset publik, seperti *Kaggle*, dan lebih berfokus pada hasil prediksi atau kinerja klasifikasi model yang digunakan.

Penekanan tersebut menyebabkan aspek interpretabilitas model sering kali terabaikan. Namun, pendekatan kecerdasan buatan yang dapat dijelaskan (*Explainable Artificial Intelligence*/XAI) memberikan peluang untuk mengevaluasi pengaruh variabel klinis, khususnya melalui analisis pentingnya fitur pada algoritma *Random Forest*. Studi pada[7] berfokus pada evaluasi kinerja algoritma tanpa menganalisis kontribusi masing-masing fitur klinis, sehingga interpretasi terhadap hasil prediksi masih terbatas. Selain itu, [8] menyoroti pengujian keakuratan *Random Forest* dengan pendekatan data mining yang hanya sedikit membahas aspek klinis terkait stroke. Keterbatasan tersebut membuka peluang untuk memanfaatkan *Random Forest* tidak hanya sebagai metode klasifikasi, tetapi juga sebagai alat analisis dalam mengidentifikasi prioritas klinis untuk upaya pencegahan dini.

Meningkatnya jumlah penderita stroke di Indonesia mengindikasikan keterbatasan metode tradisional dalam memprediksi risiko secara akurat. Oleh karena itu, metode *Random Forest* dipilih karena kemampuannya dalam menangani hubungan nonlinier antar variabel serta mengurangi risiko *overfitting* pada data. Studi pada[9] menunjukkan efektivitas *Random Forest* dalam kasus stroke, sementara [10] menekankan pentingnya klasifikasi risiko berbasis data. Berdasarkan hal tersebut, penelitian ini bertujuan untuk membangun model prediksi sekaligus mengidentifikasi faktor klinis yang paling berpengaruh terhadap risiko stroke.

Penelitian ini mengembangkan model prediksi risiko stroke menggunakan algoritma *Random Forest* berdasarkan data klinis yang tersedia secara publik. Selain mengevaluasi kinerja model, penelitian ini juga menganalisis faktor klinis yang paling berpengaruh terhadap risiko stroke melalui pendekatan interpretasi model. Hasil penelitian diharapkan dapat mendukung pengembangan sistem skrining berbasis *machine learning* serta meningkatkan pemanfaatan kecerdasan buatan dalam bidang kesehatan.

Kontribusi utama penelitian ini terletak pada integrasi metode SMOTEENN untuk mengatasi ketidakseimbangan data serta penerapan SHAP untuk meningkatkan interpretabilitas model dalam prediksi risiko stroke. Pendekatan ini tidak hanya berfokus pada peningkatan kinerja model, tetapi juga memberikan pemahaman terhadap faktor klinis yang berpengaruh, yang masih terbatas dibahas secara bersamaan dalam penelitian sebelumnya.

Namun, penelitian ini memiliki beberapa keterbatasan. Data yang digunakan merupakan dataset publik dari platform *Kaggle* sehingga mungkin belum sepenuhnya merepresentasikan kondisi klinis yang lebih luas. Selain itu, variabel yang digunakan terbatas pada fitur yang tersedia dalam dataset, sehingga belum mencakup seluruh faktor risiko stroke secara komprehensif.

METODE

Dataset Penelitian

Dataset yang digunakan dalam penelitian ini adalah *Stroke Prediction Dataset* yang diperoleh dari platform Kaggle (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset). Dataset ini berisi data klinis pasien terkait faktor risiko stroke dengan total 5.110 data pasien. Atribut yang tersedia meliputi usia, jenis kelamin, hipertensi, penyakit jantung, rata-rata kadar glukosa, indeks massa tubuh (BMI), serta status merokok. Variabel target dalam dataset ini adalah kejadian stroke yang diklasifikasikan menjadi dua kategori, yaitu stroke dan tidak stroke. Distribusi data menunjukkan ketidakseimbangan kelas, di mana jumlah kasus stroke lebih sedikit dibandingkan non-stroke.

Pengumpulan Dataset

Data yang digunakan dalam penelitian ini diperoleh dari dataset publik *Stroke Prediction Dataset* yang tersedia di platform *Kaggle*. Dataset ini digunakan sebagai dasar dalam pengembangan model prediksi risiko stroke.

Eksplorasi Data

Berdasarkan hasil eksplorasi data, distribusi kelas pada dataset menunjukkan ketidakseimbangan yang signifikan antara kelas stroke dan non-stroke, di mana sekitar 95%

data merupakan non-stroke dan hanya sekitar 5% merupakan stroke. Kondisi ini berpotensi menyebabkan model cenderung bias terhadap kelas mayoritas sehingga diperlukan penanganan khusus terhadap data tidak seimbang. Temuan ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa distribusi kasus stroke umumnya tidak seimbang dengan dominasi kelas non-stroke [11]. Selain itu, analisis awal menunjukkan bahwa variabel seperti usia, kadar glukosa, dan hipertensi memiliki kecenderungan berbeda antara kedua kelas, sehingga berpotensi menjadi faktor penting dalam prediksi risiko stroke.

Preprocessing Data

Pada tahap preprocessing, dilakukan beberapa langkah untuk mempersiapkan data sebelum proses pemodelan [12]. Pembersihan data dilakukan dengan menghapus atribut yang tidak relevan seperti id karena tidak berpengaruh terhadap proses prediksi [13]. Selanjutnya, dilakukan seleksi fitur dengan mempertahankan variabel klinis yang relevan, yaitu usia, jenis kelamin, hipertensi, penyakit jantung, kadar glukosa rata-rata, indeks massa tubuh (BMI), dan status merokok guna meningkatkan kinerja model [14]. Setelah itu, dilakukan transformasi data dengan mengonversi atribut kategorikal menjadi format numerik menggunakan teknik *one-hot encoding* [15]. Tahap berikutnya adalah penanganan nilai yang hilang pada variabel BMI menggunakan metode imputasi median untuk menjaga kestabilan distribusi data [16].

Pembagian Data

Dataset yang telah melalui tahap *preprocessing* kemudian dibagi menjadi data latih dan data uji dengan perbandingan 80:20. Pembagian ini dilakukan menggunakan teknik *stratified* sampling untuk menjaga proporsi distribusi kelas pada kedua subset data. Selain itu, digunakan parameter random_state sebesar 42 untuk memastikan proses pembagian data dapat direproduksi. Data latih digunakan untuk membangun model *Random Forest*, sedangkan data uji digunakan untuk mengevaluasi kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya, sehingga dapat mengurangi risiko overfitting [17].

Penanganan Ketidakseimbangan Data

Karena dataset memiliki distribusi kelas yang tidak seimbang, maka dilakukan penanganan menggunakan metode SMOTEENN dengan memanfaatkan *library imbalanced-learn*. Metode ini menggabungkan teknik SMOTE untuk menambah data pada kelas minoritas dan *Edited Nearest Neighbor* (ENN) untuk menghapus data yang dianggap sebagai noise [18].

Pemodelan

Tahap pemodelan dilakukan menggunakan algoritma *Random Forest* dengan memanfaatkan *library imbalanced-learn*, yaitu *BalancedRandomForestClassifier.* Model dibangun dengan parameter jumlah pohon (n_estimators) sebanyak 100, kedalaman maksimum pohon (max_depth) bernilai None (*default*), serta random_state sebesar 42 untuk memastikan konsistensi hasil. Model dilatih menggunakan data latih yang telah melalui proses penyeimbangan kelas, kemudian diuji menggunakan data uji untuk mengevaluasi performa prediksi.

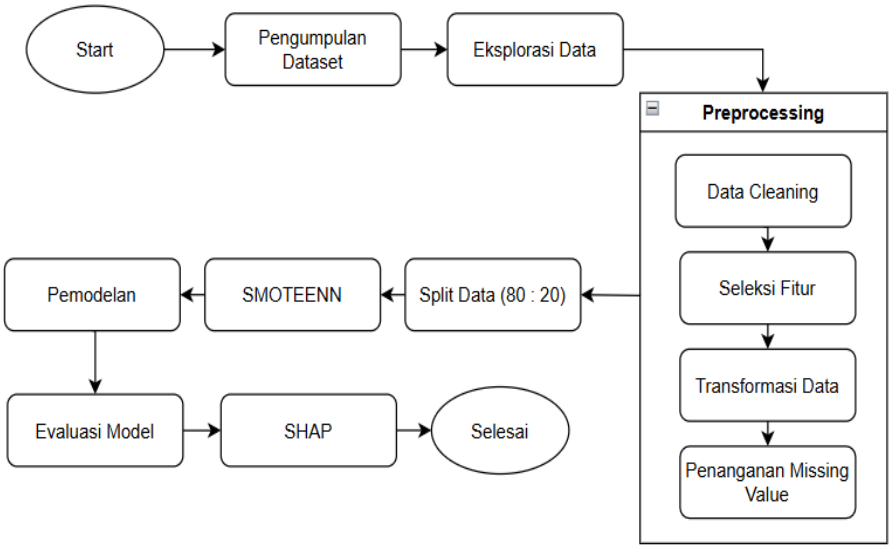
Evaluasi Model

Kinerja model dinilai dengan berbagai metrik evaluasi seperti akurasi, presisi, *recall* , F1-score, dan ROC-AUC untuk mengukur efektivitas model dalam proses klasifikasi [19]. Metrik tersebut digunakan untuk menilai seberapa baik model dapat mengkategorikan data dan membedakan antara kelas stroke dan non-stroke.

Interpretasi Model (SHAP)

Setelah model dibangun, dilakukan interpretasi model menggunakan metode SHAP (*SHapley Additive Explanations*) untuk mengetahui kontribusi masing-masing fitur terhadap hasil prediksi model [20].

Rangkuman tahapan tersebut ditunjukkan pada Gambar 1



Gambar 1. Alur penelitian

HASIL DAN PEMBAHASAN

Dataset

Dataset yang digunakan dalam penelitian ini merupakan dataset publik dari platform Kaggle yang berisi 5.110 data pasien dengan variabel target stroke (0 = tidak stroke, 1 = stroke). Distribusi kelas pada dataset awal ditunjukkan pada tabel berikut.

Tabel 1. Distribusi dataset

Kelas	Jumlah
Tidak Stroke (0)	4861
Stroke(1)	249
Total	5110

Berdasarkan Tabel 1, jumlah data yang tidak mengalami stroke adalah sebanyak 4.861 data (95,13%), sedangkan data yang mengalami stroke berjumlah 249 data (4,87%) dari total 5.110 data. Penyebaran ini menunjukkan bahwa data yang ada bersifat tidak seimbang karena kelas yang tidak mengalami stroke jauh lebih dominan dibandingkan kelas yang mengalami stroke.

Penanganan Imbalance (SMOTEENN)

Dataset yang digunakan dalam penelitian ini memiliki distribusi kelas yang tidak seimbang antara kelas stroke dan tidak stroke. Oleh karena itu, dilakukan penanganan ketidakseimbangan data menggunakan metode SMOTEENN untuk meningkatkan kemampuan model dalam mendeteksi kelas minoritas serta mengurangi bias model terhadap kelas mayoritas.

Tabel 2. Data setelah SMOTEENN

Kelas	Sebelum	Sesudah
Stroke	249	**3617**
Tidak Stroke	4861	**2957**

Berdasarkan Tabel 2, distribusi data setelah penerapan metode SMOTEENN menjadi lebih seimbang dibandingkan dengan distribusi data awal. Proses ini membantu model dalam mempelajari pola pada kelas stroke sebagai kelas minoritas. Hal ini meningkatkan kemampuan model dalam mengenali pola pada kelas minoritas yang sebelumnya terabaikan, sehingga kinerja klasifikasi menjadi lebih optimal.

Implementasi Random Forest

Pada tahap ini, model *Random Forest* diimplementasikan menggunakan data latih yang telah melalui proses penyeimbangan kelas dengan metode SMOTEENN. Proses penyeimbangan ini dilakukan untuk mengatasi permasalahan ketidakseimbangan kelas pada dataset, sehingga model dapat belajar secara lebih baik dalam mengenali pola pada kedua kelas, yaitu stroke dan tidak stroke. Model kemudian dilatih untuk membedakan kedua kelas tersebut berdasarkan fitur-fitur klinis yang tersedia pada dataset. Setelah proses pelatihan selesai, model diuji menggunakan data uji yang tidak terlibat dalam proses pelatihan guna mengevaluasi kemampuan model dalam melakukan prediksi. Evaluasi awal dilakukan dengan menggunakan nilai *threshold default* sebesar 0,5 untuk menentukan klasifikasi hasil prediksi. Hasil evaluasi awal model yang diperoleh dari proses pengujian tersebut ditunjukkan pada tabel berikut.

Tabel 3. Hasil evaluasi awal (*Threshold* 0,5)

Metrik	Nilai
Accuracy	0,9110
Precision	0,2029
Recall	0,2800
F1-Score	0,2353
ROC-AUC	0,8201

Berdasarkan Tabel 3, hasil evaluasi awal terlihat bahwa tingkat akurasi cukup tinggi, yaitu sekitar 91,10%. Namun demikian, nilai *recall* untuk kelas stroke hanya sekitar 28,00%. Ini menunjukkan bahwa model belum mampu mendeteksi sebagian besar kasus stroke secara optimal. Oleh karena itu, perlu dilakukan penyesuaian *threshold* untuk meningkatkan sensitivitas model terhadap kelas minoritas pada tahap selanjutnya.

Optimasi *Threshold*

Pada model klasifikasi, nilai probabilitas prediksi umumnya menggunakan *threshold* sebesar 0,5 sebagai batas untuk menentukan kelas. Namun, pada dataset yang memiliki distribusi

kelas tidak seimbang, penggunaan *threshold default* tersebut sering kali menyebabkan model kurang optimal dalam mendeteksi kelas minoritas, yaitu kasus stroke.

Berdasarkan hasil evaluasi awal pada *threshold* 0,5, nilai *recall* untuk kelas stroke masih relatif rendah. Hal ini menunjukkan bahwa model belum mampu mendeteksi sebagian besar kasus stroke secara optimal. Oleh karena itu, dilakukan pengujian beberapa nilai *threshold* untuk meningkatkan kemampuan model dalam mengenali kelas stroke sebagai kelas minoritas.

Tabel 4. Perbandingan hasil evaluasi berdasarkan *Threshold*

Threshold	*Accuracy*	*Precision*	*Recall*	F1-Score
0,5	0,9110	0,2029	0,2800	**0,2353**
0,10	0,7162	0,1273	0,8200	**0,2204**
0,20	0,7926	0,1568	0,7400	**0,2587**

Berdasarkan hasil pengujian pada beberapa nilai *threshold*, nilai 0,20 dipilih sebagai *threshold* yang paling optimal. Nilai ini memberikan keseimbangan yang lebih baik antara *precision* dan *recall* serta meningkatkan kemampuan model dalam mendeteksi kasus stroke sebagai kelas minoritas.

Evaluasi Kinerja Model

Setelah menentukan nilai *threshold* optimal, tahap selanjutnya adalah melakukan evaluasi kinerja model untuk mengetahui kemampuan model *Random Forest* dalam memprediksi risiko stroke. Evaluasi dilakukan menggunakan beberapa metrik klasifikasi, yaitu *accuracy*, *precision*, *recall* , F1-score, serta ROC-AUC.

Confusion Matrix

Confusion Matrix digunakan untuk melihat jumlah prediksi benar dan salah yang dihasilkan oleh model pada masing-masing kelas. Berdasarkan hasil pengujian yang menggunakan *threshold* sebesar 0,20, diperoleh *Confusion Matrix* sebagai berikut.

Tabel 5. *Confusion Matrix* hasil prediksi model *Random Forest*

	Prediksi Tidak Stroke (0)	Prediksi Stroke (1)	Total Aktual
Aktual Tidak Stroke (0)	773 (TN)	199 (FP)	972
Aktual Stroke (1)	13 (FN)	37 (TP)	50
Total Prediksi	786	236	1022

Berdasarkan Tabel 5, hasil *Confusion Matrix* menunjukkan bahwa model menghasilkan nilai *True Negative* (TN) sebesar 773, yang berarti sebanyak 773 data pasien yang tidak mengalami stroke berhasil diprediksi dengan benar sebagai tidak stroke. Sementara itu, nilai *False Positive* (FP) sebesar 199 menunjukkan bahwa terdapat 199 data pasien yang sebenarnya tidak stroke namun diprediksi sebagai stroke oleh model.

Pada kelas stroke, diperoleh nilai *False Negative* (FN) sebesar 13, yang menunjukkan jumlah kasus stroke yang tidak berhasil dideteksi oleh model dan diprediksi sebagai tidak stroke. Selain itu, nilai *True Positive* (TP) sebesar 37 menunjukkan jumlah data pasien yang benar-benar mengalami stroke dan berhasil diprediksi dengan benar oleh model. Dari total 50 data stroke pada data uji, model mampu mendeteksi 37 kasus stroke, sehingga menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengenali kelas stroke sebagai kelas minoritas.

Dalam konteks prediksi risiko stroke, nilai *False Negative* (FN) menjadi perhatian penting karena merepresentasikan kasus stroke yang tidak teridentifikasi oleh sistem, sehingga berpotensi menimbulkan risiko apabila digunakan sebagai dasar pengambilan keputusan.

Metrik Evaluasi Model

Berdasarkan hasil evaluasi model menggunakan *threshold* optimal, diperoleh beberapa metrik evaluasi yang digunakan untuk mengukur kinerja model dalam melakukan klasifikasi. Nilai metrik tersebut meliputi *accuracy*, *precision*, *recall* , F1-score, dan ROC-AUC seperti yang ditunjukkan pada tabel berikut.

Tabel 6. Hasil evaluasi model

Metrik	Nilai
Accuracy	**79,26%**
Precision	**15,68%**
Recall	**74%**
F1-Score	**25,87%**
ROC-AUC	**82,01%**

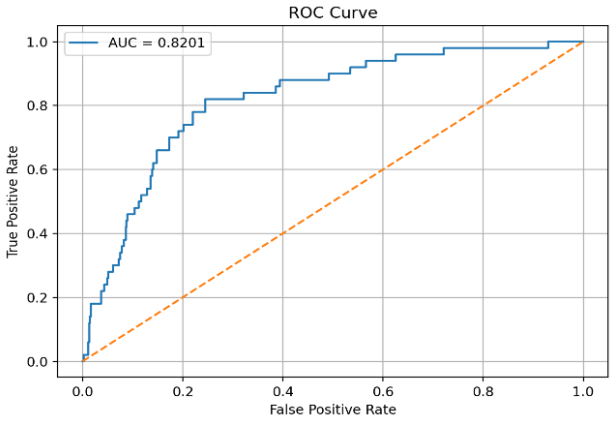
Berdasarkan hasil evaluasi pada Tabel 6, model *Random Forest* memperoleh nilai *accuracy* sebesar 79,26%. Nilai *precision* sebesar 15,68% tergolong rendah, yang disebabkan oleh penggunaan *threshold* yang relatif rendah serta penerapan metode SMOTEENN yang meningkatkan sensitivitas model terhadap kelas minoritas. Kondisi ini menyebabkan meningkatnya jumlah *false positive*, di mana model cenderung mengklasifikasikan lebih banyak data sebagai stroke. Namun, nilai *recall* sebesar 74% menunjukkan bahwa model mampu mendeteksi sebagian besar kasus stroke. Hal ini mencerminkan adanya *trade-off* antara *precision* dan *recall* , di mana model dioptimalkan untuk meningkatkan kemampuan deteksi kelas minoritas. Dalam konteks medis, pendekatan ini masih dapat diterima karena kesalahan dalam tidak mendeteksi kasus stroke (*false negative*) memiliki risiko yang lebih besar dibandingkan *false positive*. Nilai F1-score sebesar 25,87% mencerminkan keseimbangan antara *precision* dan *recall* , sedangkan ROC-AUC sebesar 82,01% menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam membedakan kedua kelas. Hasil ini sejalan dengan penelitian sebelumnya [4] dan [5] yang menunjukkan bahwa *Random Forest* efektif dalam klasifikasi data medis, namun dalam penelitian ini dilakukan penyesuaian *threshold* untuk meningkatkan deteksi kelas minoritas.

ROC Curve

Kurva ROC digunakan untuk mengevaluasi kemampuan model dalam membedakan antara kelas stroke dan tidak stroke pada berbagai nilai *threshold*. Kurva ini menggambarkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai ambang keputusan yang digunakan dalam proses klasifikasi. Nilai *Area Under Curve* (AUC) digunakan sebagai ukuran performa model secara keseluruhan, di mana semakin mendekati nilai 1 menunjukkan kemampuan model yang semakin baik dalam membedakan kedua kelas tersebut.

Berdasarkan Gambar 2, model ini memperoleh nilai ROC-AUC sebesar 0,82, yang mengindikasikan bahwa model memiliki kemampuan yang cukup baik dalam membedakan antara kelompok pasien yang mengalami stroke dan yang tidak. Nilai AUC yang mendekati 1 menunjukkan bahwa model mampu melakukan klasifikasi dengan tingkat diskriminasi

yang baik antara kedua kelas tersebut, sehingga model dapat digunakan sebagai pendekatan yang cukup efektif dalam prediksi risiko stroke berdasarkan data klinis pasien

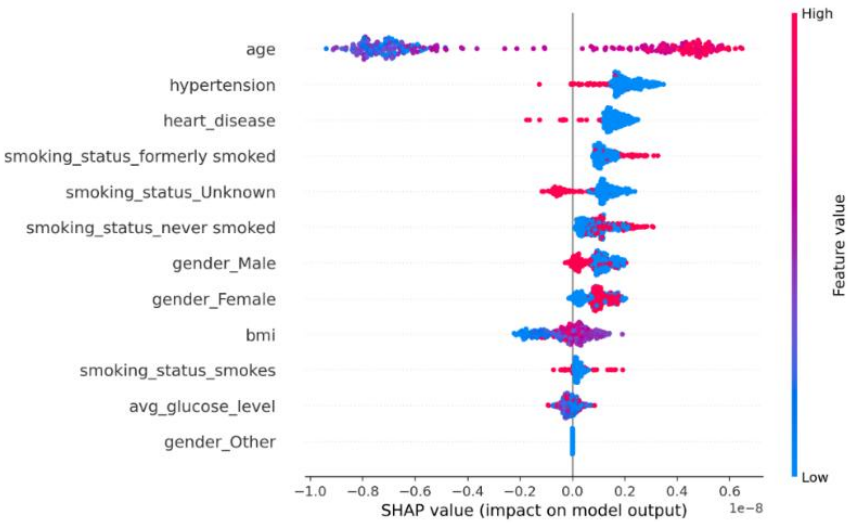


Gambar 2. Grafik ROC curve

Nilai ROC-AUC yang diperoleh dalam penelitian ini sejalan dengan beberapa penelitian sebelumnya yang menunjukkan bahwa *Random Forest* memiliki kemampuan klasifikasi yang baik dalam data medis, khususnya pada kasus prediksi stroke [4] [9]. Hal ini menunjukkan bahwa model yang dibangun memiliki perforsma yang kompetitif dibandingkan dengan pendekatan serupa.

Interpretasi Model Menggunakan SHAP

Untuk mengetahui peran masing-masing fitur dalam hasil prediksi model, dilakukan analisis memakai pendekatan SHAP (SHapley Additive Explanations). Pendekatan ini bertujuan untuk menjelaskan dampak setiap fitur dalam proses prediksi model Random Forest.



Gambar 3. SHAP Summary Plot

Gambar 3 menunjukkan kontribusi setiap fitur terhadap hasil prediksi model *Random Forest*. Fitur yang berada pada posisi paling atas menunjukkan tingkat pengaruh yang lebih besar terhadap prediksi risiko stroke. Warna merah menunjukkan nilai fitur yang tinggi, sedangkan warna biru menunjukkan nilai fitur yang rendah. Semakin jauh posisi titik dari garis tengah, semakin besar pengaruh fitur tersebut terhadap output model. Berdasarkan grafik tersebut, fitur age, hypertension, heart disease, dan BMI memiliki kontribusi yang lebih dominan

dalam menentukan prediksi risiko stroke. Hal ini menunjukkan bahwa faktor-faktor klinis tersebut memiliki peran penting dalam meningkatkan atau menurunkan kemungkinan terjadinya stroke pada pasien. Dengan adanya analisis SHAP ini, interpretasi model menjadi lebih transparan sehingga dapat membantu memahami hubungan antara faktor klinis dan prediksi yang dihasilkan oleh model.

SIMPULAN

Penelitian ini menunjukkan bahwa model yang dibangun mampu memberikan performa yang cukup baik dalam memprediksi risiko stroke berdasarkan faktor klinis, terutama dalam meningkatkan kemampuan deteksi pada kelas minoritas. Namun, masih terdapat kelemahan pada ketepatan prediksi positif, yang menunjukkan adanya prediksi yang tidak tepat. Kondisi ini mencerminkan adanya trade-off antara kemampuan deteksi kasus stroke dan tingkat ketepatan prediksi. Selain itu, hasil analisis menunjukkan bahwa faktor seperti usia, kadar glukosa, hipertensi, dan indeks massa tubuh (BMI) merupakan variabel yang berpengaruh terhadap risiko stroke. Meskipun demikian, penelitian ini masih memiliki keterbatasan karena menggunakan dataset publik dengan jumlah dan variasi fitur yang terbatas. Oleh karena itu, penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih beragam serta mengembangkan pendekatan yang dapat meningkatkan ketepatan model.

DAFTAR PUSTAKA

[1] World Stroke Organization, "World Stroke Organization Annual Report 2023," World Stroke Organization, Geneva, Switzerland, Annual Report Annual Report 2023, 2023. [Online]. Available: https://www.world-stroke.org/

[2] S. Sutrisno, C. N. Widayati, and U. Rukanah, "Hubungan Kecepatan Pertolongan Pertama Keluarga Penderita Hipertensi Dengan Kejadian Stroke Pada Penderita Hipertensi Di Wilayah Uptd Puskesmas Purwodadi I," *Shine Cahaya Dunia -1 Keperawatan*, vol. 7, no. 2, Dec. 2022, doi: 10.35720/tscs1kep.v7i2.389.

[3] S. S. Kasim *et al.*, "Validation of the general Framingham Risk Score (FRS), SCORE2, revised PCE and WHO CVD risk scores in an Asian population," *Lancet Reg. Health - West. Pac.*, vol. 35, p. 100742, Jun. 2023, doi: 10.1016/j.lanwpc.2023.100742.

[4] A. R. Dana, R. V. Kristananda, M. B. S. Wibowo, and D. A. Prasetya, "Perbandingan Algoritma Decision Tree dan *Random Forest* dengan Hyperparameter Tuning dalam Mendeteksi Penyakit Stroke," vol. 4, 2024.

[5] H. A. Nabila and Endang Wahyu Pamungkas, "Perbandingan Algoritma Machine Learning: Svm, Random Forest, Dan Xgboost Untuk Prediksi Stroke," *Rabit J. Teknol. Dan Sist. Inf. Univrab*, vol. 10, no. 2, pp. 1098–1110, Jul. 2025, doi: 10.36341/rabit.v10i2.6444.

[6] Ary Prandika Siregar, Dwi Priyadi Purba, Jojor Putri Pasaribu, and Khairul Reza Bakara, "Implementasi Algoritma *Random Forest* Dalam Klasifikasi Diagnosis Penyakit Stroke," *J. Penelit. Rumpun Ilmu Tek.*, vol. 2, no. 4, pp. 155–164, Nov. 2023, doi: 10.55606/juprit.v2i4.3039.

[7] Y. Aulia, A. Andriyansyah, S. Suharjito, and S. W. Nensi, "Analisis Prediksi Stroke dengan Membandingkan Tiga Metode Klasifikasi Decision Tree, Naïve Bayes, dan Random Forest," *J. Ilmu Komput. Dan Inform.*, vol. 3, no. 2, pp. 89–98, Jan. 2024, doi: 10.54082/jiki.90.

[8] Y. Azhar, A. K. Firdausy, and P. J. Amelia, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," *SINTECH Sci. Inf. Technol. J.*, vol. 5, no. 2, pp. 191–197, Oct. 2022, doi: 10.31598/sintechjournal.v5i2.1222.

[9] Gullam Almuzadid and Egia Rosi Subhiyakto, "Stroke Risk Classification Using the Ensemble Learning Method of XGBoost and Random Forest," *J. Appl. Inform. Comput.*, vol. 9, no. 3, pp. 828–837, Jun. 2025, doi: 10.30871/jaic.v9i3.9528.

[10] W. A. Pamungkas, "Penerapan Klasifikasi Data Mining Untuk Prediksi Dengan Metode Algoritma Decission Tree," vol. 2, no. 2, 2025.

[11] M. Putri, "Prediksi Penyakit Stroke Menggunakan Machine Learning Dengan Algoritma Random Forest".

[12] Bhargavi Konda, "The impact of data preprocessing on data mining outcomes," *World J. Adv. Res. Rev.*, vol. 15, no. 3, pp. 540–544, Sep. 2022, doi: 10.30574/wjarr.2022.15.3.0931.

[13] L. Santoso, "Mengoptimalkan Proses Pembersihan Data dalam Analisis Big Data Menggunakan Pipeline Berbasis AI," vol. 17, no. 2, 2024.

[14] M. Sutcu, D. Jouda, B. Yildiz, J. Katrib, and K. M. Almustafa, "Predicting Stroke Risk Using Machine Learning: A Data-Driven Approach to Early Detection and Prevention," *Stroke Res. Treat.*, vol. 2025, no. 1, p. 2892726, Jan. 2025, doi: 10.1155/srat/2892726.

[15] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y.-I. Cho, "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms," *Mathematics*, vol. 12, no. 16, p. 2553, Aug. 2024, doi: 10.3390/math12162553.

[16] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transit. Proc.*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[17] V. Ignatenko, A. Surkov, and S. Koltcov, "Random forests with parametric entropy-based information gains for classification and regression problems," *PeerJ Comput. Sci.*, vol. 10, p. e1775, Jan. 2024, doi: 10.7717/peerj-cs.1775.

[18] R. Nursyahfitri, C. Rozikin, and R. I. Adam, "Penerapan Metode SMOTE dalam Klasifikasi Daerah Rawan Banjir di Karawang Menggunakan Algoritma Naive Bayes," *J. Sist. Dan Teknol. Inf. JustIN*, vol. 10, no. 4, p. 339, Dec. 2022, doi: 10.26418/justin.v10i4.46935.

[19] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[20] Y. Dubey, Y. Tarte, N. Talatule, K. Damahe, P. Palsodkar, and P. Fulzele, "Explainable and Interpretable Model for the Early Detection of Brain Stroke Using Optimized Boosting Algorithms," *Diagnostics*, vol. 14, no. 22, 2024, doi: 10.3390/diagnostics14222514.