

IMPLEMENTASI RANDOM FOREST DAN SHAP PADA PREDIKSI RESIKO STROKE BERDASARKAN POLA KONSUMSI NUTRISI

Muh Fadhil Ahmad^{1*}, Chyquitha Danuputri², Ida Mulyadi³

^{1,2,3} Universitas Muhammadiyah Makassar, Indonesia

¹105841109722@student.unismuh.ac.id

²chyquitadanuputri@unismuh.ac.id

³idadamulyadi@unismuh.ac.id

*Corresponding Author

Info Artikel

Pengajuan : 31-07-2026
Peninjauan : 10-08-2026
Revisi : 26-08-2026
Diterima : 17-09-2026
Diterbitkan : 24-09-2026

Kata Kunci:

Random Forest, SHAP,
klasifikasi risiko stroke, pola
konsumsi nutrisi,
explainable artificial
intelligence.

Abstrak

Stroke merupakan salah satu penyebab utama kematian dan disabilitas di dunia, dengan pola konsumsi nutrisi yang tidak seimbang sebagai salah satu faktor risiko utamanya. Penelitian ini mengembangkan model klasifikasi tingkat risiko stroke berbasis algoritma Random Forest menggunakan data pola konsumsi nutrisi dan indikator kesehatan pasien Puskesmas Kanjilo, Kabupaten Gowa, Sulawesi Selatan, serta menerapkan metode SHAP (SHapley Additive exPlanations) untuk menginterpretasikan kontribusi setiap fitur terhadap hasil klasifikasi. Label kategori risiko (Aman, Berisiko, Sangat Berisiko) dibentuk melalui skema skor berbasis aturan klinis (rule-based scoring) atas 12 indikator kesehatan dan nutrisi, sehingga hasil evaluasi model perlu dibaca sebagai ukuran konsistensi model terhadap aturan pelabelan tersebut, bukan sebagai bukti prediksi klinis yang telah tervalidasi. Model Random Forest yang dibangun menunjukkan performa klasifikasi yang sangat tinggi dan konsisten pada pengujian data uji maupun validasi silang 5-fold, dengan kestabilan performa yang baik pada ketiga kategori risiko. Analisis SHAP menunjukkan bahwa konsumsi serat, konsumsi lemak, BMI, tekanan darah, dan status hipertensi merupakan fitur dengan kontribusi terbesar terhadap hasil klasifikasi, konsisten dengan feature importance bawaan Random Forest maupun dengan literatur epidemiologis mengenai peran nutrisi dan tekanan darah dalam risiko stroke. Model semacam ini berpotensi menjadi alat bantu skrining awal yang transparan bagi tenaga kesehatan di fasilitas pelayanan primer seperti Puskesmas untuk mengidentifikasi pasien berisiko dan mengarahkan edukasi gizi, dengan catatan bahwa penerapan klinisnya tetap memerlukan validasi lebih lanjut oleh tenaga medis serta pengujian pada data di luar skema pelabelan internal yang digunakan.

How to Cite : Ahmad, M. F., Danuputri, C., & Mulyadi, I. (2026). Implementasi Random Forest Dan SHAP Pada Prediksi Risiko Stroke Berdasarkan Pola Konsumsi Nutrisi. Journal of Computer Science and Information Technology (JCSIT), 3(4), 518–530. <https://doi.org/10.70248/jcsit.v3i4.4489>

DOI : <https://doi.org/10.70248/jcsit.v3i4.4489>

This is an open access article under <https://creativecommons.org/licenses/by/4.0/>
Copyright© 2026 by Author. Published by Yayasan Nuraini Ibrahim Mandiri



PENDAHULUAN

Stroke merupakan entitas patologis pada sistem saraf pusat yang ditandai oleh disrupsi suplai darah ke jaringan otak secara akut, sehingga memicu kerusakan seluler dan gangguan fungsi neurologis yang bersifat fokal maupun global. Secara global, stroke tetap menjadi penyebab dominan mortalitas dan disabilitas jangka panjang dengan insidensi yang terus meningkat setiap tahun (Feigin et al., 2021; WHO, 2023). Studi case-control multinasional INTERSTROKE menunjukkan bahwa sepuluh faktor risiko yang berpotensi dimodifikasi berkontribusi terhadap sekitar 90% population

attributable risk stroke di berbagai kawasan dunia, dengan hipertensi dan pola makan tidak sehat sebagai dua kontributor terbesar (O'Donnell et al., 2016). Tingginya angka kejadian tersebut menimbulkan dampak kesehatan, sosial, dan ekonomi yang signifikan, sehingga upaya preventif melalui identifikasi faktor risiko sejak dini menjadi sangat penting (Johnson et al., 2016).

Berbagai faktor seperti hipertensi, diabetes mellitus, obesitas, dan pola konsumsi nutrisi yang tidak seimbang berkontribusi terhadap meningkatnya risiko stroke (Kernan et al., 2013; Sacco et al., 2013). Studi berbasis data Riskesdas pada populasi dewasa Indonesia menunjukkan bahwa konsumsi makanan berisiko (tinggi garam, lemak, dan gula) serta rendahnya konsumsi buah dan sayur berhubungan signifikan dengan peningkatan prevalensi stroke, dengan mekanisme yang melibatkan resistensi insulin, dislipidemia, dan aterosklerosis (Syauqy et al., 2023). Secara patofisiologis, stroke iskemik terjadi akibat sumbatan pembuluh darah akibat aterosklerosis yang dipengaruhi tingginya asupan lemak (Tadi & Lui, 2022), sedangkan stroke hemoragik sering berkaitan dengan hipertensi kronis akibat konsumsi garam berlebih (Unnithan et al., 2022). Faktor-faktor tersebut bersifat multidimensional dan saling berkaitan, sehingga hubungan antarvariabel sering kali kompleks dan tidak linier, membuat pendekatan statistik tradisional kurang optimal dalam menangkap pola interaksi fitur secara menyeluruh.

Dalam beberapa tahun terakhir, pendekatan machine learning berkembang pesat dan semakin luas diimplementasikan dalam pemodelan prediktif di bidang kesehatan (Breiman, 2001; Bzdok et al., 2018), termasuk dalam klasifikasi risiko stroke. Di antara sejumlah algoritma berbasis pohon keputusan, Random Forest dipilih pada penelitian ini karena dua pertimbangan komputasi utama. Pertama, Random Forest secara alami menangani data tabular dengan variabel campuran (kategorikal dan kontinu) tanpa memerlukan asumsi distribusi tertentu, serta relatif tahan terhadap outlier dan skala fitur yang berbeda-beda (Breiman, 2001). Kedua, dibandingkan algoritma boosting seperti XGBoost (Chen & Guestrin, 2016) atau LightGBM (Ke et al., 2017) yang performanya sangat bergantung pada penalaan hyperparameter (learning rate, jumlah iterasi boosting, kedalaman pohon) dan lebih rentan overfitting pada dataset berukuran sedang tanpa regularisasi yang cermat, Random Forest melalui mekanisme bagging dan agregasi sejumlah pohon keputusan independen cenderung lebih stabil dengan konfigurasi default, sekaligus tetap kompetitif dalam tugas klasifikasi risiko stroke pada berbagai studi terdahulu (Dritsas & Trigka, 2022; Kumar & Kaur, 2021; Chawla et al., 2002).

Namun demikian, model machine learning dengan performa tinggi sering kali bersifat black box sehingga kurang transparan dalam menjelaskan alasan di balik hasil klasifikasi (Beretta & Santaniello, 2016). Dalam konteks kesehatan, transparansi sangat penting agar faktor risiko dominan dapat dipahami dengan jelas oleh tenaga medis maupun masyarakat, dan agar keluaran model dapat dipertanggungjawabkan secara etis dan legal sebelum diadopsi dalam praktik klinis (Lundberg et al., 2020). Oleh karena itu, pendekatan Explainable Artificial Intelligence (XAI) menjadi relevan untuk meningkatkan interpretabilitas model (Miller, 2019). Salah satu metode yang banyak digunakan adalah SHAP (SHapley Additive exPlanations), yang mampu mengukur kontribusi masing-masing fitur secara kuantitatif terhadap hasil klasifikasi baik secara global maupun pada level prediksi individual (Miller, 2019; Lundberg & Lee, 2017).

Penelitian ini menggunakan data dari Puskesmas Kanjilo, Kecamatan Barombong, Kabupaten Gowa, sebagai lokasi studi. Pemilihan lokasi ini didasarkan pada

pertimbangan praktis berupa ketersediaan data pola konsumsi nutrisi dan indikator kesehatan pasien secara terstruktur di fasilitas tersebut, serta relevansinya sebagai representasi fasilitas pelayanan kesehatan primer (Puskesmas) yang menjadi ujung tombak skrining faktor risiko penyakit tidak menular di tingkat masyarakat. Meskipun demikian, penelitian ini menyadari bahwa desain studi tunggal-lokasi (single-site) membatasi generalisasi temuan terhadap populasi yang lebih luas, sehingga hasil penelitian ini perlu dipahami sebagai studi kasus komputasional dan bukan representasi epidemiologis nasional. Cakupan variabel nutrisi dalam penelitian ini juga dibatasi pada empat indikator pola konsumsi harian yang tersedia dalam catatan Puskesmas, yaitu konsumsi garam, gula, lemak, dan serat, tanpa mencakup penilaian gizi menyeluruh (seperti recall makanan 24 jam atau food frequency questionnaire yang tervalidasi) sebagaimana lazim digunakan pada studi gizi epidemiologis berskala nasional (Syauqy et al., 2023).

Perlu ditegaskan pula bahwa skor risiko yang digunakan untuk membentuk label kategori (Aman, Berisiko, Sangat Berisiko) pada penelitian ini merupakan instrumen ambang batas internal yang disusun berdasarkan nilai rujukan klinis umum pada masing-masing indikator, dan belum divalidasi secara formal oleh dokter spesialis saraf maupun dibandingkan dengan instrumen skor risiko stroke yang telah baku secara klinis, seperti Framingham Stroke Risk Profile (Wolf et al., 1991). Dengan demikian, skema pelabelan pada penelitian ini diposisikan sebagai kerangka kerja (framework) awal yang bersifat eksploratif, bukan sebagai instrumen diagnostik yang siap pakai, sehingga validasi oleh tenaga medis tetap menjadi prasyarat sebelum penerapan lebih lanjut.

Meskipun berbagai penelitian telah mengembangkan model klasifikasi stroke berbasis machine learning, sebagian besar masih berfokus pada variabel klinis seperti tekanan darah, riwayat penyakit, dan hasil pemeriksaan laboratorium, sementara kajian yang secara khusus menempatkan pola konsumsi nutrisi sebagai variabel utama serta mengintegrasikan Random Forest dengan SHAP masih relatif terbatas (Athallah et al., 2025; Widiani & Yasa, 2023). Berdasarkan hal tersebut, penelitian ini mengimplementasikan Random Forest dan SHAP untuk mengklasifikasikan tingkat risiko stroke berdasarkan pola konsumsi nutrisi. Tahapan penelitian meliputi pengolahan data, pembangunan model klasifikasi, evaluasi menggunakan metrik akurasi, precision, recall, F1-score, dan confusion matrix, serta analisis interpretasi fitur menggunakan SHAP, dengan pembahasan kritis mengenai keterbatasan metodologis yang melekat pada skema pelabelan berbasis aturan yang digunakan.

METODE PENELITIAN

Penelitian ini dilaksanakan di Puskesmas Kanjilo, Kecamatan Barombong, Kabupaten Gowa, Sulawesi Selatan, dengan memanfaatkan data pola konsumsi nutrisi dan indikator kesehatan pasien sebagai dasar analisis. Seluruh tahapan penelitian dilakukan menggunakan bahasa pemrograman Python melalui Jupyter Notebook, dengan memanfaatkan pustaka Pandas dan NumPy untuk pengolahan data, Scikit-learn untuk pemodelan dan evaluasi, Matplotlib untuk visualisasi, serta SHAP untuk interpretasi model.

Label kategori risiko (Aman, Berisiko, Sangat Berisiko) dibentuk melalui skor kumulatif atas 12 indikator klinis dan pola konsumsi nutrisi (BMI, tekanan darah sistolik dan diastolik, riwayat hipertensi, diabetes, kolesterol, jantung, kadar garam, gula, lemak, serat, dan status merokok), dengan ambang kategori Aman untuk skor ≤ 3 ,

Berisiko untuk skor 4–5, dan Sangat Berisiko untuk skor >5. Sebagaimana telah dikemukakan pada bagian Pendahuluan, ambang batas ini merupakan instrumen internal yang disusun dari nilai rujukan klinis umum dan belum divalidasi terhadap instrumen skor risiko stroke yang baku seperti Framingham Stroke Risk Profile (Wolf et al., 1991); penelitian ini memposisikannya sebagai kerangka eksploratif awal.

Perlu digarisbawahi sebuah keterbatasan metodologis penting: karena variabel penyusun skor (BMI, tekanan darah, kadar garam/gula/lemak/serat, riwayat penyakit) adalah variabel yang sama yang kemudian digunakan sebagai fitur prediktor untuk melatih model Random Forest, terdapat risiko circular reasoning — model pada dasarnya mempelajari kembali aturan aritmetika yang telah digunakan untuk membentuk labelnya sendiri, bukan semata-mata menemukan pola hubungan biologis laten antara nutrisi dan risiko stroke dari data observasional independen. Kondisi ini menjadi salah satu penyebab utama akurasi klasifikasi yang sangat tinggi (mendekati sempurna) pada bagian Hasil dan Pembahasan. Fenomena bias pada ukuran feature importance akibat struktur fitur yang saling berkorelasi maupun yang terlibat langsung dalam pembentukan target juga telah didokumentasikan pada literatur Random Forest (Strobl et al., 2007). Oleh karena itu, hasil evaluasi performa pada penelitian ini sebaiknya dimaknai sebagai bukti bahwa model berhasil mempelajari aturan pelabelan secara konsisten, dan bukan sebagai bukti independen atas kemampuan prediktif klinis model terhadap kejadian stroke yang sesungguhnya. Validasi lanjutan pada data dengan label diagnosis klinis aktual (bukan skor sintesis) diperlukan sebelum model ini dapat diklaim memiliki nilai prediktif klinis.

Sebelum pelatihan model, dilakukan pemeriksaan missing value yang menunjukkan 928 dari 4.603 baris data (sekitar 20,2%) memiliki nilai kosong pada satu atau lebih atribut. Pada penelitian ini, baris dengan nilai kosong dihapus secara listwise (listwise deletion) sehingga tersisa 3.675 baris data. Pendekatan ini dipilih karena kesederhanaan implementasi, namun disadari berpotensi menimbulkan selection bias apabila pola hilangnya data (missingness) tidak sepenuhnya acak (not missing completely at random), karena unit data yang dibuang bisa jadi memiliki karakteristik klinis tertentu yang sistematis (Sterne et al., 2009). Sebagai catatan metodologis, teknik imputasi seperti mean imputation atau k-nearest neighbor imputation dapat memanfaatkan seluruh data yang tersedia dengan lebih baik dan berpotensi mengurangi bias tersebut (Beretta & Santaniello, 2016), namun penerapan dan perbandingan teknik imputasi terhadap listwise deletion pada dataset ini belum dilakukan dan diusulkan sebagai agenda penelitian lanjutan.

Atribut kategorikal (jenis kelamin, hipertensi, diabetes, kolesterol, jantung, aktivitas, dan status merokok) ditransformasikan ke bentuk numerik menggunakan LabelEncoder dari pustaka scikit-learn. Atribut ID dan Nama dihapus karena bersifat identitas, sehingga tersisa 17 atribut prediktor (X) dan satu variabel target (y) berupa label Risiko. Pembagian data latih (2.941 data, 80%) dan data uji (736 data, 20%) dilakukan menggunakan train_test_split dengan random_state=42 dan stratify=y untuk menjaga proporsi kelas tetap seimbang.

Model Random Forest Classifier dilatih dengan 100 pohon keputusan (n_estimators=100), kriteria pemisahan Gini Impurity, serta parameter class_weight=balanced untuk menangani ketidakseimbangan jumlah data antar kelas sebagai alternatif terhadap teknik oversampling seperti SMOTE (Chawla et al., 2002). Parameter max_depth pada konfigurasi awal ini diatur None (tanpa batas), yang secara umum meningkatkan risiko overfitting pada pohon keputusan individual, khususnya

ketika target berasal dari aturan deterministik seperti pada penelitian ini. Sebagai mitigasi parsial, kinerja model diverifikasi melalui 5-fold cross validation pada seluruh data untuk menilai konsistensi antar-fold. Meskipun demikian, penalaan hyperparameter yang lebih sistematis (misalnya grid search atau random search dengan pembatasan max_depth dan min_samples_leaf) belum dilakukan pada penelitian ini dan disarankan sebagai perbaikan pada penelitian selanjutnya untuk memastikan performa model tidak semata-mata mencerminkan kemampuan menghafal aturan pelabelan.

Evaluasi kinerja klasifikasi dilakukan menggunakan confusion matrix sebagai dasar analisis, yang menggambarkan hubungan antara hasil klasifikasi model dengan kondisi aktual data pada setiap kategori risiko (Sokolova & Lapalme, 2009). Berdasarkan confusion matrix tersebut, dihitung metrik accuracy, precision, recall, dan F1-score dengan pendekatan rata-rata berbobot (weighted average) untuk mengakomodasi klasifikasi multikelas. Interpretasi model dilakukan menggunakan metode SHAP melalui TreeExplainer untuk mengukur kontribusi setiap fitur terhadap hasil klasifikasi, baik secara global maupun lokal, sehingga dapat ditelusuri variabel nutrisi yang paling berpengaruh terhadap tingkat risiko stroke beserta arah kontribusinya.

HASIL PENELITIAN DAN PEMBAHASAN

Dataset Penelitian

Data yang digunakan pada penelitian ini diperoleh dari Puskesmas Kanjilo, Kecamatan Barombong, Kabupaten Gowa, Sulawesi Selatan, berupa berkas Microsoft Excel (Dataset_Stroke.xlsx) yang terdiri atas 4.603 baris data pasien dengan 19 atribut, meliputi identitas pasien, data antropometri, tekanan darah, riwayat penyakit penyerta, serta pola konsumsi nutrisi harian. Variabel Risiko Stroke merupakan variabel target (label) yang diklasifikasikan oleh model Random Forest, sedangkan variabel Umur, Berat Badan, BMI, Sistolik, Diastolik, Konsumsi Garam, Konsumsi Gula, Konsumsi Lemak, dan Konsumsi Serat merupakan fitur utama yang digunakan dalam penelitian ini. Statistik deskriptif variabel penelitian disajikan pada Tabel 1.

Tabel 1. Statistik Deskriptif Variabel Penelitian

Atribut	Mean	Std	Min	Max
Umur	61,45	12,91	25	85
Berat Badan	71,64	16,67	45	100
Tinggi Badan	163,12	10,21	145	180
BMI	27,23	7,22	13,9	46,9
Sistolik	145,38	25,43	100	190
Diastolik	85,43	14,80	60	110
Konsumsi Garam (g)	8,79	3,45	3	15
Konsumsi Gula (g)	71,30	28,96	20	119,8
Konsumsi Lemak (g)	70,89	29,11	20,4	119,9
Konsumsi Serat (g)	20,38	8,63	5	35

Berdasarkan Tabel 1, rata-rata usia pasien dalam dataset adalah 61,45 tahun, dengan rentang usia antara 25 hingga 85 tahun. Rata-rata Indeks Massa Tubuh (BMI) sebesar 27,23 menunjukkan bahwa sebagian besar pasien berada pada kategori kelebihan berat badan (overweight). Rata-rata tekanan darah sistolik sebesar 145,38 mmHg dan diastolik sebesar 85,43 mmHg mengindikasikan bahwa sebagian besar pasien memiliki tekanan darah yang relatif tinggi. Dari aspek pola konsumsi nutrisi,

rata-rata konsumsi garam harian sebesar 8,79 gram, konsumsi gula sebesar 71,30 gram, konsumsi lemak sebesar 70,89 gram, dan konsumsi serat sebesar 20,38 gram, dengan standar deviasi yang relatif besar pada konsumsi gula dan lemak, yang menunjukkan adanya variasi pola konsumsi nutrisi antarpasien.

Preprocessing Data

Sebagaimana dijelaskan pada bagian Metode Penelitian, pemeriksaan missing value menunjukkan 928 dari 4.603 baris (20,2%) memiliki nilai kosong, yang kemudian dihapus secara listwise sehingga tersisa 3.675 baris data. Atribut kategorikal ditransformasikan ke bentuk numerik menggunakan LabelEncoder. Dataset bersih kemudian disimpan sebagai data_clean3.csv. Setelah atribut ID dan Nama dihapus, tersisa 17 atribut prediktor (X) dan satu variabel target (y). Pembagian data latih (2.941 data, 80%) dan data uji (736 data, 20%) dilakukan menggunakan train_test_split dengan random_state=42 dan stratify=y.

Pelatihan dan Evaluasi Model Random Forest

Proses klasifikasi tingkat risiko stroke menggunakan algoritma Random Forest Classifier yang diimplementasikan melalui pustaka scikit-learn, dengan konfigurasi parameter seperti pada Tabel 2. Sebagaimana telah didiskusikan pada Metode Penelitian, konfigurasi max_depth=None berpotensi meningkatkan risiko overfitting, sehingga hasil pada subbab ini perlu dibaca bersamaan dengan catatan keterbatasan mengenai skema pelabelan berbasis aturan.

Tabel 2. Parameter Model Random Forest

Parameter	Nilai
n_estimators	100
criterion	gini
max_depth	None
min_samples_split	2
min_samples_leaf	1
random_state	42
class_weight	Balanced

Setelah proses pelatihan selesai, model diuji menggunakan data uji untuk menghasilkan prediksi yang selanjutnya dibandingkan dengan label aktual guna mengukur kinerja model. Evaluasi dilakukan menggunakan empat metrik utama, yaitu accuracy, precision, recall, dan F1-score dengan metode rata-rata weighted. Hasil evaluasi disajikan pada Tabel 3.

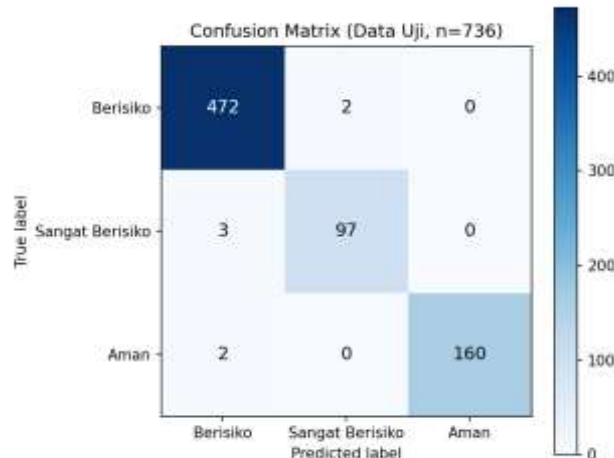
Tabel 3. Hasil Evaluasi Kinerja Model Random Forest

Metrik	Nilai
Accuracy	0,9905 (99,05%)
Precision (weighted)	0,9905 (99%)
Recall (weighted)	0,9905 (99%)
F1-Score (weighted)	0,9905 (99%)

Rincian confusion matrix hasil pengujian terhadap 736 data uji disajikan pada Tabel 4 dan divisualisasikan secara konsisten pada Gambar 1.

Tabel 4. Confusion Matrix Hasil Klasifikasi Model Random Forest

Aktual \ Prediksi	Berisiko	Sangat Berisiko	Aman	Total
Berisiko	472	2	0	474
Sangat Berisiko	3	97	0	100
Aman	2	0	160	162
Total	477	99	160	736



Gambar 1. Confusion Matrix Model Random Forest (angka telah disinkronkan dengan Tabel 4)

Berdasarkan Tabel 4 dan Gambar 1, dari 474 data aktual berlabel Berisiko, model berhasil mengklasifikasikan dengan benar sebanyak 472 data (2 data salah diklasifikasikan sebagai Sangat Berisiko). Dari 100 data aktual berlabel Sangat Berisiko, model berhasil mengklasifikasikan dengan benar 97 data (3 data salah diklasifikasikan sebagai Berisiko). Dari 162 data aktual berlabel Aman, model berhasil mengklasifikasikan dengan benar 160 data (2 data salah diklasifikasikan sebagai Berisiko). Secara keseluruhan, model berhasil mengklasifikasikan dengan benar 729 dari 736 data uji (99,05%), dengan kesalahan klasifikasi (7 data) yang seluruhnya terjadi di antara kelas Berisiko dan Sangat Berisiko yang berbatasan langsung pada skema skor, bukan pada kelas yang berjauhan seperti Aman dan Sangat Berisiko — pola ini konsisten dengan dugaan bahwa model mempelajari batas numerik skema pelabelan secara sistematis.

Untuk menilai konsistensi performa model terhadap satu skema pembagian data tertentu, dilakukan validasi silang 5-fold pada seluruh data, dengan hasil disajikan pada Tabel 6. Perlu ditegaskan bahwa nilai akurasi rata-rata validasi silang (99,6%) dan akurasi pada data uji tunggal (99,05%) merupakan dua besaran yang berasal dari skema evaluasi berbeda — yang pertama adalah rata-rata performa pada lima kombinasi pembagian data pada keseluruhan dataset, sedangkan yang kedua adalah performa pada satu kali pembagian 80:20 — sehingga keduanya bukan merupakan hasil yang saling bertentangan atau tidak konsisten, melainkan dua sudut pandang evaluasi yang saling melengkapi. Konsistensi nilai akurasi antar-fold (rentang 99,4%–99,8%) menunjukkan bahwa model secara stabil mereproduksi pola label pada berbagai partisi data, sejalan dengan penjelasan pada Metode Penelitian bahwa stabilitas ini sebagian besar mencerminkan konsistensi aturan pelabelan, bukan semata-mata kemampuan generalisasi klinis.

Tabel 5. Classification Report Model Random Forest

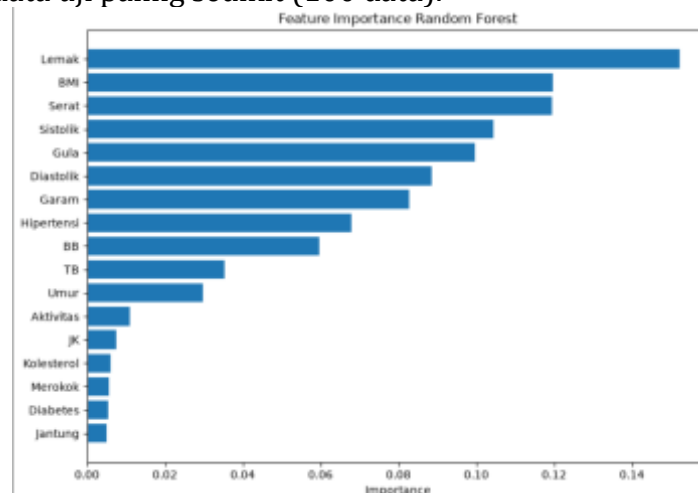
Kelas	Precision	Recall	F1-Score	Support
-------	-----------	--------	----------	---------

Berisiko	0,99	1,00	0,99	474
Sangat Berisiko	1,00	0,97	0,98	100
Aman	0,99	0,99	0,99	162
Accuracy	-	-	0,99	736
Macro Average	0,99	0,98	0,99	736
Weighted Average	0,99	0,99	0,99	736

Tabel 6. Hasil 5-Fold Cross Validation

Fold	Accuracy
1	0,998
2	0,995
3	0,997
4	0,997
5	0,994
Rata-rata	0,996 (99,6%)

Rincian kinerja model pada masing-masing kelas ditampilkan melalui classification report pada Tabel 5. Nilai macro average (0,98–0,99) yang hampir setara dengan nilai weighted average (0,99) mengindikasikan bahwa performa model relatif konsisten pada kelas mayoritas maupun minoritas, dengan bantuan parameter `class_weight=balanced` dalam menjaga performa pada kelas Sangat Berisiko yang memiliki jumlah data uji paling sedikit (100 data).



Gambar 2. Feature Importance Random Forest

Analisis tingkat kepentingan fitur (feature importance) bawaan Random Forest divisualisasikan pada Gambar 2. Fitur konsumsi lemak menunjukkan tingkat kepentingan tertinggi, diikuti BMI dan konsumsi serat, sejalan dengan indikator klinis seperti tekanan darah sistolik dan diastolik. Sebagaimana didiskusikan pada Metode Penelitian, hasil ini turut dipengaruhi oleh keterlibatan langsung variabel-variabel tersebut dalam skema pembentukan skor risiko (Strobl et al., 2007), sehingga tingginya feature importance perlu dimaknai sebagai indikasi kesesuaian model dengan aturan pelabelan, bukan murni sebagai bukti kekuatan hubungan biologis independen.

Interpretasi SHAP (Explainable AI)

Untuk memberikan interpretasi yang lebih mendalam terhadap keputusan model Random Forest, penelitian ini menerapkan metode SHAP menggunakan TreeExplainer. Nilai kepentingan fitur secara global diperoleh dengan menghitung rata-rata nilai absolut SHAP pada seluruh data dan seluruh kelas, sebagaimana disajikan pada Tabel 7.

Tabel 7. Nilai Mean |SHAP value| per Fitur (Global, Seluruh Kelas)

Fitur	Mean SHAP value
Serat	0,1101
Lemak	0,1023
BMI	0,1007
Diastolik	0,0882
Hipertensi	0,0853
Sistolik	0,0852
Berat Badan (BB)	0,0292
Kolesterol	0,0205
Tinggi Badan (TB)	0,0086
Garam	0,0044

Berdasarkan Tabel 7, fitur Serat menunjukkan kontribusi terbesar (mean |SHAP value| = 0,1101), diikuti Lemak (0,1023) dan BMI (0,1007), dengan Diastolik, Hipertensi, dan Sistolik juga menunjukkan kontribusi yang cukup besar. Hasil ini konsisten dengan feature importance bawaan Random Forest, memperkuat bahwa kedua metode menangkap pola yang sama dalam data — namun sebagaimana ditegaskan berulang kali pada bagian sebelumnya, konsistensi ini juga sejalan dengan fakta bahwa keenam fitur tersebut merupakan komponen langsung penyusun skor label, sehingga konsistensi antar-metode ini pertama-tama membuktikan konsistensi internal model terhadap skema skor, dan baru secara sekunder dapat dibaca sebagai indikasi relevansi klinis fitur tersebut.

Dari sudut pandang literatur biomedis, temuan bahwa serat dan lemak menonjol sejalan dengan sejumlah studi epidemiologis independen (yang tidak menggunakan skema pelabelan seperti pada penelitian ini). Meta-analisis studi kohort prospektif melaporkan hubungan dosis-respons terbalik antara asupan serat dan risiko stroke, dengan penurunan risiko sekitar 7–12% untuk setiap peningkatan konsumsi serat 7–10 gram per hari (Threapleton et al., 2013; Zhang et al., 2013; Chen et al., 2013). Secara mekanistik, asupan serat yang tinggi diketahui menurunkan kadar kolesterol LDL, memperbaiki sensitivitas insulin, dan menurunkan tekanan darah, sehingga menghambat proses aterosklerosis; sebaliknya, konsumsi lemak jenuh yang tinggi meningkatkan kadar kolesterol LDL dan memicu pembentukan plak aterosklerosis yang dapat menyumbat pembuluh darah otak (Syauqy et al., 2023). Studi berbasis populasi Indonesia juga melaporkan bahwa konsumsi makanan berlemak tinggi merupakan salah satu faktor risiko stroke terbesar pada dewasa, melalui mekanisme peningkatan kolesterol total dan penurunan elastisitas pembuluh darah (Syauqy et al., 2023). Dengan demikian, meskipun besaran kontribusi fitur pada model ini sebagian dibentuk oleh struktur skor label, arah hubungan yang ditangkap SHAP — serat sebagai faktor protektif dan lemak sebagai faktor risiko — tetap selaras secara kualitatif dengan bukti biomedis independen, sementara peran tekanan darah dan hipertensi sebagai prediktor kuat juga konsisten dengan studi INTERSTROKE yang menempatkan hipertensi sebagai kontributor tunggal terbesar risiko stroke secara global (O'Donnell et al., 2016). Hal ini menunjukkan bahwa meskipun akurasi numerik model dipengaruhi oleh circular reasoning pada skema pelabelan, arah dan peringkat relatif fitur yang teridentifikasi SHAP tidak sepenuhnya bertentangan dengan pengetahuan medis yang ada.

Tabel 8. Nilai Mean |SHAP value| per Fitur — Kelas Sangat Berisiko

Fitur	Mean SHAP value (Sangat Berisiko)
-------	-------------------------------------

Lemak	0,1325
BMI	0,1101
Diastolik	0,1064
Sistolik	0,0998
Serat	0,0799
Hipertensi	0,0566

Pada kelas Sangat Berisiko secara khusus, urutan kepentingan fitur berubah dengan Lemak menjadi paling dominan, diikuti BMI, Diastolik, Sistolik, Serat, dan Hipertensi (Tabel 8), sejalan dengan kombinasi ambang skor yang digunakan pada tahap pelabelan (tekanan darah tinggi, BMI ≥ 25 , konsumsi lemak tinggi, dan konsumsi serat rendah). Selain interpretasi global, SHAP juga digunakan untuk menjelaskan keputusan model pada level individu melalui waterfall plot dan force plot, yang memungkinkan setiap hasil klasifikasi ditelusuri alasannya secara spesifik per pasien.

PEMBAHASAN

Model Random Forest yang dikembangkan pada penelitian ini menunjukkan performa klasifikasi yang sangat tinggi (99,05% pada data uji dan 99,6% pada validasi silang) dan stabil pada ketiga kategori risiko. Namun, sebagaimana ditegaskan pada Metode Penelitian, tingginya akurasi ini sebagian besar merupakan konsekuensi dari skema pelabelan berbasis aturan yang menggunakan variabel prediktor yang sama, sehingga angka tersebut sebaiknya dimaknai sebagai bukti bahwa model berhasil mempelajari fungsi skor secara konsisten, dan bukan sebagai bukti kemampuan prediksi klinis independen terhadap kejadian stroke aktual.

Meskipun demikian, integrasi metode SHAP memberikan nilai tambah dalam mengonfirmasi bahwa keputusan model didasarkan pada fitur-fitur yang secara substansi relevan dengan literatur risiko stroke — asupan serat, asupan lemak, tekanan darah, BMI, dan status hipertensi (O'Donnell et al., 2016; Threapleton et al., 2013; Syauqy et al., 2023; Zhang et al., 2013) — dan bukan pada pola data yang bersifat kebetulan (spurious correlation) atau artefak encoding yang tidak bermakna secara klinis. Temuan ini memiliki implikasi praktis: model semacam ini berpotensi dikembangkan menjadi alat bantu skrining awal yang transparan di fasilitas pelayanan kesehatan primer seperti Puskesmas, membantu tenaga kesehatan mengidentifikasi pasien dengan kombinasi faktor risiko nutrisi dan klinis tertentu untuk diprioritaskan dalam edukasi gizi atau rujukan lanjutan, dengan tetap menempatkan keputusan klinis akhir pada tenaga medis yang berwenang.

Penelitian ini memiliki sejumlah keterbatasan yang perlu ditindaklanjuti. Pertama, label risiko dibentuk melalui skema skor berbasis aturan yang menimbulkan potensi circular reasoning terhadap performa model, sebagaimana telah dibahas secara rinci pada bagian Metode Penelitian dan sepanjang bagian ini. Kedua, ambang batas skor belum divalidasi oleh dokter spesialis saraf atau dibandingkan dengan instrumen baku seperti Framingham Stroke Risk Profile (Wolf et al., 1991). Ketiga, penanganan missing value melalui listwise deletion terhadap lebih dari 20% data berpotensi menimbulkan selection bias yang belum diuji melalui teknik imputasi pembandingan (Sterne et al., 2009; Beretta & Santaniello, 2016). Keempat, konfigurasi hyperparameter Random Forest belum melalui proses penalaan sistematis untuk mengurangi risiko overfitting pada data hasil skema skor. Kelima, desain studi tunggal-lokasi di Puskesmas Kanjilo membatasi generalisasi model terhadap populasi dan fasilitas kesehatan lain. Oleh karena itu, penelitian lanjutan yang melibatkan kolaborasi dengan praktisi kesehatan

atau ahli saraf untuk validasi lapangan, penggunaan label diagnosis klinis aktual, penerapan teknik imputasi data, penalaan hyperparameter, serta pengujian pada dataset multi-fasilitas sangat diperlukan sebelum model ini dapat dipertimbangkan untuk mendukung pengambilan keputusan klinis yang sesungguhnya.

KESIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma Random Forest untuk mengklasifikasikan tingkat risiko stroke ke dalam tiga kategori (Aman, Berisiko, dan Sangat Berisiko) berdasarkan data pola konsumsi nutrisi dan indikator kesehatan pasien Puskesmas Kanjilo, serta menerapkan metode SHAP untuk menginterpretasikan kontribusi setiap fitur terhadap hasil klasifikasi secara kuantitatif dan transparan.

Model yang dibangun menunjukkan performa klasifikasi yang tinggi dan konsisten pada pengujian data uji maupun validasi silang, dengan fitur konsumsi serat, lemak, BMI, tekanan darah, dan status hipertensi teridentifikasi sebagai kontributor utama, konsisten antara feature importance Random Forest dan nilai SHAP, serta selaras secara kualitatif dengan literatur epidemiologis mengenai faktor risiko stroke.

Perlu ditegaskan bahwa performa numerik yang sangat tinggi pada penelitian ini sebagian besar mencerminkan konsistensi model terhadap skema pelabelan berbasis aturan yang digunakan, bukan bukti independen atas kemampuan prediksi klinis. Oleh karena itu, penerapan model ini sebagai alat bantu skrining di fasilitas kesehatan primer memerlukan kolaborasi lebih lanjut dengan tenaga medis dan ahli saraf untuk validasi lapangan, penggunaan label diagnosis klinis aktual, serta pengujian pada data multi-fasilitas sebelum dapat diadopsi dalam praktik klinis yang sesungguhnya.

Pernyataan mengenai Penggunaan Kecerdasan Buatan (AI)

AI digunakan dalam penyusunan naskah ini. Penulis menggunakan perangkat kecerdasan buatan semata-mata untuk membantu perbaikan bahasa, pemeriksaan tata bahasa, serta penyuntingan demi kejelasan dan keterbacaan. Penulis tetap bertanggung jawab penuh atas akurasi, orisinalitas, integritas, dan konten akademis naskah tersebut. Seluruh konten yang dibantu oleh AI telah ditinjau dan diverifikasi oleh penulis sebelum naskah diserahkan.

Konflik Kepentingan

Penulis menyatakan tidak ada konflik kepentingan.

Kontribusi Penulis

Konseptualisasi, pengumpulan data, validasi data, analisis formal, penulisan – penyusunan draf awal, penulisan – peninjauan dan penyuntingan, serta visualisasi dilakukan oleh Muh Fadhil Ahmad. Seluruh penulis telah membaca dan menyetujui versi akhir naskah ini.

Ucapan Terima Kasih

“Penelitian ini didukung oleh Puskesmas Kanjilo, Kecamatan Barombong, Kabupaten Gowa dan Program Studi Informatika, Fakultas Teknik, Universitas Muhammadiyah Makassar. Penulis mengucapkan terima kasih atas dukungan, fasilitas, dan sumber daya yang diberikan untuk memfasilitasi penyelesaian penelitian ini.”

DAFTAR PUSTAKA

- Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1), Article 310.
- American Diabetes Association. (2022). Standards of medical care in diabetes—2022. *Diabetes Care*, 45(Suppl. 1), S1–S264.
- Athallah, D., Fathoni, F., & Rachmad, M. I. F. (2025). Klasifikasi risiko stroke menggunakan algoritma Random Forest dengan teknik Knowledge Discovery in Database. *Indonesian Journal of Computer Science*, 4(1), 38–44.
- Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16(Suppl. 3), Article 74.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Bzdok, D., Krzywinski, M., & Altman, N. (2018). Machine learning: Supervised methods. *Nature Methods*, 15(1), 5–6.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chen, G. C., Lv, D. B., Pang, Z., Dong, J. Y., & Liu, Q. F. (2013). Dietary fiber intake and stroke risk: A meta-analysis of prospective cohort studies. *European Journal of Clinical Nutrition*, 67(1), 96–100.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
- Dritsas, E., & Trigka, M. (2022). Stroke risk prediction with machine learning techniques. *Applied Sciences*, 12(9), Article 4670.
- Feigin, V. L., et al. (2021). Global, regional, and national burden of stroke and its risk factors, 1990–2019. *The Lancet Neurology*, 20(10), 795–820.
- Feigin, V. L., et al. (2025). World Stroke Organization: Global stroke fact sheet 2025. *International Journal of Stroke*, 20(2), 132–144.
- Johnson, W., et al. (2016). Stroke: A global response is needed. *Bulletin of the World Health Organization*, 94(9), 634–634A.
- Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30).
- Kernan, W. N., et al. (2013). Guidelines for the prevention of stroke in patients with stroke and transient ischemic attack. *Stroke*, 45(7), 2160–2236.
- Kumar, V., & Kaur, A. (2021). Stroke prediction using machine learning algorithms. *Journal of Healthcare Engineering*, 2021, 1–9.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30).
- Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- O'Donnell, M. J., et al. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): A case-control study. *The Lancet*, 388(10046), 761–775.
- Pratiwi, A., & Rahmayani, D. (2021). Hubungan pola makan dengan kejadian stroke pada pasien di Indonesia. *Jurnal Kesehatan Masyarakat*.

- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.
- Sacco, R. L., et al. (2013). An updated definition of stroke for the 21st century. *Stroke*, 44(7), 2064–2089.
- Sari, D. P., & Hidayat, T. (2022). Faktor risiko hipertensi terhadap kejadian stroke pada pasien rumah sakit di Indonesia. *Jurnal Ilmu Kesehatan*, 10(1), 45–52.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.
- Sterne, J. A. C., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 338, Article b2393.
- Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1), Article 25.
- Syauqy, A., Wiragapa, L. R., Soekatri, M. Y. E., Ernawati, F., Nissa, C., & Dieny, F. F. (2023). Hubungan antara pola makan dan kondisi penyerta dengan prevalensi stroke pada usia dewasa di Indonesia: Analisis data Riskesdas 2018. *Gizi Indonesia*, 46(1), 121–132.
- Tadi, P., & Lui, F. (2022). *Ischemic stroke*. StatPearls Publishing.
- Threapleton, D. E., et al. (2013). Dietary fiber intake and risk of first stroke: A systematic review and meta-analysis. *Stroke*, 44(5), 1360–1368.
- Unnithan, A. K. A., et al. (2022). *Hemorrhagic stroke*. StatPearls Publishing.
- Widiani, G. A. R., & Yasa, I. M. M. (2023). Korelasi tingkat pengetahuan terhadap kemampuan deteksi dini gejala stroke dengan sikap keluarga terhadap penanganan pre hospital. *Bina Generasi Jurnal Kesehatan*, 2(14), 25–30.
- Wolf, P. A., D'Agostino, R. B., Belanger, A. J., & Kannel, W. B. (1991). Probability of stroke: A risk profile from the Framingham Study. *Stroke*, 22(3), 312–318.
- World Health Organization. (2023). *Noncommunicable diseases [Fact sheet]*. WHO.
- Zhang, Z., Xu, G., Liu, D., Zhu, W., Fan, X., & Liu, X. (2013). Dietary fiber consumption and risk of stroke. *European Journal of Epidemiology*, 28(2), 119–130.