Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

Analisis Metode Ensemble Berbasis Random Forest untuk Klasifikasi Kejadian Stroke pada Dataset Publik

Viki Mei Adi Saputra[1,*], Mula Agung Barata[1], Denny Nurdiansyah[2,]
[1]Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Nahdlatul Ulama Sunan Giri, Bojonegoro, Indonesia
[2]Statistika, Fakultas Sains dan Teknologi ,Universitas Nahdlatul Ulama Sunan Giri, Bojonegoro , Indonesia
Email: [1,*]vikimeiadisaputra04@gmail.com, [2]mula.ab26@gmail.com, [3]denny.nur@unugiri.ac.id
Email Penulis Korespondensi: vikimeiadisaputra04@gmail.com[*]
Submitted: **24/01/2026**; Accepted: **19/02/2026**; Published: **31/03/2026**

Abstrak– Stroke merupakan salah satu penyebab utama disabilitas dan kematian *global*, sehingga diperlukan pendekatan berbasis data untuk mendukung klasifikasi kejadian stroke secara sistematis. Penelitian ini menganalisis variasi metode *ensemble* berbasis *Random Forest* pada dataset publik *healthcare-dataset-stroke-data* dari *Kaggle* yang terdiri dari 5.110 data pasien dengan 11 variabel demografis dan faktor risiko kardiovaskular. Tahapan prapemrosesan meliputi imputasi nilai hilang pada atribut *bmi* menggunakan *median*, penanganan *outlier* dengan metode *interquartile range* (IQR), serta penyeimbangan kelas menggunakan *SMOTE*. Tiga skenario model dikembangkan dalam satu pipeline yang seragam, yaitu *Random Forest* sebagai *baseline*, *Bagging Random Forest*, dan *AdaBoost Random Forest*. Evaluasi dilakukan menggunakan *5-Fold Cross Validation* dengan metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil analisis menunjukkan adanya perbedaan nilai metrik evaluasi antar skema ensemble, dengan konfigurasi AdaBoost Random Forest menghasilkan nilai akurasi sebesar 94,70% pada konfigurasi pengujian yang digunakan. Studi ini memfokuskan analisis pada variasi strategi ensemble dalam satu kerangka Random Forest dengan pipeline prapemrosesan yang seragam, sehingga menghasilkan evaluasi yang terkontrol dan reprodusibel.

Kata Kunci: Random Forest; Ensemble Learning; Klasifikasi Stroke; SMOTE; Healthcare Dataset

Abstract– Stroke remains one of the leading causes of global disability and mortality, highlighting the need for data-driven approaches to support systematic stroke event classification. This study analyzes variations of Random Forest-based ensemble methods using the public healthcare-dataset-stroke-data from Kaggle, which consists of 5,110 patient records with 11 demographic and cardiovascular risk variables. The preprocessing stages include median imputation for missing values in the bmi attribute, outlier handling using the interquartile range (IQR) method, and class balancing through SMOTE. Three model scenarios were developed within a unified pipeline framework: baseline Random Forest, Bagging Random Forest, and AdaBoost Random Forest. Model evaluation was conducted using 5-Fold Cross Validation with accuracy, precision, recall, and F1-score as performance metrics. The analysis results indicate differences in evaluation metric values across the ensemble schemes, with the AdaBoost Random Forest configuration achieving an accuracy of 94.70% under the applied testing setup. This study focuses on analyzing variations of ensemble strategies within a single Random Forest framework using a uniform preprocessing pipeline, thereby ensuring a controlled and reproducible evaluation.

Keywords: Random Forest; Ensemble Learning; Stroke Classification; SMOTE; Healthcare Dataset

1. PENDAHULUAN

Stroke tetap menjadi salah satu masalah kesehatan global dengan kontribusi besar terhadap angka disabilitas dan kematian. Laporan *Heart Disease and Stroke Statistics—2025* Update dari *American Heart Association* menunjukkan bahwa setiap tahun terdapat sekitar 15 juta kasus stroke di dunia, dengan hampir 6 juta di antaranya berujung pada kematian, sehingga menempatkan stroke sebagai salah satu penyebab utama kematian *global* [1]. Meskipun berbagai strategi pencegahan dan penatalaksanaan telah dikembangkan, tantangan dalam deteksi dini dan pengendalian faktor risiko masih terjadi, terutama di negara berkembang. Di Indonesia, *Profil Kesehatan Indonesia 2023* menempatkan stroke sebagai fokus utama pengendalian penyakit tidak menular, dengan penekanan pada pencegahan faktor risiko dan peningkatan deteksi dini di fasilitas pelayanan kesehatan [2]. Namun, keterbatasan identifikasi dini dan variasi karakteristik klinis pasien menyebabkan stroke sering terdeteksi pada kondisi yang telah berkembang lebih lanjut. Keragaman faktor risiko yang melibatkan usia, tekanan darah, kadar glukosa, indeks massa tubuh, serta riwayat penyakit kardiovaskular menunjukkan bahwa stroke bersifat multifaktorial. Oleh karena itu, diperlukan pendekatan berbasis data melalui pemanfaatan *machine learning* pada dataset publik dengan variabel klinis terstruktur untuk mendukung klasifikasi kejadian stroke secara lebih sistematis dan terukur. *Machine learning* merupakan pendekatan komputasional yang memungkinkan sistem mempelajari pola dari data dan membangun model prediktif tanpa bergantung pada aturan eksplisit yang ditentukan sebelumnya [3].

Dalam studi klasifikasi kejadian stroke pada dataset publik berbasis variabel klinis terstruktur, *Random Forest* merupakan salah satu algoritma *ensemble* berbasis pohon keputusan yang banyak digunakan pada data tabular karena mampu mengurangi varians melalui mekanisme *bootstrap sampling* dan pemilihan fitur secara acak pada setiap proses pemisahan *node* [4]. Pendekatan ini menghasilkan model yang relatif stabil serta tetap menyediakan informasi pentingnya fitur untuk keperluan interpretasi. Penelitian oleh Akinwumi dkk. (2025) dan El-Geneedy dkk. (2025) menunjukkan bahwa *Random Forest* memberikan kinerja yang konsisten [5],[6]. Studi komparatif oleh Imran dkk. (2022) juga menempatkan *Random Forest* sebagai *baseline* yang kuat, sementara variasi *ensemble* seperti *boosting* dapat meningkatkan performa pada konfigurasi tertentu [7]. Selain itu, Amarya

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

dkk. (2025) melaporkan bahwa *Random Forest* menunjukkan performa yang stabil dan kompetitif pada klasifikasi stroke, dan Fulazzaky dkk. (2024) menegaskan bahwa strategi penyeimbangan kelas dalam skema *ensemble* dapat meningkatkan sensitivitas terhadap kelas minoritas pada data yang tidak seimbang [8].

Penelitian oleh Kokkotis dkk. (2022) menunjukkan bahwa performa klasifikasi stroke yang tinggi tidak hanya ditentukan oleh pemilihan algoritma, tetapi juga oleh rancangan *pipeline* yang mencakup penanganan *missing value* dan penyeimbangan kelas. Temuan tersebut menegaskan bahwa desain prapemrosesan dan strategi evaluasi memiliki peran krusial dalam meningkatkan kinerja model klasifikasi pada data stroke yang tidak seimbang. Oleh karena itu, penelitian ini menekankan konsistensi rancangan prapemrosesan dan skema evaluasi dalam membandingkan metode *ensembl*e berbasis *Random Forest* [9]. Ushasree dkk. (2024) melaporkan bahwa pendekatan *ensemble* yang dikombinasikan dengan teknik penyeimbangan kelas mampu meningkatkan kinerja klasifikasi stroke pada data tidak seimbang [10]. El-Geneedy dkk. (2025) menunjukkan bahwa integrasi beberapa *classifier* dengan strategi pengelolaan distribusi kelas dapat meningkatkan sensitivitas sekaligus menjaga interpretabilitas model [6]. Astari dkk. (2025) menegaskan bahwa penerapan teknik sampling dalam skema *ensemble* berbasis pohon keputusan efektif meningkatkan *balanced accuracy* pada kondisi ketidakseimbangan ekstrem [11]. Fulazzaky dkk. (2024) juga menunjukkan bahwa variasi *ensemble* berbasis *Random Forest* mampu mengurangi bias terhadap kelas mayoritas dan meningkatkan stabilitas performa [8]. Secara umum, berbagai penelitian menunjukkan potensi pendekatan *ensemble* berbasis *Random Forest* dalam klasifikasi stroke berbasis data tabular.

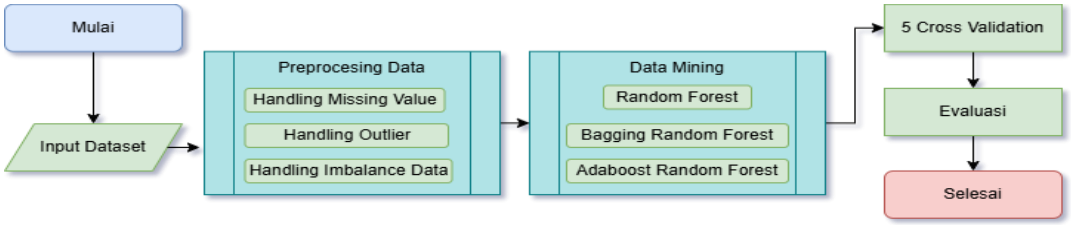
Sejumlah penelitian menunjukkan bahwa pendekatan berbasis pohon keputusan dan variasi *ensemble* memberikan performa yang stabil pada klasifikasi stroke berbasis data tabular publik. Sahriar dkk. (2024), Kitova dkk. (2024), dan Sitompul dkk. (2025) melaporkan bahwa metode *ensemble* berbasis pohon, termasuk *Random Forest* dan variasinya, cenderung lebih konsisten dibandingkan model tunggal maupun linear [12],[13],[14]. Pada data yang tidak seimbang, Erlin dkk. (2022) menunjukkan bahwa penerapan *SMOTE* meningkatkan sensitivitas kelas minoritas tanpa penurunan performa yang signifikan [15]. Sementara Fachruddin dkk. (2024) menegaskan bahwa *Random Forest* tetap kompetitif dan stabil pada klasifikasi stroke [16]. Laili dkk. (2023), Barata dkk. (2021), Barata dkk. (2024), serta Yaqin dkk. (2025) menunjukkan bahwa penerapan *machine learning*, khususnya melalui pendekatan *ensemble* berbasis pohon keputusan, berkontribusi terhadap peningkatan stabilitas dan akurasi model klasifikasi pada data kesehatan tabular [17],[18],[19],[20]. Temuan-temuan tersebut menunjukkan konsistensi performa pendekatan ensemble berbasis pohon keputusan pada klasifikasi stroke berbasis data tabular. Meskipun berbagai penelitian telah membandingkan beragam algoritma *machine learning* untuk prediksi stroke, sebagian besar studi tersebut belum secara khusus memfokuskan analisis pada konfigurasi strategi *ensemble* dalam satu kerangka model dengan pipeline prapemrosesan yang seragam. Oleh karena itu, penelitian ini memfokuskan analisis pada konfigurasi strategi *ensemble* dalam kerangka *Random Forest* dengan pipeline prapemrosesan yang konsisten guna memastikan evaluasi yang terkontrol dan reprodusibel.

Berdasarkan latar belakang dan tinjauan penelitian terdahulu tersebut, penelitian ini bertujuan untuk menganalisis performa metode *ensemble* berbasis *Random Forest* yang terdiri atas *Random Forest* dasar, *Bagging Random Forest*, dan *AdaBoost Random Forest* dalam klasifikasi kejadian stroke pada dataset publik menggunakan satu pipeline prapemrosesan yang konsisten. Secara metodologis, penelitian ini menekankan evaluasi yang terkontrol untuk menggambarkan karakteristik kinerja masing-masing konfigurasi *ensemble* dalam konteks data tabular yang tidak seimbang. Hasil penelitian ini diharapkan dapat memberikan gambaran empiris mengenai karakteristik performa strategi *ensemble* berbasis *Random Forest* pada klasifikasi kejadian stroke berbasis data kesehatan publik.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Adapun tahapan penelitian ini meliputi: (1) *input dataset* stroke sebagai data masukan, (2) *preprocessing* data yang mencakup penanganan *missing value*, penanganan *outlier*, dan penanganan ketidakseimbangan data, (3) proses *data mining* menggunakan tiga skenario model yaitu *Random Forest*, *Random Forest* dengan *Bagging*, serta *AdaBoost* dengan *Random Forest*, (4) validasi model menggunakan *5-fold cross validation*, dan (5) evaluasi kinerja model hingga tahap selesai. Alur lengkap skema tahapan penelitian tersaji pada Gambar 1.



Gambar 1. Tahapan Penelitian

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

2.2 Input Dataseta

Data yang digunakan dalam penelitian ini merupakan dataset publik *"healthcare-dataset-stroke-data.csv"* yang diperoleh dari platform *Kaggle*. Dataset ini tidak berasal dari rekam medis rumah sakit atau fasilitas layanan kesehatan tertentu, sehingga penelitian difokuskan pada evaluasi metodologis model klasifikasi kejadian stroke menggunakan data tabular publik.Dataset ini terdiri dari 12 atribut, yaitu 11 variabel independen dan 1 variabel dependen berupa label kejadian stroke. Variabel independennya meliputi id, jenis kelamin *(gender)*, umur *(age)*, riwayat hipertensi *(hypertension)*, riwayat penyakit jantung *(heart_disease)*, status perkawinan *(ever_married)*, jenis pekerjaan *(work_type)*, tipe tempat tinggal *(Residence_type)*, kadar rata-rata glukosa darah *(avg_glucose_level)*, indeks massa tubuh *(bmi)*, serta status merokok *(smoking_status)*; sedangkan variabel dependennya adalah atribut stroke yang menyatakan apakah pasien mengalami stroke (1) atau tidak (0). Pemilihan dataset ini sejalan dengan berbagai penelitian yang memanfaatkan data stroke klinis terstruktur dengan komposisi fitur demografis dan faktor risiko *kardiovaskular* yang serupa untuk pengembangan model klasifikasi berbasis *ensemble* pada penyakit stroke [12],[13],[14]. Rincian masing-masing fitur yang digunakan dalam penelitian tersaji pada Tabel 1.

Tabel 1. Dataset

No	SEX	age	hypertension	HD	EM	W_type	R_type	avg_glu_lvl	bmi	SS	stroke
0	0	67.0	0	1	0	0	0	228.69	36.6	2	1
1	1	61.0	0	0	0	1	1	202.21	NaN	0	1
2	0	80.0	0	1	0	0	1	105.92	32.5	0	1
3	1	49.0	0	0	0	0	0	171.23	34.4	3	1
4	1	79.0	1	0	0	1	1	174.12	24.0	0	1
...	...	...	...	...	...	...	...	...	...	...	...
5105	1	80.0	1	0	0	0	0	83.75	NaN	0	0
5106	1	81.0	0	0	0	1	0	125.20	40.0	0	0
5107	1	35.0	0	0	0	1	1	82.99	30.6	0	0
5108	0	51.0	0	0	0	0	1	166.29	25.6	2	0
5109	1	44.0	0	0	0	3	0	85.28	26.2	1	0

2.3 Preprocessing Data

Setelah dataset dimasukkan, tahapan berikutnya adalah *Preprocessing* Data. Pada tahap ini dilakukan serangkaian prosedur yang mencakup penanganan *missing value*, penanganan *outlier*, dan penanganan ketidakseimbangan data. Penanganan *missing value* dilakukan dengan mengidentifikasi atribut yang memiliki nilai hilang dan menerapkan teknik *imputasi* yang sesuai agar tidak menimbulkan bias dan tidak terlalu mengurangi jumlah sampel. Penanganan *outlier* difokuskan pada atribut *numerik* (seperti usia, tekanan darah, atau kadar glukosa) dengan cara mengidentifikasi nilai-nilai ekstrem yang berpotensi mengganggu pembentukan batas keputusan model, kemudian menyesuaikannya melalui *capping*, *transformasi*, atau mekanisme lain yang tetap rasional secara klinis. Selanjutnya, ketidakseimbangan kelas antara kasus stroke dan non-stroke ditangani melalui teknik penyeimbangan data (misalnya *oversampling* kelas *minoritas* atau pendekatan lain yang selaras dengan rancangan eksperimen), sehingga model tidak hanya terpaku pada kelas mayoritas dan menjadi lebih sensitif terhadap kasus stroke. Selain tiga tahapan utama tersebut, dilakukan pula penyesuaian pada fitur kategorikal, khususnya atribut *gender*. Pada dataset awal terdapat kategori *"Other"* dengan jumlah kasus satu *other* jumlah yang jauh lebih sedikit dibandingkan dengan kategori *"Male"* dan *"Female"*, sehingga berpotensi menimbulkan *sparsity* ketika dilakukan proses pengkodean *(encoding)* dan pelatihan model. Untuk menyederhanakan representasi data dan menjaga kestabilan estimasi, kategori *"Other"* digabungkan ke dalam kelas *"Female"* sehingga atribut *gender* hanya terdiri dari dua kelas utama. Hill dkk. (2023) menunjukkan bahwa penggabungan kategori langka pada variabel kategorikal merupakan bagian penting dari tahapan *data preprocessing* dalam pemodelan *prognostik* berbasis *machine learning* karena dapat meningkatkan konsistensi performa model pada data klinis [21].

2.4 Data Mining

Tahap berikutnya adalah *Data Mining* sebagaimana digambarkan pada Gambar 1. Pada tahap ini, data yang telah melalui *preprocessing* digunakan untuk membangun tiga skenario model klasifikasi. Skenario pertama menggunakan *Random Forest* sebagai model dasar *(baseline)* karena kemampuannya yang baik dalam menangani data tabular klinis, mengurangi *overfitting* melalui mekanisme *bagging*, serta menyediakan *feature importance* untuk keperluan interpretasi. Secara matematis, keputusan akhir *Random Forest* (termasuk varian *Random Forest Bagging)* dapat dinyatakan sebagai *voting mayoritas* dari himpunan pohon keputusan:

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{modeh_b}(\boldsymbol{x})_{\boldsymbol{b=1}}^{\boldsymbol{B}} \quad (\mathbf{1})$$

Dimana :

$\hat{y}(x)$: *Prediksi akhir (output) model Random Forest untuk sampel* x.
x : *Vektor fitur satu data pasien (umur, tekanan darah, gula darah, dsb.).*
$h_b(x)$: *Prediksi pohon keputusan ke-b dalam Random Forest untuk sampel* x.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

B : *Jumlah total pohon dalam ensemble Random Forest.*
$modehb(x)b = 1B$: *Memilih kelas yang paling sering diprediksi (voting mayoritas).*

Skenario kedua memanfaatkan *Random Forest* yang dipadukan dengan pendekatan *bagging* secara *eksplisit* untuk menguji sejauh mana penguatan proses *bootstrap aggregating* dapat meningkatkan stabilitas dan akurasi model dibanding *Random Forest* dasar. Konsep matematisnya tetap mengikuti Persamaan (1), namun setiap pohon dilatih pada *bootstrap sample* yang dibangkitkan secara *eksplisit* sesuai desain skenario bagging.Skenario ketiga menempatkan *Random Forest* sebagai *base learner* dalam kerangka *boosting (AdaBoost)* untuk mengevaluasi apakah penekanan bobot pada contoh-contoh yang sulit diklasifikasikan mampu memberikan kinerja deteksi stroke yang lebih baik daripada *Random Forest* tunggal maupun varian dengan *bagging.* Fungsi keputusan pada skema *AdaBoost Random Forest* dapat dituliskan sebagai kombinasi terboboti dari beberapa *base learner*:

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{2}$$

Dengan prediksi akhir :

$$\hat{y}(x) = sign\big(F(x)\big) \tag{3}$$

Dimana:

x : *Vektor fitur satu sampel (data pasien).*
$ht(x)$: *Prediksi base learner ke-t (Random Forest ke-t).*
α_t : *Bobot base learner ke-t makin bagus model, makin besar αt.*
T : *Jumlah iterasi boosting / jumlah base learner dalam ensemble.*
$F(x)$: *Skor gabungan terboboti $\sum_{t=1}^{T} \alpha_t h_t(x)$, masih berupa nilai kontinu.*
$\hat{y}(x)$: *Prediksi kelas akhir, diperoleh dari tanda nilai $F(x)$.*
$sig\, n(\cdot)$: *Fungsi tanda yang menentukan kelas berdasarkan apakah $F(x)$ positif atau negatif.*

Secara ringkas, Persamaan (2)–(3) menyatakan bahwa prediksi akhir *AdaBoost Random* Forest diperoleh dari kombinasi terboboti banyak *Random Forest,* di mana model yang lebih akurat diberi bobot lebih besar dalam penentuan keputusan akhir. Seluruh skenario model pada tahap data *mining* kemudian dilakukan pengujian menggunakan skema *k-fold cross validation.* Dataset dibagi menjadi lima lipatan *(fold*), di mana pada setiap iterasi empat *fold* digunakan sebagai data latih dan satu *fold* sebagai data uji. Proses ini diulang hingga setiap *fold* pernah berperan sebagai data uji, sehingga diperoleh estimasi kinerja yang lebih stabil dan tidak terlalu bergantung pada satu pembagian data tertentu, sebagaimana umum diterapkan pada evaluasi model klasifikasi berbasis *machine learning* [18]. Dengan demikian, perbandingan kinerja antara *Random Forest, Random Forest* dengan *Bagging,* dan *AdaBoost* dengan *Random Forest* menjadi lebih adil dan *reliabel.*

2.5 Evaluasi

Tahap terakhir pada Gambar 1 adalah Evaluasi dan Selesai. Pada tahap evaluasi, hasil prediksi dari masing-masing model dihitung metrik kinerjanya, seperti akurasi, presisi, recall, dan F1-score, kemudian dianalisis menggunakan *confusion matrix* untuk menilai kemampuan model dalam membedakan kelas stroke dan non-stroke, sebagaimana umum diterapkan pada evaluasi algoritma klasifikasi berbasis *machine learning* [22]. Kinerja ketiga skenario model dibandingkan untuk menentukan konfigurasi metode *ensemble* berbasis *Random Forest* yang paling efektif. Tahap Selesai mencakup penarikan kesimpulan, perumusan *implikasi* terhadap sistem deteksi dini stroke, serta penyusunan rekomendasi pengembangan lebih lanjut agar model yang dihasilkan dapat diintegrasikan ke dalam praktik klinis maupun sistem pendukung keputusan di layanan kesehatan.

3. HASIL DAN PEMBAHASAN

3.1 Praprocesing Data

Langkah kedua di tahap *preprocessing* data adalah melakukan pengecekan kualitas data pada dataset secara berurutan mulai dari missing value, outlier, hingga ketidakseimbangan kelas *(imbalance data)*. Pertama, dilakukan identifikasi nilai hilang *(missing value),* data yang terduplikasi, serta konsistensi tipe data pada setiap fitur. Pada tahap ini dilakukan pemeriksaan jumlah baris duplikat dan distribusi nilai kosong untuk masing-masing atribut. Hasil pengecekan menunjukkan bahwa tidak terdapat data yang terduplikasi, namun fitur *bmi* memiliki sejumlah nilai yang hilang sehingga perlu dilakukan proses *imputasi* agar tidak mengurangi jumlah sampel yang digunakan dalam pemodelan. Kedua, dilakukan analisis *outlier* pada fitur-fitur numerik seperti age, *avg_glucose*_level, dan *bmi* untuk mengidentifikasi nilai-nilai ekstrem yang berpotensi memengaruhi pembentukan batas keputusan model. Ketiga, diperiksa distribusi kelas pada variabel target stroke untuk menilai tingkat ketidakseimbangan data sebagai dasar penerapan teknik penyeimbangan data pada langkah berikutnya . Ringkasan jumlah data yang hilang pada setiap fitur.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

Tabel 2. Penanganan Missing Value

Atribut	
gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	201
smoking_status	0
stroke	0

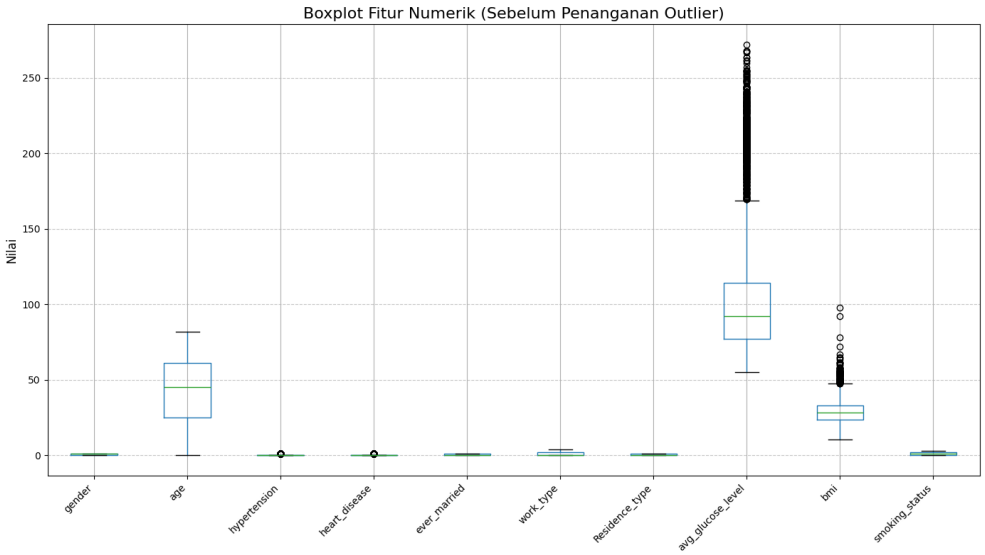
Dari total 5110 record pada dataset, kolom *bmi* memiliki 201 nilai yang hilang *(missing)*. *Missing value* pada fitur *bmi* tersebut kemudian ditangani dengan *imputasi* menggunakan nilai tengah *(median)* dari kolom *bmi*, sehingga tidak ada baris data yang harus dihapus. Selain itu, kolom *id* dihilangkan karena hanya berperan sebagai identitas unik dan tidak memberikan informasi prediktif terhadap kejadian stroke. Hasil dari langkah *penanganan missing value* dan penghapusan kolom *id* tersebut tersaji pada Tabel 3.

Tabel 3. Penanganan Missing Value

Atribut	
gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	0
smoking_status	0
stroke	0

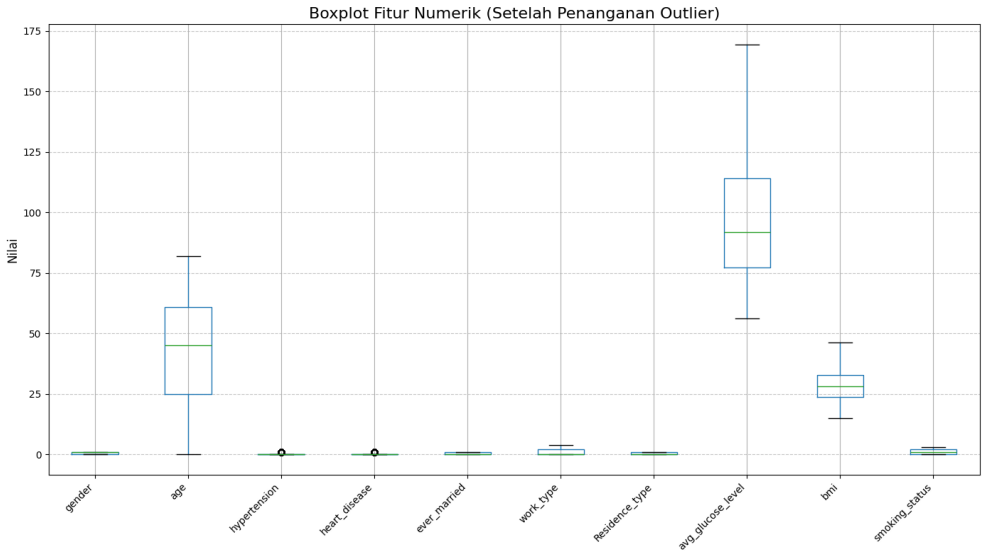
Di tahapan *outlier* bisa di lihat pada Gambar 2 ditampilkan *boxplot* seluruh fitur *numerik* sebelum dilakukan penanganan *outlier*. Terlihat bahwa sebagian besar fitur *biner* hasil pengkodean kategorikal *(gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status)* memiliki sebaran yang relatif sempit dan tidak menunjukkan nilai ekstrem yang mencolok. Sebaliknya, dua fitur kontinu utama, yaitu *avg_glucose_level* dan *bmi*, memperlihatkan banyak titik data yang berada jauh di atas rentang utama kotak, menandakan keberadaan *outlier* pada ekor atas distribusi. Pada *avg_glucose_level* tampak deretan titik yang memanjang hingga mendekati nilai sekitar 270, sementara pada *bmi* terdapat beberapa nilai di atas kisaran 50–90, yang secara klinis maupun statistik berpotensi mendistorsi estimasi parameter dan mempengaruhi pembentukan batas keputusan model klasifikasi.Untuk mencegah dominasi nilai ekstrem tersebut dalam proses pemodelan, *outlier* pada fitur *avg_glucose_level* dan *bmi* kemudian ditangani menggunakan pendekatan *interquartile range (IQR)*. Nilai di luar batas bawah dan batas atas yang diturunkan dari *Q1* dan *Q3* tidak dihapus, tetapi disesuaikan *(capping)* ke nilai ambang yang masih dianggap wajar. Pendekatan ini menjaga ukuran sampel tetap utuh sekaligus mengurangi pengaruh nilai yang terlalu jauh dari pola umum data, sehingga model *ensemble* berbasis *Random Forest* yang dibangun pada tahap berikutnya dapat distribusi data yang lebih representatif dan stabil.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496



Gambar 2. Boxplot Sebelum Penanganan Outlier

Pada Gambar 3 ditampilkan *boxplot* fitur-fitur *numerik* setelah dilakukan penanganan *outlier* menggunakan pendekatan *IQR* dengan *capping* pada batas bawah dan batas atas. Tampak bahwa sebaran nilai pada dua fitur kontinu utama, yaitu *avg_glucose_level* dan *bmi*, menjadi jauh lebih kompak dibandingkan sebelum perlakuan: *whisker* tidak lagi memanjang sangat jauh ke atas dan jumlah titik *ekstrem* di luar rentang utama hampir tidak terlihat. Rentang *interkuartil* kedua fitur tersebut masih mencerminkan variasi klinis yang wajar, tetapi tanpa dominasi nilai yang sangat tinggi yang berpotensi menarik garis keputusan model ke arah yang tidak representatif. Sementara itu, pola sebaran fitur lain seperti *age*, *hypertension*, *heart_disease*, dan variabel hasil pengkodean kategorikal relatif tetap stabil, menandakan bahwa prosedur penanganan *outlier* memang difokuskan pada atribut yang benar-benar memiliki nilai ekstrem. Secara keseluruhan, perubahan bentuk *boxplot* setelah penanganan *outlier* menunjukkan bahwa distribusi data *numerik* menjadi lebih terkontrol dan seimbang.



Gambar 3. Boxplot Setelah Penanganan Outlier

Pada tahap *handling imbalance data*, terlebih dahulu dianalisis distribusi kelas pada variabel target stroke. Hasil perhitungan menunjukkan bahwa pada dataset asli terdapat 4.861 data pasien tidak stroke (kelas 0) dan hanya 249 data pasien stroke (kelas 1), sehingga rasio kelas mendekati 95:5 dan berpotensi menimbulkan bias model terhadap kelas *mayoritas*. Untuk mengatasi ketidakseimbangan kelas, digunakan teknik *SMOTE* dengan cara menghasilkan sampel sintetis pada kelas minoritas berdasarkan tetangga terdekat di ruang fitur. Setelah proses resampling dengan *SMOTE*, distribusi kelas menjadi seimbang, yaitu 4.861 data untuk kelas 0 dan 4.861 data untuk kelas 1, sehingga model *ensemble* berbasis *Random Forest* yang dilatih memiliki peluang lebih besar untuk mempelajari pola kasus stroke tanpa mengabaikan kelas minoritas sebagaimana tersaji pada Tabel 4. Pendekatan ini sejalan dengan temuan Fulazzaky dkk. [8] yang menunjukkan bahwa kombinasi teknik penyeimbangan kelas dan metode *ensemble* berbasis pohon keputusan dapat meningkatkan stabilitas dan sensitivitas model pada data yang tidak seimbang.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

Tabel 4. Hasil Handling Imbalance Data

ukuran dataset asli:		stroke
0	4861	
1	249	
name:	count, dtype:	int64
1	4861	
0	4861	
name:	count, dtype:	int64

3.1.1 Data Mining

Setelah seluruh tahapan *preprocessing* data yang meliputi penanganan *missing value*, analisis *outlier*, dan penyeimbangan kelas dengan *SMOTE* selesai dilakukan dan diperoleh dataset akhir yang bersih serta seimbang, langkah berikutnya adalah tahap *data mining* dengan membangun tiga skenario model klasifikasi, yaitu *Random Forest*, *Bagging Random Forest*, dan *AdaBoost Random Forest*. Pada tahap inilah setiap model dilatih dan diuji dengan skema *k-fold cross validation* dengan nilai $k = 5$, sehingga diperoleh lima skenario pelatihan dan pengujian dari data yang berbeda. Hasil dari 5 skenario *Random Forest* tersaji pada Tabel 5.

Tabel 5. Hasil Random Forest Basline

Fold	Accuracy	Precision	Recall	F1_Score
1	0.927378	0.928368	0.927378	0.927348
2	0.940231	0.940408	0.940231	0.940207
3	0.949196	0.949633	0.949196	0.949169
4	0.942765	0.943789	0.942765	0.942742
5	0.922186	0.922940	0.922186	0.922196
Rata-Rata	0.936351	0.937028	0.936351	0.936333

Rata-rata pengujian *Random Forest* dengan *5-fold cross-validation* menghasilkan akurasi 93,64%, presisi 93,70%, recall 93,64%, dan F1-score 93,63%, sebagaimana ditunjukkan pada Tabel 5. Sebagai variasi terhadap model *Random Forest* tunggal, selanjutnya dilakukan pemodelan menggunakan algoritma *Bagging Random Forest* dengan skema validasi silang berstrata 5-fold yang sama seperti pada baseline. Hasil evaluasi kinerja Bagging *Random Forest*, meliputi akurasi, presisi, recall, dan F1-score pada setiap *fold* beserta nilai rata-ratanya, tersaji pada Tabel 6.

Tabel 6. Hasil Bagging Random Forest

Fold	Accuracy	Precision	Recall	F1_Score
1	0.918380	0.920365	0.918380	0.918304
2	0.926093	0.926301	0.926093	0.926058
3	0.936977	0.938083	0.936977	0.936908
4	0.933762	0.935817	0.933762	0.933701
5	0.915756	0.917693	0.915756	0.915738
Rata-Rata	0.926194	0.927652	0.926194	0.926142

Rata-rata pengujian *Bagging Random Forest* dengan *5-fold cross-validation* menghasilkan akurasi 92,62%, presisi 92,77%, recall 92,62%, dan F1-score 92,61%, sebagaimana ditunjukkan pada Tabel 6. Sebagai variasi *ensemble* lain yang memberikan bobot lebih besar pada sampel-sampel yang sulit diklasifikasikan, selanjutnya dilakukan pemodelan menggunakan algoritma *AdaBoost* dengan *Random Forest* sebagai *base learner* menggunakan skema validasi silang berstrata *5-fold* yang sama seperti pada dua model sebelumnya. Hasil evaluasi kinerja *AdaBoost Random Forest*, yang meliputi akurasi, presisi, recall, dan F1-score pada setiap fold beserta nilai rata-ratanya, tersaji pada Tabel 7.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

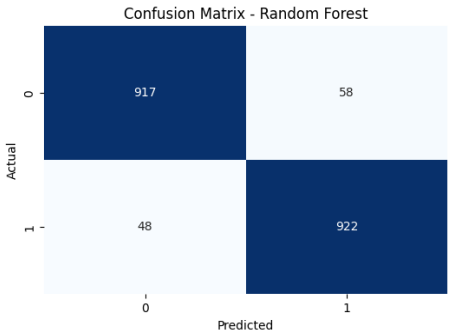
Tabel 7. Hasil Adaboost Random Forest

Fold	Accuracy	Precision	Recall	F1_Score
1	0.930591	0.931642	0.930591	0.930561
2	0.942802	0.943031	0.942802	0.942776
3	0.949839	0.950393	0.949839	0.949807
4	0.939550	0.940513	0.939550	0.939527
5	0.931833	0.932638	0.931833	0.931841
Rata-Rata	0.938923	0.939644	0.938923	0.938902

Rata-rata pengujian *AdaBoost Random Forest* dengan *5-fold cross-validation* menghasilkan akurasi 93,89%, presisi 93,96%, recall 93,89%, dan F1-score 93,89%, sebagaimana ditunjukkan pada Tabel 7.

3.1.2. Evaluasi

Setelah ketiga skenario model, yaitu *Random Forest*, *Bagging Random Forest*, dan *AdaBoost Random Forest*, diuji menggunakan skema *5-fold cross validation*, dilakukan evaluasi lanjutan dengan menggunakan *confusion matrix* untuk melihat pola klasifikasi secara lebih rinci. Pada tahapan ini dibandingkan *confusion matrix* dari *fold-1* hasil uji *5-fold cross validation* sebagai representasi kinerja masing-masing model; *confusion matrix fold-1* untuk model *Random Forest* disajikan pada Gambar 4 dan menjadi dasar pembahasan mengenai kemampuan model dalam membedakan kelas stroke dan non-stroke.



Gambar 4. Confusion Matrix Random Forest

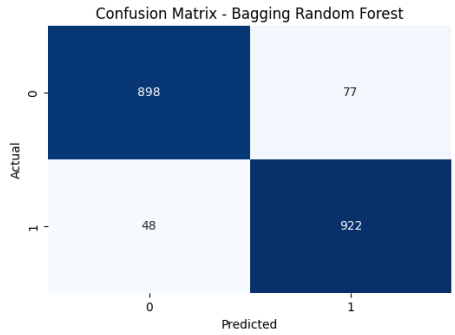
Gambar 4 menunjukkan *confusion matrix* model *Random Forest* pada klasifikasi stroke. Model mengklasifikasikan sebagian besar data dengan benar, dengan 917 data non-stroke dan 922 data stroke diprediksi tepat, serta jumlah kesalahan yang relatif rendah. Rendahnya *false negative* menunjukkan bahwa model memiliki sensitivitas yang baik terhadap kelas stroke secara metodologis, sehingga meminimalkan kasus yang terlewatkan dalam proses klasifikasi.

Tabel 8. Hasil Evaluasi Random Forest

Kelas	Accuracy	Precision	Recall	F1 Score
Tidak	0.9405	0.9503	0.9405	0.9454
Ya	0.9505	0.9408	0.9505	0.9456
Rata-Rata	0.9455	0.9455	0.9455	0.9455

Tabel 8 menunjukkan bahwa *model Random Forest* memiliki kinerja yang tinggi dan seimbang pada klasifikasi stroke, dengan *accuracy* rata-rata sebesar 0,9455. Nilai *precision* yang tinggi pada kelas non-stroke (0,9503) serta *recall* yang tinggi pada kelas stroke (0,9505), yang tercermin pada nilai *F1-score* yang hampir identik, menegaskan *Random Forest* sebagai *baseline* yang kuat dalam evaluasi metodologis klasifikasi stroke.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496



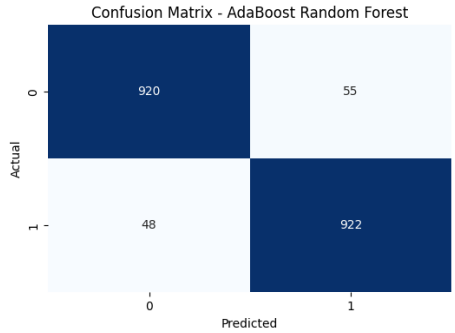
Gambar 5. Confusion Matrix Bagging Random Forest

Sebagai langkah kedua dalam analisis, Gambar 5 menampilkan *confusion matrix* model *Bagging Random Forest* pada klasifikasi stroke. Model mengklasifikasikan sebagian besar data dengan benar, dengan 898 data non-stroke dan 922 data stroke diprediksi tepat, namun menghasilkan *false positive* yang lebih tinggi (77 kasus) dibandingkan *Random Forest* dasar. Meskipun jumlah *false negative* tetap rendah (48 kasus), temuan ini menunjukkan bahwa penerapan skema *bagging* belum meningkatkan kemampuan diskriminasi kelas, sehingga performanya berada di bawah *Random Forest* dasar dalam evaluasi metodologis klasifikasi stroke.

Tabel 9. Hasil Evaluasi Bagging Random Forest

Kelas	Accuracy	Precision	Recall	F1 Score
Tidak	0.9210	0.9493	0.9210	0.9349
Ya	0.9505	0.9229	0.9505	0.9365
Rata-Rata	0.9357	0.9361	0.9358	0.9357

Tabel 9 menunjukkan kinerja model *Bagging Random Forest* pada klasifikasi stroke dengan *accuracy* rata-rata sebesar 0,9357, yang lebih rendah dibandingkan *Random Forest* dasar. Nilai *precision* yang tinggi pada kelas non-stroke (0,9493) dan *recall* yang tinggi pada kelas stroke (0,9505) tercermin pada *F1-score* rata-rata sebesar 0,9357, namun hasil ini menunjukkan bahwa penerapan skema *bagging* belum mampu meningkatkan kinerja model secara keseluruhan dibandingkan *Random Forest* dasar dalam evaluasi metodologis yang dilakukan.



Gambar 6. Confusion Matrix Adaboost Random Forest

Sebagai langkah terakhir dalam analisis, Gambar 6 menampilkan *confusion matrix* model *AdaBoost Random Forest* pada klasifikasi stroke. Model mengklasifikasikan sebagian besar data dengan benar, dengan 920 data non-stroke dan 922 data stroke diprediksi tepat, serta jumlah kesalahan yang relatif rendah. Dibandingkan dengan *Random Forest dasar* dan *Bagging Random Forest*, jumlah *false positive* yang lebih rendah (55 kasus) menunjukkan peningkatan kemampuan diskriminasi kelas, sehingga penerapan skema *boosting* memberikan kinerja yang lebih baik secara metodologis dalam klasifikasi stroke.

Tabel 10. Hasil Evaluasi Adaboost Random Forest

Kelas	Accuracy	Precision	Recall	F1 Score
Tidak	0.9436	0.9504	0.9436	0.9470
Ya	0.9505	0.9437	0.9505	0.9471
Rata-Rata	0.9470	0.9471	0.9471	0.9470

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

Tabel 10 menunjukkan bahwa model *AdaBoost Random Forest* menghasilkan nilai akurasi rata-rata sebesar 0,9470, dengan *precision* dan *recall* yang relatif seimbang pada kedua kelas, yang konsisten dengan nilai *F1-score* sebesar 0,9470. Temuan ini mengindikasikan bahwa penerapan skema *boosting* dalam kerangka *Random Forest* memberikan karakteristik kinerja yang stabil pada konfigurasi pengujian yang digunakan. Perbedaan nilai metrik antar pendekatan *ensemble* menunjukkan adanya perbedaan karakteristik kinerja dalam satu pipeline prapemrosesan yang seragam.

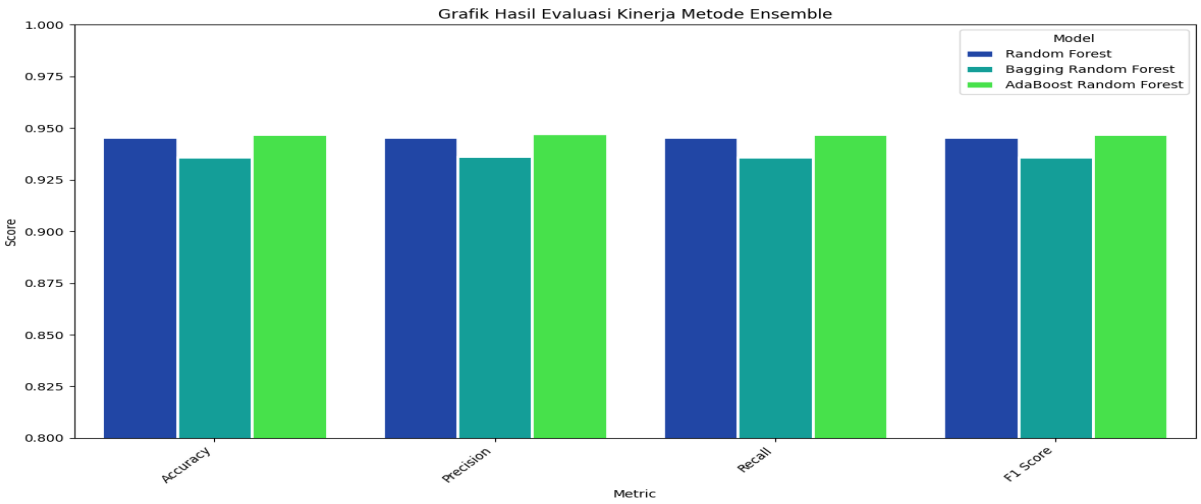
3.2 Pembahasan

Analisis kinerja pada tiga variasi metode *ensemble* berbasis *Random Forest*, yaitu *Random Forest* dasar, *Bagging Random Forest*, dan *AdaBoost Random Forest*, dilakukan melalui pengamatan nilai akurasi, presisi, *recall*, dan *F1-score*. Ringkasan hasil evaluasi ketiga konfigurasi model disajikan pada Tabel 11.

Tabel 11..Nilai Metrik Evaluasi pada Tiga Skema Ensemble

Model	Accuracy	Precision	Recall	F1_Score
RF	0.945501	0.945550	0.945501	0.945501
Bagging + RF	0.935733	0.936125	0.935733	0.935721
Adaboost + RF	0.947044	0.947068	0.947044	0.947044

Hasil pengujian pada tiga skema metode ensemble berbasis *Random Forest* menunjukkan adanya perbedaan nilai metrik evaluasi pada dataset stroke yang telah diseimbangkan menggunakan *SMOTE*. Konfigurasi *AdaBoost Random Forest* menghasilkan nilai akurasi sebesar 94,70%, diikuti *Random Forest* dasar sebesar 94,55% dan *Bagging Random Forest* sebesar 93,57%. Pola ini menunjukkan bahwa penerapan mekanisme *boosting* dalam kerangka *Random Forest* memberikan karakteristik kinerja yang berbeda pada konfigurasi pengujian yang digunakan. Temuan ini sejalan dengan penelitian Imran dkk. [7] yang melaporkan bahwa pendekatan berbasis *boosting* pada klasifikasi stroke menggunakan skema *k5-fold cross validation* menunjukkan akurasi sekitar 95,0%, sehingga mengindikasikan potensi peningkatan performa melalui mekanisme pembobotan bertahap terhadap kesalahan klasifikasi. Meskipun demikian, perbedaan konfigurasi dataset, komposisi fitur, serta tahapan prapemrosesan yang digunakan pada masing-masing penelitian dapat memengaruhi capaian nilai metrik evaluasi yang diperoleh Grafik perbandingan metrik evaluasi untuk ketiga skenario model tersebut ditunjukkan pada Gambar 7.



Gambar 7. Grafik Nilai Metrik Evaluasi pada Tiga Skema Ensemble

Gambar tersebut menunjukkan pola yang konsisten, di mana konfigurasi *AdaBoost Random Forest* menghasilkan nilai metrik evaluasi tertinggi pada skenario pengujian yang dilakukan. Hal ini mengindikasikan bahwa mekanisme *boosting* dalam kerangka *Random Forest* memberikan karakteristik kinerja yang berbeda melalui proses pembobotan pada sampel yang sulit diklasifikasikan. Temuan ini sejalan dengan penelitian Imran dkk.[7] yang melaporkan bahwa pendekatan berbasis *boosting* menunjukkan kinerja yang kompetitif dalam klasifikasi stroke pada skema validasi silang. Sementara itu, konfigurasi *Bagging Random Forest* memperlihatkan nilai metrik yang berbeda dalam konfigurasi yang diuji, yang menunjukkan adanya variasi karakteristik performa antar skema *ensemble* dalam satu pipeline prapemrosesan yang seragam.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

4. KESIMPULAN

Penelitian ini menyimpulkan bahwa pendekatan *ensemble* berbasis *Random Forest* menunjukkan karakteristik kinerja yang stabil dalam klasifikasi kejadian stroke pada dataset publik. Hasil eksperimen memperlihatkan adanya variasi nilai metrik evaluasi antar pendekatan, dengan konfigurasi *AdaBoost Random Forest* menghasilkan nilai metrik tertinggi dalam skenario pengujian yang dilakukan. Temuan ini menunjukkan bahwa penerapan strategi *boosting* dalam kerangka *Random Forest* menghasilkan karakteristik kinerja yang berbeda dibandingkan variasi *ensemble* lainnya pada data yang telah diseimbangkan. Kontribusi penelitian ini terletak pada penyajian analisis dalam satu alur prapemrosesan dan evaluasi yang seragam dan reprodusibel, sehingga memungkinkan pengamatan karakteristik kinerja masing-masing variasi *ensemble* secara terkontrol. Meskipun penelitian ini menggunakan satu dataset publik dan difokuskan pada satu keluarga algoritma *ensemble*, hasil yang diperoleh memberikan dasar metodologis bagi pengembangan dan pengujian lanjutan model klasifikasi stroke pada dataset dengan karakteristik berbeda.

REFERENCES

[1] S. S. Martin *et al.*, *2025 Heart Disease and Stroke Statistics: A Report of US and Global Data from the American Heart Association*, vol. 151, no. 8. 2025. doi: 10.1161/CIR.0000000000001303.

[2] P. K. Indonesia, *Profil Kesehatan*. Jalan HR. Rasuna Said Blok X-5 Kav 4-9, Jakarta 12950, 2024. [Online]. Available: http://www.kemkes/go.id

[3] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for Machine Learning*, First Edit. Cambridge University Press. doi: 10.1017/9781108679930.

[4] T. Dietterich, C. Bishop, D. Heckerman, M. Jordan, M. Kearns, and A. Editors, *Probabilistic machine learning*. Cambridge, Massachusetts :, [2022] Series: [Online]. Available: https://lccn.loc.gov/2021027430

[5] P. O. Akinwumi, S. Ojo, T. I. Nathaniel, J. Wanliss, O. Karunwi, and M. Sulaiman, "Evaluating machine learning models for stroke prediction based on clinical variables," *Front. Neurol.*, vol. 16, no. September, 2025, doi: 10.3389/fneur.2025.1668420.

[6] M. El-Geneedy, H. El-Din Moustafa, H. Khater, S. Abd-Elsamee, and S. A. Gamel, "A comprehensive explainable AI approach for enhancing transparency and interpretability in stroke prediction," *Sci. Rep.*, vol. 15, no. 1, pp. 1–23, 2025, doi: 10.1038/s41598-025-11263-9.

[7] B. Imran, E. Wahyudi, A. Subki, S. Salman, and A. Yani, "Classification of stroke patients using data mining with adaboost, decision tree and random forest models," *Ilk. J. Ilm.*, vol. 14, no. 3, pp. 218–228, 2022, doi: 10.33096/ilkom.v14i3.1328.218-228.

[8] T. Fulazzaky, A. Saefuddin, and A. M. Soleh, "Evaluating Ensemble Learning Techniques for Class Imbalance in Machine Learning: A Comparative Analysis of Balanced Random Forest, SMOTE-RF, SMOTEBoost, and RUSBoost," *Sci. J. Informatics*, vol. 11, no. 4, pp. 969–980, 2024, doi: 10.15294/sji.v11i4.15937.

[9] M. O. Ullah, S. A. Raju, M. I. Nazir, A. Akter, and M. S. Rahman, "An Innovative Machine Learning Pipeline for Stroke Prediction on Imbalanced Data," *2023 Int. Conf. Inf. Commun. Technol. Sustain. Dev. ICICT4SD 2023 - Proc.*, pp. 153–157, 2023, doi: 10.1109/ICICT4SD59951.2023.10303319.

[10] D. Ushasree, A. V. Praveen Krishna, and C. Mallikarjuna Rao, "Enhanced stroke prediction using stacking methodology (ESPESM) in intelligent sensors for aiding preemptive clinical diagnosis of brain stroke," *Meas. Sensors*, vol. 33, no. November 2023, p. 101108, 2024, doi: 10.1016/j.measen.2024.101108.

[11] R. A. Astari, I. M. Sumertajaya, and A. M. Soleh, "A Hybrid Sampling Approach for Handling Data Imbalance in Ensemble Learning Algorithms," *Sci. J. Informatics*, vol. 12, no. 2, pp. 247–258, 2025, doi: 10.15294/sji.v12i2.19163.

[12] S. Sahriar *et al.*, "Unlocking stroke prediction: Harnessing projection-based statistical feature extraction with ML algorithms," *Heliyon*, vol. 10, no. 5, p. e27411, 2024, doi: 10.1016/j.heliyon.2024.e27411.

[13] K. Kitova, I. Ivanov, and V. Hooper, "Stroke Dataset Modeling: Comparative Study of Machine Learning Classification Methods," *Algorithms*, vol. 17, no. 12, pp. 1–16, 2024, doi: 10.3390/a17120571.

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

[15] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.

[16] Y. P. Fachruddin1*, Errissya Rasywir2, "Random Forest Algorithm with Mutual Information Feature Selection," *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 8, no. 4, pp. 555–562, 2024, doi: https://doi.org/10.29207/resti.v8i4.5795 1.

[17] E. F. Laili *et al.*, "Komparasi Algoritma Decision Tree Dan Support Vector Machine (Svm) Dalam," *J. Sist. Inf. dan Inform.*, vol. 8, no. 1, pp. 67–76, 2025.

[18] M. Barata, "Improving the Accuracy of C4.5 Algorithm with Chi-Square Method on Pure Tea Classification Using Electronic Nose," *Resti*, vol. 1, no. 10, pp. 19–25, 2021.

[19] M. Barata, Dwi Irnawati, Ifnu Wisma Dwi Prastya, and Dwi Issadari Hastuti, "Hydrogen Sulfide Leak Detection Using the C4.5 Algorithm: Optimizing Feature Extraction for Enhanced Accuracy," *PROCEEDING AL GHAZALI Int. Conf.*, vol. 2, pp. 348–358, 2025, doi: 10.52802/aicp.v1i1.1352.

[20] A. A. Yaqin, M. A. Barata, and N. Mahmudah, "Implementation of the Random Forest Algorithm with Optuna Optimization in Lung Cancer Classification," *Sistemasi*, vol. 14, no. 2, p. 561, 2025, doi: 10.32520/stmsi.v14i2.4877.

[21] H. A. Hill *et al.*, "Integrative Prognostic Machine Learning Models in Mantle Cell Lymphoma," vol. 3, no. August, pp. 1435–1446, 2023, doi: 10.1158/2767-9764.CRC-23-0083.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9496

[22] E. Helmud, E. Helmud, and P. Romadiana, “Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix,” vol. 13, pp. 92–97, 2024.