Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

Evaluasi Seleksi Fitur Algoritma Genetika pada Klasifikasi Penyakit Stroke Menggunakan K-Nearest Neighbor pada Dataset Overlapping

Aqmal Syarif Fadilah, Siska Kurnia Gusti*, Iwan Iskandar, Iis Afrianty

Fakultas Sains dan Teknologi, Prodi Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]12250113564@students.uin-suska.ac.id, [2,*]siskakurniagusti@uin-suska.ac.id, [3]iwan.iskandar@uin-suska.ac.id, [4]iis.afrianty@uin-suska.ac.id
Email Penulis Korespondensi: siskakurniagusti@uin-suska.ac.id

Abstrak−Stroke merupakan penyakit gangguan fungsi otak akibat terhambatnya aliran darah dan membutuhkan deteksi dini untuk mengurangi risiko kecacatan maupun kematian. Namun, kompleksitas dataset penyakit stroke sering kali menghasilkan representasi data dengan tingkat tumpang tindih (overlapping) kelas yang tinggi, sehingga memicu bias dan menurunkan akurasi klasifikasi. Salah satu metode klasifikasi yang banyak digunakan dalam prediksi penyakit stroke adalah K-Nearest Neighbor (KNN), namun algoritma ini sensitif terhadap fitur tidak relevan dan distribusi data yang kompleks. Penelitian ini bertujuan mengevaluasi efektivitas seleksi fitur menggunakan Algoritma Genetika (GA) terhadap performa klasifikasi KNN pada dataset prediksi stroke yang memiliki karakteristik overlapping tinggi antar kelas. Dataset yang digunakan terdiri dari 15.000 data pasien dengan 21 atribut awal, yang setelah tahapan preprocessing berubah menjadi 38 atribut. Evaluasi model dilakukan menggunakan 10-Fold Stratified Cross Validation pada tiga rasio pembagian data, yaitu 90:10, 80:20, dan 70:30, dengan parameter K=13. Hasil penelitian menunjukkan bahwa GA efektif mereduksi dimensi data secara signifikan menjadi 15–17 fitur terpilih serta meningkatkan rata-rata nilai fitness validasi internal sekitar 3%. Namun demikian, evaluasi pada data uji menunjukkan bahwa performa model KNN maupun GA-KNN tetap terhenti pada rentang akurasi 48–52% dengan nilai ROC-AUC mendekati 0,500. Kondisi ini menunjukkan bahwa GA cenderung menemukan solusi lokal optimal yang sensitif terhadap pembagian data latih, sehingga model kesulitan melakukan generalisasi saat memproses data baru. Ekstremnya tingkat overlapping pada dataset ini dibuktikan lebih lanjut melalui eksperimen pembanding menggunakan rekayasa fitur turunan (Feature Engineering) dan algoritma Random Forest, yang seluruhnya ikut terhenti di kisaran akurasi 50%. Hasil tersebut menunjukkan bahwa seleksi fitur GA sukses menyederhanakan kompleksitas data, namun tidak secara otomatis mampu menyelesaikan masalah pemisahan kelas apabila tumpang tindih merupakan karakteristik bawaan dari data itu sendiri.

Kata Kunci: Algoritma Genetika; K-Nearest Neighbor; Overlapping Dataset; Seleksi Fitur; Stroke

Abstract−Stroke is a disease that disrupts brain function due to impaired blood flow and requires early detection to reduce the risk of disability and death. However, the complexity of stroke datasets often results in data representation with a high level of class overlap, thus triggering bias and reducing classification accuracy. One classification method widely used in stroke prediction is K-Nearest Neighbor (KNN), but this algorithm is sensitive to irrelevant features and complex data distribution. This study aims to evaluate the effectiveness of feature selection using Genetic Algorithms (GA) on the performance of KNN classification in stroke prediction datasets characterized by high overlap between classes. The dataset used consists of 15,000 patient data with 21 initial attributes, which after preprocessing stages are changed to 38 attributes. Model evaluation was carried out using 10-Fold Stratified Cross Validation at three data split ratios, namely 90:10, 80:20, and 70:30, with parameter K = 13. The results showed that GA effectively reduced the data dimensionality significantly to 15–17 selected features and increased the average internal validation fitness value by approximately 3%. However, evaluation of the test data showed that the performance of both the KNN and GA-KNN models remained stagnant at an accuracy range of 48–52%, with ROC-AUC values approaching 0.500. This condition indicates that GA tends to find local optimal solutions that are sensitive to the division of the training data, making it difficult for the model to generalize when processing new data. The extreme level of overlap in this dataset was further demonstrated through comparative experiments using feature engineering and the Random Forest algorithm, both of which also stalled at around 50% accuracy. These results indicate that GA feature selection successfully simplifies data complexity, but is not automatically capable of solving class separation problems when overlap is an inherent characteristic of the data itself.

Keywords: Genetic Algorithm; K-Nearest Neighbor; Overlapping Dataset; Stroke; Feature Selection

1. PENDAHULUAN

Stroke adalah penyakit yang mengganggu kerja fungsi otak sehingga mengakibatkan gangguan pada aliran darah menuju otak [1]. Stroke disebabkan oleh adanya sumbatan gumpalan darah sehingga mengganggu fungsi syaraf [2]. Stroke terbagi menjadi dua jenis yaitu *stroke iskemik* disebabkan oleh penyumbatan pembuluh darah dan *stroke hemoragik* disebabkan oleh pecahnya pembuluh darah [3]. Penyakit stroke merupakan penyebab kematian nomor tiga di dunia setelah penyakit jantung dan kanker, kemudian juga penyebab kecacatan nomor satu secara global [4]. Penyebab terjadinya stroke dipengaruhi beberapa faktor yang tidak dapat dirubah yaitu genetik, usia, cacat bawaan, dan keturunan. Sedangkan faktor yang dapat dirubah yaitu hipertensi, penyakit jantung, obesitas, merokok, konsumsi alkohol, dan kurangnya aktivitas [5]. Namun, dari banyaknya faktor penyebab stroke hipertensi merupakan faktor yang paling dominan dalam meningkatkan risiko terjadinya stroke [6]. Di indonesia data Riskesdas tahun 2018 penyakit stroke berada di peringkat kedua setelah penyakit jantung yang menyebabkan kematian [7]. Sehingga, deteksi dini penyakit stroke sangat penting karena gejala awal sering sekali tidak dapat dikenali [6]. Salah satu upaya untuk mendeteksi penyakit stroke adalah dengan menggunakan klasifikasi, klasifikasi yang akurat dapat membantu tenaga medis dalam mengambil keputusan dan tindakan yang tepat [8].

Salah satu algoritma klasifikasi yang sederhana namun kuat dan efektif dalam mengklasifikasikan data adalah algoritma K-Nearest Neighbor (KNN) [4]. Algoritma ini mengklasifikasikan data baru berdasarkan jarak dari

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

kemiripannya dengan data latih [9]. Penelitian oleh [10] melakukan perbandingan algoritma Naive Bayes dan KNN dalam mengklasifikasikan penyakit stroke dengan dataset yang berjumlah 5110 baris data rekam medis dengan 11 atribut. Dalam pengujiannya menunjukkan bahwa algoritma KNN secara konsisten memberikan hasil yang lebih baik dibandingkan dengan algoritma Naive Bayes pada berbagai rasio *split data*, dengan akurasi tertinggi mencapai 95,77% pada rasio split data 60:40. Penelitian lain oleh [4] menerpakan algoritma KNN untuk mengklasifikasikan risiko stroke dengan dataset berjumlah 5110 baris data rekam medis dan 11 atribut. Melalui pengujian berbagai nilai tetangga terdekat (K) dari 1 hingga 10 dengan metrik *Euclidean Distance*, penelitian ini menunjukkan bahwa KNN mampu mencapai akurasi tertinggi sebesar 93,54% pada nilai K=1. Ini menegaskan bahwa pemilihan parameter K yang tepat sangat krusial dalam memaksimalkan potensi KNN dan dalam mengenali pola data apabila dataset yang memiliki fitur yang cukup representatif terhadap label target.

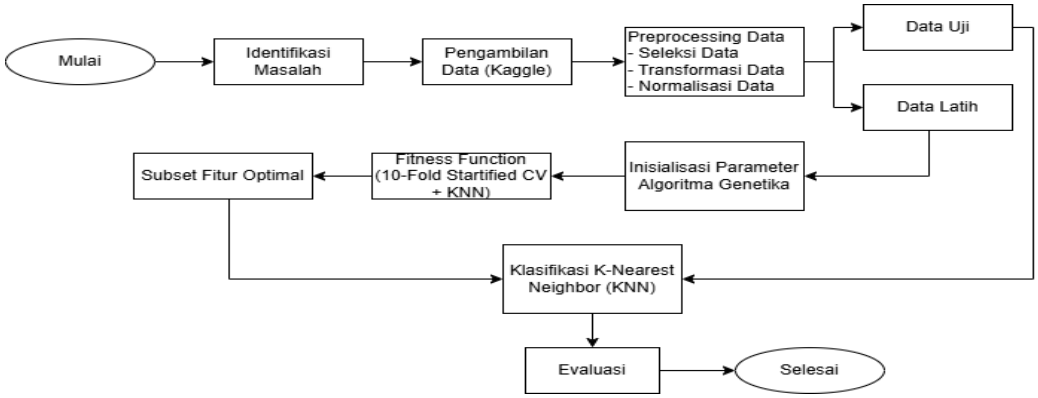
Algoritma KNN memiliki keunggulan dalam proses klasifikasi karena mampu bekerja secara efektif pada data dengan pola yang jelas. Namun, KNN juga memiliki kelemahan yaitu sangat sensitif terhadap fitur yang kurang relevan [11]. Selain itu, KNN juga sensitif terhadap distribusi data dan skala fitur, sehingga performa algoritma bergantung pada struktur data [12]. Oleh karena itu, untuk mengurangi penggunaan fitur yang kurang relevan dapat dilakukan degan teknik seleksi fitur [13]. Salah satu teknik seleksi fitur efektif dalam menangani permasalahan tersebut adalah Algoritma Genetika (GA). GA merupakan metode pencarian metaheuristik yang meniru proses evolusi alamiah untuk menemukan subset fitur paling optimal dalam ruang pencarian yang kompleks [14]. Dalam konteks kesehatan, penggunaan GA terbukti mampu meningkatkan performa klasifikasi dengan mengeliminasi fitur-fitur yang tidak relevan dan redundan [14]. Penelitian oleh [15] menunjukkan bahwa seleksi fitur berbasis GA efektif dalam memangkas variabel yang redundan pada dataset stroke. Dengan menyaring empat atribut yaitu *heart disease*, *ever married*, *work type*, dan *residence type* GA berhasil meningkatkan akurasi model dari 94,81% menjadi 95,13% pada data kaggle dan 93,30% menjadi 98,20% pada data mendeley. Penelitian lain oleh [12] mengoptimasi deteksi penyakit diabetes dengan hasil pengujian dari dataset *Pima Indians Diabetes*, GA berhasil meningkatkan akurasi dari 74,2% menjadi 79,1%. Sementara pada dataset *Early Stage Diabetes*, GA berhasil meningkatkan akurasi diperoleh adalah 98,2% dari 97,5%. Oleh karena itu, penerapan GA diharapkan dapat mereduksi dimensi data dan meminimalkan noise pada dataset stroke, sehingga menghasilkan model prediksi yang lebih efisien [16].

Meskipun berbagai penelitian sebelumnya membuktikan bahwa algoritma KNN maupun kombinasi GA-KNN mampu mencapai tingkat akurasi yang sangat tinggi hingga berkisar antara 93% sampai 98%, sebagian besar eksperimen tersebut dilakukan pada karakteristik dataset yang memiliki batas pemisahan kelas cenderung jelas. Perbedaan mendasar antara penelitian ini dengan penelitian-penelitian terdahulu terletak pada karakteristik internal dataset rekam medis yang digunakan. Penelitian terdahulu belum mengevaluasi performa kombinasi metode ini pada dataset stroke dengan tingkat overlapping yang sangat ekstrem. Pada kondisi overlapping yang tinggi, subset fitur yang disaring oleh Algoritma Genetika sering kali terjebak pada solusi lokal optimal yang hanya sensitif pada partisi data latih tertentu, sehingga performanya menurun atau mandek saat diuji menggunakan data baru. Analisis robustness check menggunakan algoritma pembanding non-jarak (Random Forest) serta rekayasa fitur (Feature Engineering) juga diperlukan untuk membuktikan apakah keterbatasan generalisasi model merupakan kegagalan algoritma atau memang karakteristik bawaan dari dataset medis itu sendiri. Berdasarkan permasalahan tersebut, tujuan dari penelitian ini adalah untuk mengevaluasi efektivitas Algoritma Genetika dalam melakukan seleksi fitur pada klasifikasi penyakit stroke menggunakan K-Nearest Neighbor serta menganalisis pengaruh karakteristik overlapping dataset terhadap kemampuan generalisasi model.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian akan melakukan beberapa tahapan seperti yang ada pada Gambar 1.



Gambar 1. Tahapan Penelitian

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

Pertama, dimulai dengan pengumpulan data, kemudian pemilihan data pada dataset, dan kemudian tahap preprocessing data yang meliputi beberapa proses yang dilakukan termasuk pembersihan data, transformasi data menggunakan Label Encoding dan One-Hot Encoding untuk data kategorikal, kemudian metode Min-Max Normalization untuk normalisasi. Selanjutnya, dilakukan pembagian data latih dan data uji yang akan divalidasi menggunakan K-fold Cross validation. Selanjutnya, pemilihan fitur menggunakan Algoritma Genetika, setelah itu klasifikasi menggunakan metode K-Nearest Neighbor. Dan selanjutnya evaluasi menggunakan Confusion Matrix. Sehingga kesimpulan dapat ditarik dalam penelitian ini.

2.2 Preprocessing

Preprocessing bertujuan untuk mengubah data mentah menjadi format yang lebih bersih dan siap digunakan dalam analisis maupun pemodelan. Pada tahap ini mencakup beberapa proses yaitu seleksi data dengan menghapus atribut *Patient ID, Patient Name, dan Symptoms* karena tidak memiliki kontribusi terhadap proses klasifikasi, sedangkan *Symptoms* dihapus karena formatnya teks bebas yang sulit diproses oleh model machine learning. Kemudian transformasi, dan normalisasi untuk membuat hasil analisis lebih akurat dan efisien [17]. Setelah proses seleksi, dataset menyisakan 18 atribut dan 1 label seperti pada Tabel 1.

Tabel 1. Atribut Hasil Seleksi

Atribut	Keterangan
Age	Usia
Gender	Jenis Kelamin
Hypertension	Riwayat Hipertensi
Heart Disease	Riwayat penyakit jantung
Marital Status	Status pernikahan
Work Type	Jenis Pekerjaan
Residence Type	Tipe tempat tinggal
Average Glucose Level	Rata-rata kadar glukosa
Body Mass Index (BMI)	Indeks masa tubuh
Smoking Status	Status merokok
Alcohol Intake	Tingkat aktivitas fisik
Physical Activity	Tingkat aktivitas fisik
Stroke History	Riwayat stroke sebelumnya
Family History of Stroke	Riwayat stroke dalam keluarga
Dietary Habits	Kebiasaan pola makan
Stress Levels	Tingkat stress
Blood Pressure Levels	Tekanan darah
Cholesterol Levels	Kadar kolesterol
Diagnosis	Label target

2.2.1 Transformasi Data

Transformasi data kategorikal dilakukan menggunakan dua metode encoding. Pertama, *Label Encoding* digunakan pada atribut kategorikal yang bersifat biner seperti *Gender, Residence Type,* dan *Family History of Stroke* dengan mengubah data ketegorikal menjadi bentuk numerik. Hal ini bertujuan agar algoritma machine learning dapat lebih efektif memahami variabel pada dataset [18]. Kedua, *One-Hot Encoding* digunakan pada atribut kategorikal dengan lebih dari dua kategori seperti *Marital Status, Work Type*, *Smoking Status*, *Alcohol Intake*, *Physical Activity*, dan *Dietary Habits.* Hal ini bertujuan untuk menghindari bias ordinal pada data nominal [19]. Selain encoding, dilakukan ekstrasi fitur terhadap dua atribut bertipe string yang tidak dapat langsung diproses oleh model. Yaitu atribut *Blood Pressure Levels* dan *Cholesterol Levels* seperti pada **Tabel 2**.

Tabel 2. Atribut Sebelum Ekstraksi Fitur

Atribut Asal	Format Asal
Blood Pressure Levels	174/81
Cholesterol Levels	HDL: 70, LDL: 137

Berdasarkan hasil ekstraksi fitur yang dilakukan, atribut asal akan dihapus dan menghasilkan atribut yang baru setelah dilakukan ekstrasi fitur seperti pada **Tabel 3**.

Tabel 3. Atribut Setelah Ekstraksi Fitur

Atribut Hasil	Format Hasil
Systolic	174
Diastolic	81
HDL	70
LDL	137

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

2.2.2 Normalisasi Data

Normalisasi data dilakukan dengan *Min-Max Normalization*, pada proses ini dilakukan transformasi linier pada atribut data asli dengan teknik normalisasi. Adapun persamaan normalisasi data dapat dilihat pada persamaan (1).

$$Xn = \frac{Xi - min\ (x)}{max\ (x) - min\ (x)} \quad (1)$$

Rumus (1) menjelaskan bahwa Xn merupakan nilai hasil dari nilai awal suatu titik data (Xi) sedangkan min(x) merupakan nilai minimum diseluruh data, dan max(x) merupakan nilai maksimum diseluruh data.

2.3 Pembagian Data dan Validasi

Tahap pembagian data dilakukan dengan memisahkan dataset menjadi dua bagian utama, yaitu data latih dan data uji. Penelitian ini menerapkan skenario *split data* dengan tiga variasi rasio untuk megevaluasi pengaruhnya terhadap performa model, yaitu 90:10, 80:20, 70:30. Prosedur pemisahan data uji secara permanen di awal proses berguna untuk menjaga validitas evaluasi [20]. Pembagian dilakukan secara *stratified* untuk memastikan distribusi kelas pada setiap subset tetap proporsional sesuai distribusi kelas pada dataset awal. Detail jumlah data pada setiap skenario ditunjukkan pada Tabel 4.

Tabel 4. Skenario Pembagian Data

Rasio Split	Jumlah Data Latih	Jumlah Data Uji
90:10	13.500	1.500
80:20	12.000	3.000
70:30	10.500	4.500

Kemudian, untuk mengevaluasi performa dan tingkat generalisasi model baseline maupun model optimasi pada data latih, dilakukan teknik 10-Fold Stratified Cross Validation. Pendekatan ini membagi data latih secara acak menjadi 10 subset berukuran sama, di mana proses pelatihan dan validasi internal diulang sebanyak 10 kali secara bergantian. [20]. Penelitian ini menggunakan nilai k=10 (*10-Fold Stratified Cross Validation*) yang diterapkan pada proses evaluasi model baseline KNN maupun pada evaluasi *fitness function* dalam Algoritma Genetika.

2.4 Analisis Distribusi dan Visualisasi Dataset

Analisis distribusi data dilakukan untuk memahami karakteristik internal dataset sebelum proses klasifikasi dilakukan. Proses ini bertujuan untuk mengidentifikasi pola persebaran data, hubungan antar kelas, serta potensi overlapping antara kelas stroke dan no stroke yang dapat mempengaruhi performa model klasifikasi. Visualisasi data menggunakan metode *Principal Component Analisis (PCA)* dan *t-distibuted Stochastic Neighbor Embedding (t-SNE)* dengan mereduksi dimensi data dalam ruang dua dimensi. PCA digunakan untuk memvisualisasikan distribusi data secara linear berdasarkan variansi terbesar, sedangkan t-SNE digunakan untuk memetakan hubungan non-linear dan mempertahankan kedekatan lokal antar data sehingga pola kluster dan overlapping antar kelas dapat diamati dengan lebih jelas [21].

2.5 Algoritma Genetika

Proses utama dalam algoritma genetika terdiri dari beberapa tahapan. Pertama, inisialisasi populasi, yaitu membangkitkan sejumlah kromosom secara acak sebagai solusi awal. Kedua, evaluasi fitness, yaitu menghitung nilai kebugaran setiap individu berdasarkan akurasi yang diperoleh dari 10-Fold Stratified Cross Validation menggunakan KNN. Ketiga, seleksi, yaitu memilih individu terbaik menggunakan metode *Tournament Selection*, di mana individu dengan nilai *fitness* tertinggi dipilih sebagai induk generasi berikutnya. Keempat, crossover, yaitu menggabungkan dua kromosom induk menggunakan metode *Two-Point Crossover*, di mana dua titik potong dipilih secara acak pada kromosom dan segmen di antaranya dipertukarkan untuk menghasilkan kromosom baru. Kelima, mutasi, yaitu membalik nilai gen secara acak dengan probabilitas tertentu untuk menjaga keberagaman populasi dan menghindari konvergensi dini. Proses evolusi diulang hingga mencapai jumlah generasi maksimum atau berhenti lebih awal apabila nilai *fitness* terbaik tidak mengalami peningkatan selama 15 generasi berturut-turut (*saturate* 15). Kriteria penghentian ini bertujuan untuk menghemat waktu komputasi tanpa mengorbankan kualitas solusi yang dihasilkan. Dalam penelitian ini Algoritma Genetika mengggunakan beberapa parameter seperti pada Tabel 5.

Tabel 5. Parameter Algoritma Genetika

Parameter	Nilai
Ukuran populasi	50 individu
Jumlah generasi	100 generasi
Jumlah Induk	25 induk
Metode seleksi	Tournament
Metode crossover	Two-point crossover
Probabilitas crossover	0.8
Metode mutase	Random
Probabilitas mutasi	0.1

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

Parameter	Nilai
Representasi kromosom	Array biner (0 atau 1)
Fitness function	Akurasi 10-fold stratified CV + KNN
Kriteria berhenti	Saturate 15 generasi

2.6 K-Nearest Neighbor

Algoritma K-Nearest Neighbor (KNN) mengklasifikasikan risiko berdasarkan kemiripan data baru dengan sejumlah *K* tetangga terdekat pada data latih. Dengan menghitung kemiripan menggunakan *Euclidean Distance* seperti pada persamaan berikut (2).

$$d = \sqrt{\sum_{i=1}^{n} (X_i - y_i)^2} \quad (2)$$

Dimana *d* adalah jarak euclidean, *n* adalah dimensi data atau jumlah atribut, *i=1* merupakan indeks atribut data ke=1, kemudian X_i adalah nilai data latih, sedangkan y_i adalah nilai data uji. Pengujian dilakukan secara sistematis menggunakan nilai *K* ganjil sampai menemukan tren konvergen untuk menemukan titik optimal sekaligus mengindari hasil seri dalam penentuan kelas mayoritas. Pemilihan nilai K yang tidak terlalu kecil maupun terlalu besar bertujuan untuk menjaga keseimbangan antara sensitivitas model terhadap noise dan kemampuan model dalam mempertahankan pola data. Nilai K yang terlalu kecil cenderung sensitif terhadap noise, sedangkan nilai K yang terlalu besar dapat menyebabkan batas antar kelas menjadi kurang spesifik.

2.7 Evaluasi Confusion Matrix

Dalam mengevaluasi kinerja model klasifikasi digunakan *Confusion Matrix*, yaitu bagian dari *Machine Learning* dengan mempelajari data yang ada dan mengelompokkannya sebagai data baru dengan memberikan hasil dengan variabel kategorikal nominal maupun ordinal. Dalam *Confusion Matrix* terdiri dari 4 istilah yang digunakan untuk mewakili proses hasil klasifikasi, yaitu TP (*True Positive*), FP (*False Positive*), FN (*False Negative*), TN (*True Negative*). Berdasarkan nilai pada *Confusion Matrix* tersebut, metrik evaluasi yang digunakan didefinisikan pada persamaan (3) sampai (6).

Akurasi $$\frac{(TP + TN)}{TP + TN + FP + FN} X100\% \quad (3)$$

Presisi $$\frac{TP}{TP + FP} X100\% \quad (4)$$

Recall $$\frac{TP}{TP + FN} X100\% \quad (5)$$

F1 Score $$\frac{2(Precission\ x\ Recall)}{(Precission + Recall} \quad (6)$$

a. Akurasi merupakan tolak ukur pada keakuratan input data dan output yang dihasilkan, dengan rasio prediksi yang benar terhadap keseluruhan data yang diuji.
b. Presisi adalah ukuran dari akurasi prediksi positif, yang merupakan rasio prediksi positif yang benar terhadap keseluruhan prediksi positif yang dibuat model.
c. Recall merupakan kemampuan model untuk mendeteksi data yang positif dengan benar, yaitu rasio data positif yang berhasil dikenali terhadap keseluruhan data positif yang sebenarnya.
d. F1-Score adalah rata-rata harmonis antara presisi dengan recall yang mencerminkan keseimbangan keduanya, F1-Score sangat berguna terutama pada data yang tidak seimbang.
e. ROC- AUC merupakan Kurva *Receiver Operating Characteristic* (ROC) menggambarkan kinerja klasifikasi dengan memplot *True Positive Rate* (sumbu Y) terhadap *False Positive Rate* (sumbu X). *Area Under Curve* (AUC) mengukur total luas di bawah kurva tersebut (rentang 0.0–1.0); semakin mendekati 1.0, semakin baik kemampuan model dalam membedakan antar kelas.
   1. 0.9-1.0 = Klasifikasi sangat baik.
   2. 0.8-0.9 = Klasifikasi baik.
   3. 0.7-0.8 = Klasifikasi rata-rata.
   4. 0.6-0.7 = Klasifikasi rendah.
   5. 0.5-0.6 = Kegagalan.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

Pada penelitian ini dataset yang digunakan adalah dataset prediksi stroke yang diperoleh dari Kaggle (https://www.kaggle.com/code/merfarukyce/stroke-prediction-analysis-eda). Dataset ini terdiri dari 15000 baris data dengan 21 atribut dan 1 label, kemudian dengan jumlah label Stroke 7468 dan No Stroke 7532. Atribut yang tersedia meliputi *patient id, patient name, age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), smoking status, alcohol intake, physical activity, stroke history, family history of stroke, dietary habits, stress levels, blood pressure levels, cholesterol levels, symptoms, diagnosis* (label). Dapat dilihat pada **Tabel 6**.

Tabel 6. Dataset Awal

Patient ID	Patient Name	Age	Gender	Hypertension	...	Cholesterol Levels	Sysmptoms	Diagnosis
18153	Mamooty Khurana	56	Male	0	...	HDL: 68, LDL: 133	Difficulty Speaking, Headache	Stroke
62749	Kaira Subramaniam	80	Male	0	...	HDL: 63, LDL: 70	Loss of Balance, Headache, Dizziness	Stroke
...	...	...	...	...	...	...	...	...
22343	Anvi Mannan	73	Male	0	...	HDL: 79, LDL: 91	Severe Fatigue, Numbness	No Stroke
11066	Gokul Trivedi	64	Female	0	...	HDL: 78, LDL: 179	Headache	Stroke

Implementasi sistem dilakukan dengan bahasa pemrograman python dan menggunakan Google Colab. Data tersebut kemudian dibagi menjadi data latih dan data uji dengan tiga variasi rasio pembagian, yaitu 90:10, 80:20, dan 70:30. Proses seleksi fitur dijalankan menggunakan Algoritma Genetika (GA) untuk mereduksi dimensi data dan mencari subset fitur optimal. Dalam proses ini, 10-Fold Cross Validation diterapkan pada data latih sebagai fungsi fitness untuk mengevaluasi setiap individu fitur yang dihasilkan oleh GA. Selanjutnya, algoritma K-Nearest Neighbor (KNN) digunakan sebagai metode klasifikasi dengan pengujian nilai K ganjil dari K = 3 sampai menunjukkan pola konvergen menggunakan perhitungan jarak euclidean. Tahap akhir mengevaluasi menggunakan metrix confusion matrix dan kurva ROC-AUC untuk menganalisis kemampuan model dalam membedakan kelas pada kondisi data yang memiliki tingkat tumpang tindih (*overlapping*) tinggi.

3.2 Hasil Preprocessing Data

Hasil preprocessing menunjukkan adanya perubahan pada struktur dataset setelah dilakukan seleksi atribut, transformasi data kategorikal, dan ekstraksi fitur. Dataset awal yang terdiri dari 21 atribut dan 1 label mengalami pengurangan atribut setelah proses seleksi data dengan menghapus atribut Patient ID, Patient Name, dan Symptoms. Selanjutnya, proses transformasi dan ekstraksi fitur menyebabkan jumlah atribut kembali bertambah akibat penerapan One-Hot Encoding dan pemecahan atribut bertipe string menjadi beberapa atribut numerik baru. Perubahan atribut hasil preprocessing ditunjukkan pada **Tabel 7.**

Tabel 7. Atribut Setelah Preprocessing

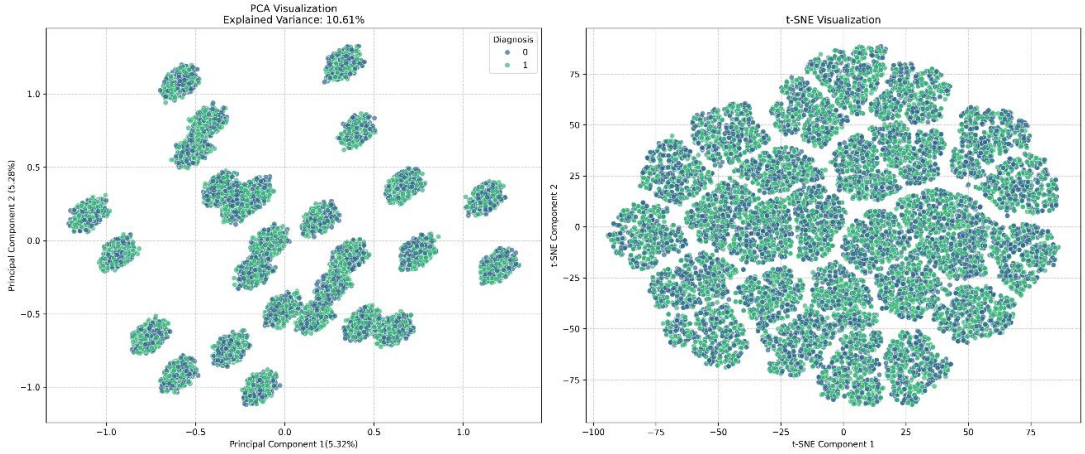
Tahapan	Jumlah Atribut	Hasil Atribut
Seleksi Data	18 atribut	Age, Gender, Hypertension, Heart Disease, Marital Status, Work Type, Residence Type, Average Glucose Level, Body Mass Index (BMI), Smoking Status, Alcohol Intake, Physical Activity, Stroke History, Family History of Stroke, Dietary Habits, Stress Levels, Blood Pressure Levels, Cholesterol Levels
Transformasi Data	38 atribut	Age, Gender, Hypertension, Heart Disease, Residence Type, Average Glucose Level, Body Mass Index (BMI), Stroke History, Family History of Stroke, Stress Levels, Systolic, Diastolic, HDL, LDL, Marital Status_Divorced, Marital Status_Married, Marital Status_Single, Work Type_Government Job, Work Type_Never Worked, Work Type_Private, Work Type_Self-employed, Smoking Status_Currently Smokes, Smoking Status_Formerly Smoked, Smoking Status_Non-smoker, Alcohol Intake_Frequent Drinker, Alcohol Intake_Never, Alcohol Intake_Rarely, Alcohol Intake_Social Drinker, Physical Activity_High, Physical Activity_Low, Physical Activity_Moderate, Dietary Habits_Gluten-Free, Dietary Habits_Keto, Dietary Habits_Non-Vegetarian, Dietary Habits_Paleo, Dietary Habits_Pescatarian, Dietary Habits_Vegan, Dietary Habits_Vegetarian

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

Peningkatan jumlah atribut terjadi karena beberapa atribut kategorikal diubah menjadi beberapa atribut biner menggunakan metode One-Hot Encoding. Selain itu, atribut Blood Pressure Levels dan Cholesterol Levels yang sebelumnya berbentuk string berhasil diekstraksi menjadi atribut numerik baru berupa Systolic, Diastolic, HDL, dan LDL sehingga data menjadi lebih representatif untuk proses klasifikasi. Dataset hasil preprocessing selanjutnya digunakan pada tahap pelatihan dan evaluasi model klasifikasi menggunakan algoritma K-Nearest Neighbor (KNN) dan seleksi fitur berbasis Algoritma Genetika.

3.3 Hasil Analisis Visualisasi Dataset

Berdasarkan visualisasi pada Gambar 2. Visualisasi PCA menjelasakan distribusi data antara kelas stroke dan no stroke masih menunjukkan overlap yang cukup tinggi sehingga pemisahan kelas secara linear belum terlihat cukup jelas. Nilai total explained variance sebesar 10,6% menunjukkan bahwa dua komponen utama PCA hanya mampu merepresentasikan sebagian kecil informasi dari keseluruhan dataset, yang artinya struktur data bersifat kompleks dan tersebar pada banyak dimensi. Sementara itu, pada t-SNE menunjukkan pembentukan kluster lokal yang lebih jelas dibandingkan PCA karena t-SNE mampu mempertahankan hubungan non linear antar data. Namun demikian, overlap antara kelas stroke dan no stroke masih terlihat pada sebagian besar area visualisasi. Hal ini menunjukkan bahwa karakteristik antar kelas masih memiliki kemiripan sehingga proses klasifikasi menjadi lebih menantang dan memerlukan optimasi fitur untuk meningkatkan performa model klasifikasi.



Gambar 2. Visualisasi PCA dan t-SNE

3.4 Pengujian Nilai K pada KNN

Untuk menentukan jumlah tetangga terdekat yang paling optimal, dilakukan pengujian nilai K secara berulang dengan angka ganjil untuk menghindari angka seri. Pengujian dilakukan dari rentang K=3 sampai pada performa menunjukkan tren konvergen. Berdasarkan hasil pengujian, yang menggunakan 10-Fold Stratified CV pada rasio 90:10 dan 70:30 nilai akurasi tertinggi diperoleh pada K=13, sedangkan pada rasio 70:30 akurasi tertinggi diperoleh pada K=19. Hasil pengujian dapat dilihat pada Tabel 8.

Tabel 8. Hasil Pengujian Nilai K

Rasio	K = 3	K = 5	K = 7	K = 9	K = 11	K = 13	K = 15	K = 17	K = 19
90:10	49.62%	48.88%	49.32%	49.28%	49.28%	**49.63%**	49.55%	49.50%	49.51%
80:20	48.89%	48.70%	49.15%	48.97%	49.02%	48.97%	48.97%	49.19%	**49.25%**
70:30	49.20%	49.38%	49.21%	49.15%	49.17%	**49.70%**	49.66%	49.92%	49.67%

Berdasarkan hasil pengujian, nilai K=13 dipilih sebagai parameter optimal dikarenakan pada dua rasio K=13 menghasilkan akurasi tertinggi. Kemudian nilai K=13 berada pada rentang yang cukup seimbang.

3.5 Seleksi Fitur Algoritma Genetika

Proses seleksi fitur menggunakan Algoritma Genetika dilakukan pada data latih dari masing-masing skenario rasio pembagian data. GA mengevaluasi setiap subset fitur menggunakan fitness function berbasis akurasi 10-Fold Stratified Cross Validation dengan KNN (K=13). Proses evolusi berjalan hingga maksimal 100 generasi atau berhenti lebih awal apabila nilai fitness tidak meningkat selama 15 generasi berturut-turut. Hasil seleksi fitur pada setiap skenario rasio ditunjukkan pada Tabel 9.

Tabel 9. Hasil Seleksi Fitur Algoritma Genetika

Rasio	Fitur Awal	Fitur Terpilih	Fitness Terbaik
90:10	38	16	51.78%
80:20	38	15	51.84%

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

70:30	38	17	52.27%

Berdasarkan Tabel 9. Pada rasio 90:10, GA berhasil mereduksi 38 fitur menjadi 16 fitur terpilih dengan nilai fitness terbaik sebesar 51,78%. Fitur yang terpilih meliputi *Heart Disease, Systolic, Marital Status_Divorced, Marital Status_Married, Work Type_Self-employed, Smoking Status_Currently Smokes, Smoking Status_Formerly Smoked, Smoking Status_Non-smoker, Alcohol Intake_Never, Alcohol Intake_Social Drinker, Physical Activity_High, Physical Activity_Low, Dietary Habits_Gluten-Free, Dietary Habits_Paleo, Dietary Habits_Vegan, Dietary Habits_Vegetarian*. Pada rasio 80:20, GA memilih 15 fitur dari 38 fitur tersedia dengan nilai fitness sebesar 51,84%, meliputi *Gender, Residence Type, Average Glucose Level, Body Mass Index (BMI), Family History of Stroke, Systolic, Diastolic, HDL, Marital Status_Married, Work Type_Government Job, Work Type_Private, Alcohol Intake_Frequent Drinker, Alcohol Intake_Never, Physical Activity_Moderate, Dietary Habits_Vegetarian*. Pada rasio 70:30, GA juga memilih 17 fitur dengan nilai fitness 52,27%, meliputi *Age, Heart Disease, Residence Type, Body Mass Index (BMI), Family History of Stroke, Stress Levels, Marital Status_Divorced, Marital Status_Married, Marital Status_Single, Work Type_Never Worked, Alcohol Intake_Frequent Drinker, Alcohol Intake_Rarely, Alcohol Intake_Social Drinker, Physical Activity_High, Dietary Habits_Gluten-Free, Dietary Habits_Paleo, Dietary Habits_Vegetarian*.

Dari ketiga skenario tersebut, terdapat dua fitur yang konsisten terpilih pada seluruh rasio pembagian data, yaitu *Marital Status_Married* dan *Dietary Habits_Vegetarian*. Konsistensi pemilihan kedua fitur ini mengindikasikan bahwa fitur-fitur tersebut memiliki kontribusi yang relatif lebih signifikan terhadap proses klasifikasi stroke dibandingkan fitur lainnya. Meskipun GA berhasil meningkatkan nilai fitness CV sebesar +-3% dibandingkan KNN baseline, fitur yang terpilih berbeda-beda di setiap rasio. Hal ini menunjukkan bahwa GA menemukan solusi lokal optimal yang berbeda tergantung pada partisi data latih yang digunakan, bukan solusi global yang stabil. Kondisi ini konsisten dengan karakteristik dataset yang memiliki tingkat overlapping tinggi antar kelas, sehingga tidak terdapat satu subset fitur tunggal yang secara konsisten optimal pada seluruh kondisi pengujian.

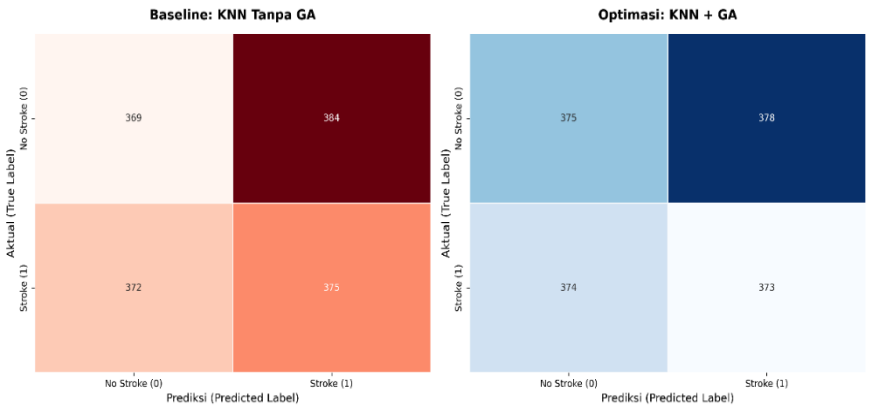
3.6 Evaluasi Performa Klasifikasi dan Robustness Check

Evaluasi performa klasifikasi dilakukan dengan membandingkan model KNN tanpa seleksi fitur dan model KNN dengan seleksi fitur Algoritma Genetika pada tiga skenario rasio pembagian data. Evaluasi menggunakan metrik akurasi, presisi, recall, F1-score, dan ROC-AUC pada data uji dengan parameter K=13. Hasil perbandingan performa kedua model ditunjukkan pada Tabel 10.

Tabel 10. Perbandingan Performa KNN dan GA-KNN

Rasio	Model	Akurasi	Presisi	Recall	F1-score	ROC-AUC
90:10	KNN	49.60%	49.41%	50.20%	49.80%	49.92%
	KNN-GA	49.87%	49.67%	49.93%	49.80%	50.17%
80:20	KNN	49.63%	49.43%	49.00%	49.21%	50.02%
	KNN-GA	49.47%	49.24%	47.99%	48.61%	49.04%
70:30	KNN	49.51%	49.27%	48.26%	48.76%	50.12%
	KNN-GA	49.91%	49.68%	48.08%	48.87%	49.43%

Berdasarkan Tabel 10 GA berhasil meningkatkan akurasi dibandingkan dengan model baseline, khusunya pada rasio 90:10 dan 70:30. Akurasi model meningkat dari 49,60% menjadi 49,87% pada rasio 90:10 dan 49,51% menjadi 49,91% pada rasio 70:30. Sementara itu, pada rasio 80:20 tidak mengalami kenaikan dari akurasi 49,63% menjadi 49,47%. Secara keseluruhan, seleksi fitur menggunakan GA mampu meningkatkan performa klasifikasi meskipun peningkatan akurasi tidak melonjak secara drastis. Hal ini dikarenakan tingginya Tingkat overlapping antar kelas pada dataset.



Gambar 3. Confusion Matrix KNN dan GA-KNN Rasio 90:10

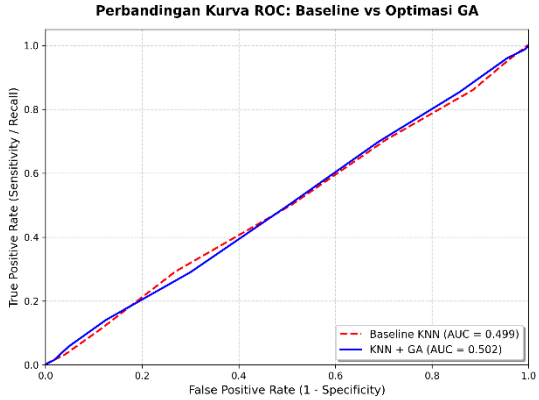
Untuk memperjelas distribusi hasil klasifikasi, visualisasi confusion matrix pada rasio 90:10 ditunjukkan pada Gambar 3. Berdasarkan confusion matrix KNN baseline, model berhasil mengklasifikasikan 369 data No Stroke dengan

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

benar (True Negative) dan 375 data Stroke dengan benar (True Positive). Namun, model juga menghasilkan 384 kesalahan False Positive (data No Stroke diprediksi Stroke) dan 372 kesalahan False Negative (data Stroke diprediksi No Stroke). Pada model GA-KNN, diperoleh True Negative sebesar 375, True Positive sebesar 373, False Positive sebesar 378, dan False Negative sebesar 374.

Meskipun terdapat sedikit peningkatan, tingginya angka False Positive dan False Negative pada kedua model yang hampir setara dengan jumlah tebakan benarnya menjelaskan bahwa model mengalami kesulitan yang signifikan dalam memisahkan kelas. Hal ini menjadi bukti bahwa adanya overlapping yang ekstrem antar kelas dalam dataset. Selain itu, penurunan performa GA-KNN pada skenario rasio 80:20, khususnya pada metrik recall (47,99%) dan F1-score (48,61%), melihatkan adanya tantangan dalam hal generalisasi model. GA berhasil mereduksi fitur dan meningkatkan nilai fitness saat cross validation, namun solusi lokal optimal yang ditemukan oleh GA tersebut cenderung sensitif terhadap variasi pembagian data, sehingga peningkatannya tidak selalu konsisten ketika data diuji dengan data yang belum terlihat.

Hasil pengujian menunjukkan nilai AUC KNN baseline sebesar 0,499 dan GA-KNN sebesar 0,502 ditunjukkan pada Gambar 4. Yang menunjukkan klasifikasi masuk dalam kategori kegagalan. Hal ini menunjukkan bahwa baik model standar maupun model hasil optimasi GA belum mampu membedakan kelas stroke dan no stroke secara efektif.



Gambar 4. Kurva ROC

Sebagai robustness check, dilakukan eksperimen pembanding menggunakan Random Forest (RF) dengan parameter *n_estimators=50, max_dept=10, min_samples_split=5, dan min_samples_leaf=2*. Hasil pengujian dapat dilihat pada Tabel 11.

Tabel 11. Pengujian Random Forest

Rasio	Model	Akurasi	Presisi	Recall	F1-score	ROC-AUC
90:10	RF	50.80%	50.61%	50.07%	50.34%	50.89%
	RF-GA	50.80%	50.58%	52.61%	51.57%	50.13%
80:20	RF	50.17%	49.96%	46.39%	48.11%	49.22%
	RF-GA	50.20%	50.00%	51.47%	50.73%	50.66%
70:30	RF	50.09%	49.87%	49.73%	49.80%	50.68%
	RF-GA	49.93%	49.69%	45.98%	47.76%	50.04%

Berdasarkan Tabel 11, secara keseluruhan akurasi masih berada sekitar 50% pada seluruh skenario pembagian data. Selain itu, nilai ROC-AUC juga menunjukkan di angka 50% yang membuktikan bahwa kemampuan pemisahan kelas masih sangat terbatas, meskipun menggunakan model yang tidak berbasis jarak seperti KNN.

Selain itu, dilakukan Feature Engineering (FE) dengan mengekstraksi lima fitur turunan baru yang mengelompokkan variable-variabel tunggal menjadi skor risiko yang lebih terstruktur. Pertama, *Cardiovascular Risk Score*, yaitu akumulasi dari riwayat kardiovaskular bawaan pasien (Hypertension, Heart Disease, Stroke History, dan Family History of Stroke). Kedua, *Cholesterol Ratio*, yang dihitung dari rasio pembagian antara kadar LDL dan HDL. Ketiga, *Metabolic Syndrome Score*, yang merupakan gabungan nilai normalisasi dari Body Mass Index (BMI), Average Glucose Level, kadar LDL, serta status hipertensi. Keempat, *Lifestyle Risk Score*, berupa pengelompokan profil gaya hidup berisiko tinggi yang mencakup perokok aktif, konsumsi alkohol intensitas tinggi (frequent drinker), aktivitas fisik rendah, dan diet non-vegetarian. Terakhir, *BP Category*, yang mengategorikan nilai tekanan darah Sistolik dan Diastolik ke dalam tiga tingkat keparahan yaitu Normal, Pre-hipertensi, dan Hipertensi. Setelah kelima fitur turunan tersebut terbentuk, dilakukan penghapusan terhadap variabel-variabel asli yang telah diekstrak dan redundan, sehingga dimensi dataset berhasil direduksi dari 43 menjadi 34 fitur. Hasil pengujian dapat dilihat pada Tabel 12.

Tabel 12. Hasil Pengujian Feature Engineering

Model	Akurasi	Presisi	Recall	F1-score	ROC-AUC
FE KNN	50.87%	50.67%	50.74%	50.70%	51.45%
FE KNN-GA	51.20%	51.02%	50.20%	50.61%	51.76%

ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 233–243
DOI : 10.47065/bulletinds.v5i3.10179
https://ejurnal.seminar-id.com/index.php/bulletinds

Model	Akurasi	Presisi	Recall	F1-score	ROC-AUC
FE RF	50.53%	50.33%	51.14%	50.73%	50.16%
FE RF-GA	49.60%	49.43%	52.61%	50.97%	49.74%

Berdasarkan Tabel 12, performa seluruh varian model tetap di kisaran 50%. Model FE KNN-GA hanya mencapai akurasi tertinggi sebesar 51,20%. Penggunaan algoritma pembanding Random Forest juga terbukti tidak memberikan pengaruh yang berarti, di mana akurasi FE RF hanya mencapai 50,53% dan FE RF-GA justru menurun ke 49,60%.

3.7 Pembahasan

Performa GA-KNN pada penelitian ini yang hanya mencapai akurasi 48-52% jauh lebih rendah dibandingkan dengan penelitian sebelumnya, seperti [4] dan [10] yang mencapai akurasi 93,54% dan 95,77%. Serta [12] dan [15] yang mengalami peningkatan akurasi signifikan menggunakan GA dengan akurasi mencapai 98,2% dan 95,13%. Perbedaan ini disebabkan oleh karakteristik internal dataset, penelitian sebelumnya umumnya menggunakan dataset dengan batas pemisah kelas yang jelas. Sedangkan pada penelitian ini menggunakan dataset yang memiliki tingkat overlapping yang tinggi antar kelas. Sebagaimana ditunjukkan oleh visualisasi PCA dan t-SNE pada Gambar 2, serta dibuktikan dengan hasil Random Forest dan Feature Engineering yang juga tetap terhenti pada akurasi 50%-52%. Hal ini menunjukkan bahwa efektivitas seleksi fitur GA sangat bergantung pada separabilitas data, dan tidak dapat secara otomatis mengatasi keterbatasan akurasi apabila overlapping merupakan karakteristik bawaan dataset itu sendiri.

4. KESIMPULAN

Penelitian ini telah mengevaluasi efektivitas Algoritma Genetika (GA) dalam mengoptimasi ruang fitur untuk klasifikasi prediksi penyakit stroke menggunakan algoritma K-Nearest Neighbor (KNN) pada data rekam medis pasien. Berdasarkan hasil pengujian, GA terbukti efektif dalam mengefisienkan komputasi dan mereduksi dimensi data. Algoritma ini juga berhasil menyaring atribut yang redundan secara signifikan, memangkas 38 atribut fitur hasil preprocessing menjadi hanya 15 hingga 17 fitur terpilih pada berbagai skenario rasio pembagian data. Selain itu, pada tahap validasi internal menggunakan data latih, kombinasi GA dan KNN mampu menunjukkan kinerja yang sangat baik dengan meningkatkan rata-rata nilai fitness sekitar tiga persen dibandingkan model baseline tanpa seleksi fitur. Meskipun GA menunjukkan kinerja unggul pada fase pelatihan, evaluasi lebih lanjut terhadap data uji yang belum pernah terlihat menunjukkan bahwa performa model tetap terhenti dan kesulitan melakukan generalisasi saat memproses data baru karena GA cenderung menemukan solusi lokal optimal yang sangat sensitif terhadap variasi partisi data latih, sehingga model kesulitan melakukan generalisasi saat memproses data baru. Berhentinya performa secara keseluruhan dipengaruhi oleh karakteristik internal dataset medis tersebut yang memiliki tingkat tumpang tindih (overlapping) sangat ekstrem antar kelas. Hal ini dibuktikan lebih lanjut secara nyata melalui eksperimen pembanding menggunakan rekayasa fitur turunan (Feature Engineering) serta pergantian ke algoritma Random Forest, yang seluruhnya tetap terhenti di kisaran akurasi 50%. Secara keseluruhan, seleksi fitur GA sukses menyederhanakan kompleksitas data, namun tidak secara otomatis mampu menyelesaikan masalah pemisahan kelas apabila tumpang tindih antara fitur dengan label merupakan karakteristik bawaan dari data itu sendiri.

REFERENCES

[1] K. Wirastuti, N. S. Riasari, D. Djannah, and M. Silviana, “Upaya Pencegahan Stroke melalui Skrining Skor Risiko Stroke dengan Intervensi Penyuluhan dan Pemeriksaan Faktor Risiko Stroke di Kelurahan Bojong Salaman Kecamatan Pusponjolo Selatan Semarang Barat,” *J. ABDIMAS-KU J. Pengabdi. Masy. Kedokt.*, vol. 2, no. 1, pp. 23–29, Jan. 2023, doi: 10.30659/abdimasku.2.1.23-29.

[2] S. Rahayu and Y. Yamasari, “Klasifikasi Penyakit Stroke dengan Metode Support Vector Machine (SVM),” *J. Informatics Comput. Sci.*, vol. 5, no. 03, pp. 440–446, Feb. 2024, doi: 10.26740/jinacs.v5n03.p440-446.

[3] K. Sari *et al.*, “Deteksi Dini Stroke Menggunakan Machine Learning,” *INSOLOGI J. Sains dan Teknol.*, vol. 4, no. 4, pp. 706–720, Aug. 2025, doi: 10.55123/INSOLOGI.V4I4.5590.

[4] Z. Zuriati and N. Qomariyah, “Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor (KNN),” *ROUTERS J. Sist. dan Teknol. Inf.*, vol. 1, no. 1, pp. 1–8, Nov. 2022, doi: 10.25181/RT.V1I1.2665.

[5] T. G. Rahayu, “Analisis Faktor Risiko Terjadinya Stroke Serta Tipe Stroke,” *Faletehan Heal. J.*, vol. 10, no. 01, pp. 48–53, Mar. 2023, doi: 10.33746/fhj.v10i01.410.

[6] A. Fitri, I. Afrianty, E. Budianita, and S. K. Gusti, “Implementation of Feature Selection Information Gain in Support Vector Machine Method for Stroke Disease Classification,” *Bull. Informatics Data Sci.*, vol. 4, no. 1, pp. 22–33, May 2025, doi: 10.61944/bids.v4i1.116.

[7] K. Wirastuti, N. Sofi, and I. Rosdiana, “Pendampingan Kader Kesehatan dengan Intervensi Penyuluhan Peduli Cegah Kenali dan Atasi Stroke (Cekatan Stroke) Kelurahan Banjardowo Kecamatan Genuk Kota Semarang,” *J. ABDIMAS-KU J. Pengabdi. Masy. Kedokt.*, vol. 3, no. 1, pp. 6–12, Jan. 2024, doi: 10.30659/abdimasku.3.1.6-12.

[8] I. O. Herlistiono and S. Violina, “Model Prediksi Risiko Stroke Menggunakan Machine Learning,” *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 7, no. 4, pp. 1230–1238, Jul. 2024, doi: 10.31539/INTECOMS.V7I4.10942.

[9] C. Paramita, C. S. Simbolon, A. S. Pamungkas, J. M. Triono, E. P. Widi Utomo, and E. R. Subhiyakto, “Analisis Pengaruh SMOTE terhadap Kinerja Model KNN untuk Prediksi Risiko Stroke,” *J. Inform. J. Pengemb. IT*, vol. 10, no. 4, pp. 978–988,

Sep. 2025, doi: 10.30591/JPIT.V10I4.8809.
[10] I. Bagus, A. Indra Iswara, G. Anandita, and M. Dahul, "Comparative Analysis of Naïve Bayes and K-Nearest Neighbor (KNN) Algorithms in Stroke Classification," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 3, pp. 1415–1424, Jul. 2024, doi: 10.47709/CNAHPC.V6I3.4395.
[11] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, Jan. 2024, doi: 10.57152/MALCOM.V4I1.1085.
[12] M. A. V. Darmawan, M. M. Al Haromainy, and A. Junaidi, "OPTIMASI ALGORITMA K-NEAREST NEIGHBOR DENGAN ALGORITMA GENETIKA PADA DETEKSI PENYAKIT DIABETES MELLITUS," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 12, no. 2, Jun. 2025, doi: 10.35957/JATISI.V12I2.11353.
[13] F. Novianti and N. Ulinnuha, "SELEKSI FITUR ALGORITMA GENETIKA DALAM KLASIFIKASI DATA REKAM MEDIS PCOS MENGGUNAKAN SVM," *NERO (Networking Eng. Res. Oper.*, vol. 9, no. 1, pp. 9–20, Apr. 2024, doi: 10.21107/NERO.V9I1.25399.
[14] A. Alassaf *et al.*, "Genetic Algorithms and Feature Selection for Improving the Classification Performance in Healthcare," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 3, pp. 737–744, Mar. 2024, doi: 10.14569/IJACSA.2024.0150375.
[15] S. Wulandari, Y. I. Mukti, and T. Susanti, "Optimization of the Artificial Neural Network Algorithm with Genetic Algorithm in Stroke Prediction," *Sink. J. dan Penelit. Tek. Inform.*, vol. 8, no. 2, pp. 1056–1063, Apr. 2024, doi: 10.33395/SINKRON.V8I2.13609.
[16] G. Feng, "Feature selection algorithm based on optimized genetic algorithm and the application in high-dimensional data processing," *PLoS One*, vol. 19, no. 5, p. e0303088, May 2024, doi: 10.1371/JOURNAL.PONE.0303088.
[17] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *MDPI*, vol. 6, no. 10, p. 257, Oct. 2025, doi: 10.3390/AI6100257.
[18] M. R. Firmansyah and Y. P. Astuti, "Stroke Classification Comparison with KNN through Standardization and Normalization Techniques," *Adv. Sustain. Sci. Eng. Technol.*, vol. 6, no. 1, p. 02401012, Jan. 2024, doi: 10.26877/ASSET.V6I1.17685.
[19] C. Herdian, A. Kamila, and I. G. A. M. Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technol. J. Ilm.*, vol. 15, no. 1, pp. 93–108, Jan. 2024, doi: 10.31602/TJI.V15I1.13457.
[20] T. Abedin, H. Xu, and S. Uddin, "The impact of K selection in K-fold cross-validation on bias and variance in supervised learning models," *Sci. Reports 2026 161*, vol. 16, no. 1, pp. 6084-, Jan. 2026, doi: 10.1038/s41598-026-37247-x.
[21] T. H. Raza, A. Ahmad, M. Z. Hussain, M. Z. Hasan, H. R. Basra, and A. Bilal, "Enhancing Multivariate Data Classification Using Graph Convolutional Networks: A Comparative Evaluation with PCA and t-SNE," *VAWKUM Trans. Comput. Sci.*, vol. 13, no. 2, pp. 149–165, Oct. 2025, doi: 10.21015/VTCS.V13I2.2052.