OPEN ACCESS

Volume 10 (3), July 2026, 1208-1216

E-ISSN:2580-1643

Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)

DOI: https://doi.org/10.35870/jtik.v10i3.6372

Optimasi Kinerja Algoritma *K-Nearest Neighbor* melalui Metode *Random Forest* untuk Klasifikasi Penyakit Ginjal

Achmad Hakim Qoirul Haq [1*], Harminto Mulyo [2], Adi Sucipto [3]

[1*,2,3] *Universitas Islam Nahdlatul Ulama Jepara, Kabupaten Jepara, Jawa Tengah, Indonesia.*

a r t i c l e i n f o

Article history:
Received 20 December 2025
Received in revised form
20 January 2026
Accepted 1 February 2026
Available online July 2026.

Keywords:
Chronic Kidney Disease; K-Nearest Neighbors; Random Forest; Classification; Machine Learning.

Kata Kunci:
Penyakit Ginjal Kronis; K-Nearest Neighbor; Random Forest; Klasifikasi; Machine Learning.

a b s t r a c t

Chronic Kidney Disease (CKD) is a chronic disease with a continuously increasing prevalence rate and requires early detection to prevent disease progression. This study aims to optimize the performance of the K-Nearest Neighbor (K-NN) algorithm in the classification of chronic kidney disease through the application of the Random Forest method. The dataset used comes from Kaggle and consists of 400 patient data with 26 clinical attributes. The research stages include data pre-processing in the form of handling missing values, categorical data transformation, feature normalization, and data division into training data and test data with a ratio of 80:20. Random Forest is used as a comparison method and optimization approach, while K-NN is used as the main classification algorithm. Model performance evaluation is carried out using accuracy, precision, recall, F1-score, and confusion matrix metrics. The test results show that the Random Forest algorithm obtains an accuracy value of 98.75%, while the K-NN algorithm produces an accuracy of 96.25%. These results prove that the application of Random Forest is able to optimize the performance of K-NN in the classification of chronic kidney disease effectively.

a b s t r a k

Penyakit Ginjal Kronis (Chronic Kidney Disease / CKD) merupakan penyakit kronis dengan tingkat prevalensi yang terus meningkat dan memerlukan deteksi dini untuk mencegah progresivitas penyakit. Penelitian ini bertujuan untuk mengoptimasi kinerja algoritma K-Nearest Neighbor (K-NN) dalam klasifikasi penyakit ginjal kronis melalui penerapan metode Random Forest. Dataset yang digunakan berasal dari Kaggle dan terdiri dari 400 data pasien dengan 26 atribut klinis. Tahapan penelitian meliputi pra-pemrosesan data berupa penanganan missing value, transformasi data kategorikal, normalisasi fitur, serta pembagian data menjadi data latih dan data uji dengan rasio 80:20. Random Forest digunakan sebagai metode pembanding dan pendekatan optimasi, sedangkan K-NN digunakan sebagai algoritma klasifikasi utama. Evaluasi kinerja model dilakukan menggunakan metrik accuracy, precision, recall, F1-score, dan confusion matrix. Hasil pengujian menunjukkan bahwa algoritma Random Forest memperoleh nilai akurasi sebesar 98,75%, sedangkan algoritma K-NN menghasilkan akurasi sebesar 96,25%. Hasil tersebut membuktikan bahwa penerapan Random Forest mampu mengoptimasi kinerja K-NN dalam klasifikasi penyakit ginjal kronis secara efektif.

Corresponding Author. Email: achmadhakim9090@gmail.com [1].

1. Pendahuluan

Penyakit Ginjal Kronis (*Chronic Kidney Disease* / CKD) merupakan masalah kesehatan masyarakat yang signifikan secara global, dengan prevalensi yang terus meningkat seiring waktu dan berdampak pada morbiditas, mortalitas, serta biaya perawatan kesehatan. CKD sering tidak menunjukkan gejala pada tahap awal sehingga pemeriksaan dini menjadi sangat penting untuk mencegah progresi ke gagal ginjal yang lebih parah dan komplikasi sistemik lainnya. Prediksi dan diagnosa CKD secara efektif memerlukan analisis data klinis yang mampu mengungkap pola kompleks dari banyak parameter medis yang tersedia (Dritsas & Trigka, 2022). Tantangan utama dalam klasifikasi data medis seperti CKD adalah adanya *curse of dimensionality*, di mana banyaknya fitur yang tidak relevan dapat menyebabkan algoritma K-NN mengalami penurunan performa karena perhitungan jarak menjadi tidak efisien (Islam, Majumder, & Hussein, 2023). Oleh karena itu, diperlukan teknik reduksi dimensi. Random Forest dipilih sebagai penyeleksi fitur karena kemampuannya dalam menangani data non-linear dan memberikan skor *Importance* yang stabil terhadap variabel medis yang paling berpengaruh.

Dengan menggabungkan kedua algoritma ini, diharapkan model klasifikasi tidak hanya akurat tetapi juga memiliki kompleksitas komputasi yang lebih rendah. Dalam dekade terakhir, pendekatan *machine learning* telah banyak diterapkan dalam klasifikasi dan prediksi CKD berdasarkan data klinis dan laboratorium pasien. Berbagai algoritma, seperti Random Forest, K-Nearest Neighbor (*K-NN*), *Support Vector Machine*, dan *ensemble methods* telah dievaluasi untuk kemampuan mereka dalam mendeteksi CKD secara akurat dari dataset publik maupun klinis. Beberapa penelitian menunjukkan bahwa Random Forest mampu menyediakan performa prediksi yang kuat dan identifikasi fitur-fitur penting dalam dataset CKD, bahkan untuk dataset berukuran besar (mis. lebih dari 40.000 rekam medis) dengan nilai area under *ROC curve* (AUC) yang signifikan, menunjukkan nilai prediktif yang tinggi dalam screening faktor risiko CKD (Siregar, Hartama, & Solikhun, 2025). Selain itu, beberapa studi juga menilai efektivitas algoritma *instance-based learning* seperti *K-NN* dalam klasifikasi CKD, di mana *K-NN* menunjukkan hasil yang kompetitif dibandingkan algoritma lain pada dataset tertentu ketika tahap pra-pemrosesan (penanganan missing value dan *feature engineering*) dilakukan secara tepat (Diqi, Ordiyasa, & Hiswati, 2023). Penelitian terdahulu menegaskan bahwa integrasi teknik seleksi fitur dan pendekatan *ensemble* atau perbaikan model dapat meningkatkan kinerja klasifikasi secara signifikan untuk deteksi CKD dibandingkan penggunaan model tunggal (Miaschi, Alzetta, Brunato, Orletta, & Venturi, 2023). Meskipun berbagai pendekatan telah diusulkan, terdapat kebutuhan akan evaluasi yang lebih mendalam terhadap strategi optimasi algoritma, khususnya dalam konteks menggabungkan metode seleksi fitur seperti Random Forest dengan algoritma klasifikasi sederhana seperti *K-NN*, untuk mengatasi keterbatasan *K-NN* dalam menghadapi dimensi fitur yang tinggi dan ketidakseimbangan data. Berdasarkan temuan tersebut, penelitian ini mengevaluasi performa *K-NN* yang dioptimasi menggunakan seleksi fitur berbasis Random Forest untuk klasifikasi CKD pada dataset publik, dengan tujuan untuk menentukan efektivitas kombinasi metode tersebut dalam meningkatkan akurasi dan stabilitas model klasifikasi penyakit ginjal kronis (Liu, Liu, Liu, Xiong, Mei, & Yuan, 2024; Khalid, Khan, Zahid Khan, Mehmood, & Shuaib Qureshi, 2023).

2. Metodologi Penelitian

Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimental dengan metode kuantitatif untuk melakukan klasifikasi penyakit ginjal kronis (*Chronic Kidney Disease* / CKD) berbasis *machine learning*. Pendekatan ini dipilih karena mampu mengevaluasi kinerja algoritma klasifikasi secara objektif melalui pengujian berbasis data klinis dan metrik evaluasi terukur. Seluruh tahapan penelitian, mulai dari pengolahan data hingga evaluasi model, dilakukan menggunakan bahasa pemrograman Python pada lingkungan *Google Colaboratory*, sebagaimana umum digunakan pada penelitian klasifikasi medis berbasis data sekunder (Taqwimi & Wahono, 2025). Alur penelitian disusun secara sistematis yang mencakup tahap pengumpulan data, pra-pemrosesan

data, pemodelan menggunakan algoritma Random Forest dan *K-Nearest Neighbor* (*K-NN*), serta evaluasi kinerja model. Diagram alur (*flowchart*) penelitian digunakan untuk menggambarkan keseluruhan proses penelitian secara visual guna memastikan keterulangan (*reproducibility*) dan konsistensi metode yang digunakan (Muhammad Ahsani & Wahono, 2025).

Pengumpulan Data

Dataset yang digunakan dalam penelitian ini merupakan dataset *Chronic Kidney Disease* yang diperoleh dari platform *Kaggle*. Dataset ini terdiri dari 400 data pasien dengan 26 atribut klinis yang mencakup parameter demografis, hasil pemeriksaan laboratorium, dan kondisi medis pasien. Dataset tersebut banyak digunakan dalam penelitian sebelumnya sebagai data rujukan untuk klasifikasi CKD karena memiliki variasi fitur klinis yang representatif terhadap kondisi pasien (Siregar, Hartama, & Solikhun, 2025). Tabel 1 menyajikan ringkasan dataset yang digunakan dalam penelitian ini, terdiri dari 400 data dengan 26 fitur yang mencakup 11 atribut numerik dan 14 atribut kategorikal, serta target klasifikasi CKD dan Non-CKD.

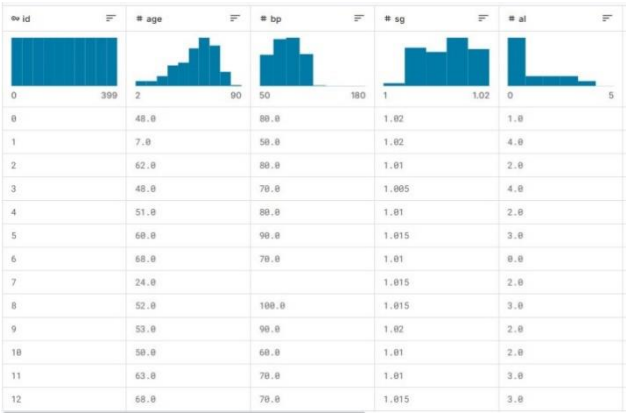
Tabel 1. Dataset Summary

Item	Value
Source	Kaggle
Total instances	400
Features	26
Numerical	11
Categorical	14
Target	CKD / Non-CKD

Penggunaan dataset sekunder dari *Kaggle* memungkinkan penelitian dilakukan secara efisien tanpa melibatkan pengumpulan data primer, serta memudahkan proses perbandingan hasil dengan penelitian-penelitian terdahulu. Dataset ini digunakan sebagai data masukan utama untuk melatih dan menguji performa model klasifikasi CKD yang dikembangkan (Rahim & Ahmar, 2022).

Sample Data

Tabel 1 menyajikan contoh beberapa data dari dataset CKD yang digunakan dalam penelitian ini. Data yang ditampilkan merupakan sebagian kecil dari dataset asli dan digunakan untuk memberikan gambaran struktur serta karakteristik data.



id	age	bp	sg	al
0 – 399	2 – 90	50 – 180	1 – 1.02	0 – 5
0	48.0	80.0	1.02	1.0
1	7.0	50.0	1.02	4.0
2	62.0	80.0	1.01	2.0
3	48.0	70.0	1.005	4.0
4	51.0	80.0	1.01	2.0
5	60.0	90.0	1.015	3.0
6	68.0	70.0	1.01	0.0
7	24.0		1.015	2.0
8	52.0	100.0	1.015	3.0
9	53.0	90.0	1.02	2.0
10	50.0	60.0	1.01	2.0
11	63.0	70.0	1.01	3.0
12	68.0	70.0	1.015	3.0

Gambar 1. Sample CKD Dataset

Data *Preprocessing*

Tahap pra-pemrosesan data dilakukan untuk meningkatkan kualitas data sebelum digunakan dalam proses pemodelan. Atribut yang tidak relevan, seperti atribut identifikasi (*ID*), dihapus karena tidak memberikan kontribusi terhadap proses klasifikasi. Selanjutnya, penanganan nilai kosong (*missing value*) dilakukan dengan mengganti nilai kosong pada atribut numerik menggunakan nilai median, serta pada atribut kategorikal menggunakan nilai modus. Pendekatan ini dipilih karena mampu menjaga distribusi data dan mengurangi bias yang dapat memengaruhi hasil klasifikasi (Jeon & Oh, 2020). Data kategorikal kemudian ditransformasikan ke dalam bentuk numerik menggunakan metode *Label Encoding* agar dapat diproses oleh algoritma *machine learning*. Setelah itu, normalisasi data dilakukan menggunakan *StandardScaler* untuk memastikan seluruh fitur berada pada skala yang sebanding. Normalisasi ini sangat penting terutama pada algoritma *K-NN* yang sensitif terhadap perbedaan skala antar fitur (Majid *et al*., 2023).

Data Encoding and Normalization

Data kategorikal kemudian ditransformasikan ke dalam bentuk numerik menggunakan metode *Label Encoding* agar dapat diproses oleh algoritma *machine learning*. Setelah proses encoding, normalisasi data dilakukan menggunakan *StandardScaler* untuk memastikan setiap fitur berada pada skala yang sebanding. Normalisasi ini sangat penting terutama pada algoritma *K-NN* yang sensitif terhadap

perbedaan skala antar fitur (Id, Schulte, & Groene, 2023). Proses normalisasi dilakukan menggunakan persamaan berikut:

$$z = \frac{x - \mu}{\sigma}$$

Di mana x merupakan nilai asli, μ adalah nilai rata-rata, dan σ adalah standar deviasi dari data.

Data *Inspection and Missing Value Identification*

Sebelum dilakukan tahap pra-pemrosesan, inspeksi awal terhadap dataset dilakukan untuk memahami struktur data, tipe atribut, serta keberadaan nilai kosong (*missing value*). Berdasarkan hasil inspeksi menggunakan fungsi *df.info()*, dataset terdiri dari 400 data dengan total 26 atribut yang mencakup atribut numerik dan kategorikal. Hasil inspeksi menunjukkan bahwa beberapa atribut, khususnya atribut hasil pemeriksaan laboratorium, memiliki jumlah data yang tidak lengkap. Kondisi ini umum ditemukan pada dataset medis dan dapat memengaruhi kinerja algoritma *machine learning* apabila tidak ditangani dengan tepat (Rahim & Ahmar, 2022).

```
#   Column          Non-Null Count  Dtype
---  ------          --------------  -----
0   id              400 non-null    int64
1   age             391 non-null    float64
2   bp              388 non-null    float64
3   sg              353 non-null    float64
4   al              354 non-null    float64
5   su              351 non-null    float64
6   rbc             248 non-null    object
7   pc              335 non-null    object
8   pcc             396 non-null    object
9   ba              396 non-null    object
10  bgr             356 non-null    float64
11  bu              381 non-null    float64
12  sc              383 non-null    float64
13  sod             313 non-null    float64
14  pot             312 non-null    float64
15  hemo            348 non-null    float64
16  pcv             330 non-null    object
17  wc              295 non-null    object
18  rc              270 non-null    object
19  htn             398 non-null    object
20  dm              398 non-null    object
21  cad             398 non-null    object
22  appet           399 non-null    object
23  pe              399 non-null    object
24  ane             399 non-null    object
25  classification  400 non-null    object
```

Gambar 2. Missing Value

Keberadaan *missing value* pada dataset dapat menyebabkan penurunan akurasi model serta bias dalam proses pembelajaran, terutama pada algoritma berbasis jarak seperti *K-Nearest Neighbor*. Oleh karena itu, diperlukan strategi pra-pemrosesan yang tepat untuk menangani data yang tidak lengkap sebelum dilakukan proses pelatihan model klasifikasi. Tahap inspeksi data ini menjadi langkah penting untuk menentukan metode penanganan *missing value* yang sesuai agar kualitas data tetap terjaga (Majid *et al*., 2023).

Data *Splitting*

Dataset yang telah melalui tahap pra-pemrosesan selanjutnya dibagi menjadi data latih (*training set*) dan data uji (*testing set*) dengan rasio 80% untuk data latih dan 20% untuk data uji menggunakan metode *train-test split*. Pembagian ini bertujuan untuk melatih model pada sebagian besar data serta menguji kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya. Teknik *stratified sampling* diterapkan pada proses pembagian data untuk menjaga proporsi kelas CKD dan non-CKD tetap seimbang pada data latih dan data uji. Pendekatan ini penting untuk menghindari bias model terhadap kelas tertentu dan meningkatkan keandalan hasil evaluasi.

K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (*K-NN*) merupakan metode klasifikasi berbasis instansi yang menentukan label suatu data berdasarkan mayoritas kategori dari k tetangga terdekatnya. Kedekatan tersebut diukur menggunakan fungsi jarak, di mana pada penelitian ini digunakan *Euclidean Distance* untuk menghitung jarak antara data uji dan data latih dalam ruang multidimensi (Islam, Majumder, & Hussein, 2023). Persamaan *Euclidean Distance* didefinisikan sebagai berikut:

$$d(x, y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)2}$$

Di mana d(x, y) adalah jarak antara data x dan y, n merupakan jumlah fitur yang telah melalui tahap seleksi, serta x_i dan y_i adalah nilai fitur pada dimensi ke-i.

Random Forest

Algoritma *K-Nearest Neighbor* (*K-NN*) merupakan metode klasifikasi berbasis instansi yang menentukan label suatu data berdasarkan mayoritas kategori dari k tetangga terdekatnya. Kedekatan tersebut diukur menggunakan fungsi jarak, di mana pada penelitian ini digunakan *Euclidean Distance* untuk menghitung jarak

antara data uji dan data latih dalam ruang multidimensi (Breiman, 2001). Persamaan *Euclidean Distance* didefinisikan sebagai berikut:

$$Gini(D) = 1 - \sum_{i=0}^{c} p_i^2$$

Di mana P_i merupakan probabilitas relatif dari kelas i pada dataset D. Mekanisme ini memungkinkan Random Forest untuk mengidentifikasi fitur medis yang paling relevan terhadap penyakit ginjal kronis.

Modeling

Pada tahap pemodelan, digunakan dua algoritma klasifikasi, yaitu Random Forest dan *K-Nearest Neighbor* (*K-NN*). Random Forest digunakan sebagai model pembanding karena kemampuannya dalam menangani data berdimensi tinggi dan menghasilkan performa klasifikasi yang stabil melalui mekanisme *ensemble learning* (Breiman, 2001). Secara matematis, prediksi kelas pada Random Forest ditentukan menggunakan mekanisme *majority voting* sebagai berikut:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\}$$

Algoritma K-NN digunakan sebagai algoritma klasifikasi utama dengan nilai k=5k = 5k=5 dan metrik jarak Minkowski. Jarak antar data dihitung menggunakan persamaan:

$$d(x, y) = \left(\sum_{i=0}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Dengan *p = 2* untuk jarak *Euclidean*. Pemilihan *K-NN* didasarkan pada kesederhanaan algoritma dan kemampuannya dalam menghasilkan performa yang baik pada data medis ketika didukung oleh pra-pemrosesan yang tepat (Siregar, Hartama, & Solikhun, 2025).

Model *Evaluation*

Evaluasi kinerja model dilakukan untuk mengukur kemampuan algoritma dalam mengklasifikasikan penyakit ginjal kronis (*Chronic Kidney Disease* / CKD) secara akurat. Evaluasi dilakukan menggunakan beberapa metrik evaluasi yang umum digunakan dalam klasifikasi biner pada bidang kesehatan, yaitu *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*. Penggunaan kombinasi metrik ini bertujuan untuk memberikan gambaran performa model yang lebih komprehensif dibandingkan hanya menggunakan satu metrik evaluasi saja (Breiman, 2001; Dritsas & Trigka, 2022).

Confusion Matrix

Confusion matrix digunakan untuk mengevaluasi performa model dengan membandingkan hasil prediksi terhadap nilai aktual. Matriks ini terdiri dari empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). TP menunjukkan jumlah data CKD yang diprediksi benar sebagai CKD, sedangkan TN menunjukkan jumlah data non-CKD yang diprediksi benar. FP dan FN masing-masing menunjukkan kesalahan prediksi positif dan negatif yang dilakukan oleh model (Islam, Majumder, & Hussein, 2023). Berdasarkan *confusion matrix*, berbagai metrik evaluasi dapat diturunkan untuk mengukur kinerja model secara kuantitatif, terutama dalam konteks klasifikasi penyakit yang menuntut tingkat kesalahan rendah (Rahim & Ahmar, 2022).

Accuracy

Accuracy merupakan rasio antara jumlah prediksi yang benar terhadap total jumlah data yang diprediksi. Metrik ini digunakan untuk mengukur tingkat ketepatan keseluruhan model dalam melakukan klasifikasi CKD dan non-CKD. Formula *accuracy* ditunjukkan pada Persamaan (Breiman, 2001).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision

Precision digunakan untuk mengukur tingkat ketepatan prediksi positif yang dilakukan oleh model. Metrik ini menunjukkan seberapa besar proporsi data yang diprediksi sebagai CKD benar-benar merupakan kasus CKD. *Precision* sangat penting dalam konteks medis untuk mengurangi kesalahan diagnosis positif (Siregar, Hartama, & Solikhun, 2025). Persamaan *precision* ditunjukkan pada Persamaan berikut:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall

Recall atau *True Positive Rate* (TPR) digunakan untuk mengukur kemampuan model dalam mendeteksi seluruh data CKD yang sebenarnya. Nilai *recall* yang

tinggi menunjukkan bahwa model mampu meminimalkan kesalahan *False Negative*, yang sangat penting dalam diagnosis penyakit kronis seperti CKD (Breiman, 2001; Rahim & Ahmar, 2022). Formula *recall* ditunjukkan pada Persamaan berikut:

$$Recall = \frac{TP}{TP+FN}$$

F1-Score

F1-score merupakan nilai rata-rata harmonik dari *precision* dan *recall* yang digunakan untuk menyeimbangkan kedua metrik tersebut. Metrik ini memberikan gambaran performa model yang lebih adil ketika terdapat perbedaan distribusi kelas atau ketika diperlukan keseimbangan antara ketepatan dan sensitivitas model (Siregar, Hartama, & Solikhun, 2025). Formula *F1-score* ditunjukkan pada Persamaan berikut:

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC and Area Under Curve (AUC)

Receiver Operating Characteristic (*ROC*) digunakan untuk menggambarkan hubungan antara *True Positive Rate* (*TPR*) dan *False Positive Rate* (*FPR*) pada berbagai nilai ambang keputusan. *FPR* didefinisikan sebagai rasio antara *False Positive* terhadap total data negatif aktual, sebagaimana ditunjukkan pada Persamaan berikut (Breiman, 2001):

$$\text{FPR} = \frac{FP}{FP+TN}$$

Nilai *Area Under Curve* (*AUC*) merepresentasikan luas area di bawah kurva *ROC* dan digunakan untuk mengukur kemampuan model dalam membedakan kelas CKD dan non-CKD. Semakin tinggi nilai *AUC*, semakin baik performa model dalam melakukan klasifikasi. Secara matematis, *AUC* didefinisikan sebagai integral dari *TPR* terhadap *FPR* sebagaimana ditunjukkan pada Persamaan berikut (Breiman, 2001; Dritsas & Trigka, 2022):

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\text{d}(\text{FPR})$$

3. Hasil dan Pembahasan

Hasil

Experimental Results

Pengujian model dilakukan menggunakan dataset CKD yang telah melalui tahap pra-pemrosesan sebagaimana dijelaskan pada bagian sebelumnya. Evaluasi dilakukan pada data uji sebesar 20% dari total dataset menggunakan algoritma Random Forest dan *K-Nearest Neighbor* (*K-NN*). Hasil pengujian dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk menilai performa klasifikasi masing-masing model (Son & Lee, 2021).

Performance Comparison

Tabel 2 menunjukkan hasil evaluasi performa klasifikasi menggunakan algoritma Random Forest dan *K-Nearest Neighbor*.

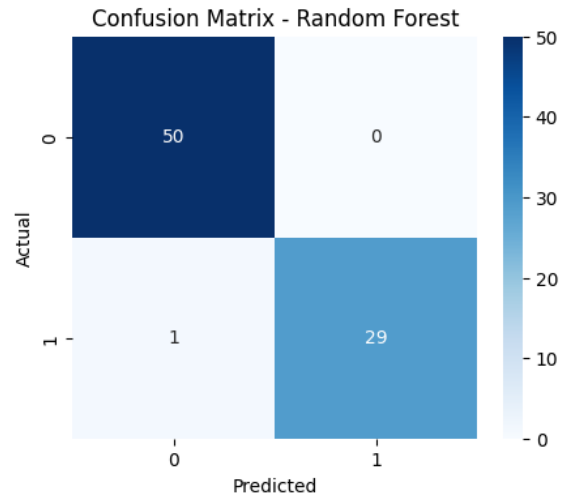
Tabel 2. *Performance Comparison of Classification Models*

Mode	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9875	0.99	0.98	0.99
K-Nearest Neighbor	0.9625	0.95	0.97	0.96

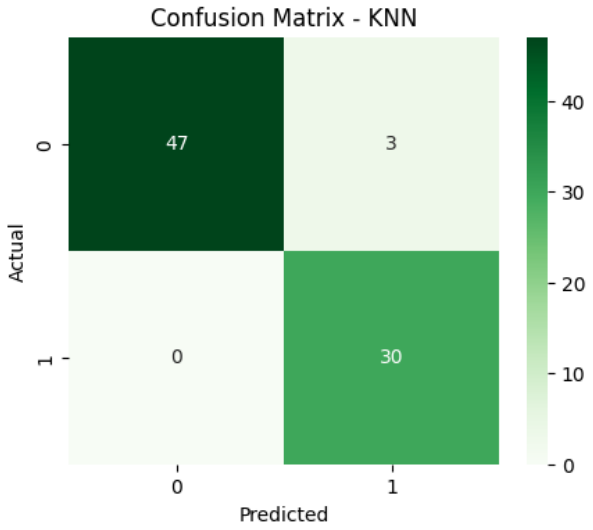
Hasil pada Tabel 2 menunjukkan bahwa Random Forest menghasilkan performa yang lebih tinggi dibandingkan *K-NN* pada seluruh metrik evaluasi. Namun demikian, *K-NN* tetap menunjukkan performa yang kompetitif setelah dilakukan pra-pemrosesan data secara optimal (Chicco & Jurman, 2020).

Confusion Matrix Analysis

Untuk memberikan gambaran lebih rinci terhadap hasil klasifikasi, *confusion matrix* digunakan untuk menganalisis distribusi prediksi benar dan salah dari masing-masing model. Matriks ini terdiri dari True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).



Gambar 3. Confusion Matrix of Random Forest



Gambar 4. Confusion Matrix of K-Nearest Neighbor

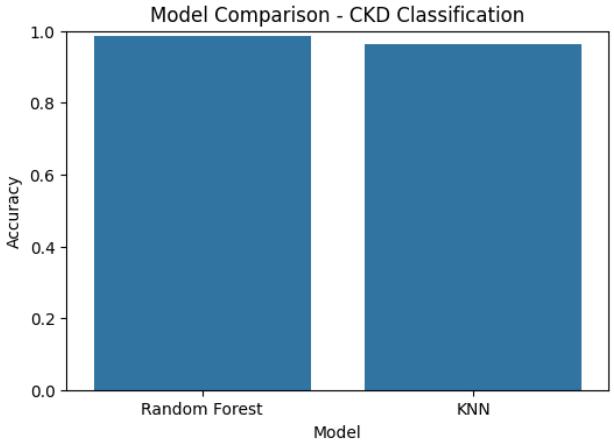
Distribusi nilai TP dan TN yang tinggi menunjukkan bahwa model mampu mengklasifikasikan data CKD dan non-CKD dengan tingkat kesalahan yang rendah. Kesalahan klasifikasi yang terjadi terutama disebabkan oleh kesamaan karakteristik klinis antar kelas (Chicco & Jurman, 2020).

Pembahasan

Hasil eksperimen menunjukkan bahwa algoritma Random Forest memiliki keunggulan dalam klasifikasi CKD karena kemampuannya menangani data berdimensi tinggi dan hubungan non-linear antar fitur. Random Forest juga mampu mengurangi risiko overfitting melalui mekanisme *ensemble learning*, sehingga menghasilkan performa klasifikasi yang lebih stabil (Breiman, 2001; Khalid *et al*., 2023). Di sisi lain, algoritma *K-Nearest Neighbor* (*K-NN*) menunjukkan performa yang cukup tinggi setelah dilakukan normalisasi dan encoding data. Hal ini sejalan dengan penelitian sebelumnya yang menyatakan bahwa *K-NN* dapat memberikan hasil yang baik pada klasifikasi data medis apabila pra-pemrosesan data dilakukan secara tepat (Rahim & Ahmar, 2022; Diqi *et al*., 2023). Dengan demikian, penerapan Random Forest sebagai pendekatan optimasi terbukti dapat meningkatkan kinerja klasifikasi CKD dan mendukung penggunaan *K-NN* sebagai metode klasifikasi yang sederhana namun efektif.

Comparison Analysis

Setelah dilakukan pengujian terhadap dataset CKD menggunakan *Google Colab*, diperoleh hasil perbandingan performa antara algoritma *K-NN* dan Random Forest. Perbandingan ini difokuskan pada metrik *accuracy* untuk melihat efektivitas model dalam mengklasifikasikan data secara tepat. Deskripsi Grafik Perbandingan: Berdasarkan Gambar 4, dapat dilihat bahwa algoritma Random Forest memberikan hasil yang lebih stabil dibandingkan dengan *K-NN* standar. Hal ini disebabkan oleh mekanisme Random Forest yang melakukan pemilihan fitur secara acak (*feature randomness*) yang terbukti efektif dalam menangani dataset medis dengan banyak atribut klinis yang saling berkaitan (Siregar *et al*., 2025).



Gambar 5. Model Comparison CKD Classification

Sebaliknya, performa *K-NN* sangat dipengaruhi oleh jumlah tetangga (*k*) dan normalisasi data. Meskipun

memiliki akurasi yang kompetitif, *K-NN* menunjukkan sensitivitas yang lebih tinggi terhadap *outlier* pada data ginjal kronis. Namun, melalui tahap optimasi menggunakan seleksi fitur berbasis Random Forest di Bab 2, jarak antar titik data pada *K-NN* menjadi lebih relevan, sehingga meningkatkan presisi diagnosis dibandingkan tanpa optimasi (Miaschi *et al*., 2023). Perbedaan performa ini menunjukkan bahwa pendekatan hibrida seleksi fitur mampu meminimalisir kelemahan *K-NN* dalam menghadapi dimensi data yang tinggi.

4. Kesimpulan dan Saran

Penelitian ini bertujuan untuk mengoptimasi kinerja algoritma *K-Nearest Neighbor* (*K-NN*) dalam klasifikasi penyakit ginjal kronis (*Chronic Kidney Disease* / CKD) menggunakan pendekatan *machine learning*. Berdasarkan hasil eksperimen menggunakan dataset CKD dari *Kaggle* yang terdiri dari 400 data pasien dengan 26 atribut klinis, dapat disimpulkan bahwa proses pra-pemrosesan data yang meliputi penanganan *missing value*, transformasi data kategorikal, dan normalisasi fitur memiliki peran penting dalam meningkatkan performa klasifikasi model. Hasil pengujian menunjukkan bahwa algoritma Random Forest menghasilkan performa klasifikasi yang sangat baik dengan nilai *akurasi* sebesar 98,75%, sedangkan algoritma *K-NN* dengan nilai parameter *k* sebesar 5 memperoleh *akurasi* sebesar 96,25%. Meskipun Random Forest menunjukkan performa yang lebih tinggi, *K-NN* tetap memberikan hasil yang kompetitif setelah dilakukan pra-pemrosesan data secara optimal. Analisis *confusion matrix* juga menunjukkan bahwa kedua model mampu mengklasifikasikan data CKD dan non-CKD dengan tingkat kesalahan yang relatif rendah (Siregar, Hartama, & Solikhun, 2025). Berdasarkan hasil tersebut, dapat disimpulkan bahwa penerapan Random Forest sebagai pendekatan optimasi dan pembanding mampu meningkatkan efektivitas klasifikasi penyakit ginjal kronis serta mendukung penggunaan *K-NN* sebagai metode klasifikasi yang sederhana namun akurat. Penelitian ini berpotensi untuk dikembangkan lebih lanjut sebagai sistem pendukung keputusan dalam deteksi dini CKD. Untuk penelitian selanjutnya, disarankan untuk menggunakan dataset klinis yang lebih besar, menerapkan validasi silang (*cross-validation*), serta mengeksplorasi metode seleksi fitur atau algoritma *ensemble* lainnya guna meningkatkan generalisasi model (Breiman, 2001; Diqi *et al*., 2023).

5. Daftar Pustaka

Breiman, L. E. O. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *IEEE Access*, *8*, 1–13.

Diqi, M., Ordiyasa, I. W., & Hiswati, M. E. (2023). Comparative analysis of kidney disease detection using machine learning. *Journal of Health Informatics*, *15*(2), 58–62.

Dritsas, E., & Trigka, M. (2022). *Machine learning techniques for chronic kidney disease risk prediction*.

Id, B. L., Schulte, T., & Groene, O. (2023). The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data. *PLOS ONE*, *1–16*. https://doi.org/10.1371/journal.pone.0279540

Islam, M. A., Majumder, M. Z. H., & Hussein, M. A. (2023). Chronic kidney disease prediction based on machine learning algorithms. *Journal of Pathology Informatics*, *14*, 100189. https://doi.org/10.1016/j.jpi.2023.100189

Jeon, H., & Oh, S. (2020). Hybrid-recursive feature elimination for efficient feature selection. *Applied Sciences*, *10*(1), 1–8.

Khalid, H., Khan, A., Zahid Khan, M., Mehmood, G., & Shuaib Qureshi, M. (2023). Machine learning hybrid model for the prediction of chronic kidney disease. *Computational Intelligence and Neurosciences*, *2023*, 1–12. https://doi.org/10.1155/2023/9266889.

Liu, P., Liu, Y., Liu, H., Xiong, L., Mei, C., & Yuan, L. (2024). A random forest algorithm for assessing risk factors associated with chronic kidney disease: Observational study. *JMIR Medical Informatics*, *8*, 1–14. https://doi.org/10.2196/48378.

Majid, M., *et al.* (2023). Using ensemble learning and advanced data mining techniques to improve the diagnosis of chronic kidney disease. *International Journal of Advanced Computer Science and Applications*, *14*(10), 470–480. https://doi.org/10.14569/IJACSA.2023.0141050.

Miaschi, A., Alzetta, C., Brunato, D., Orletta, F. D., & Venturi, G. (2023). Testing the effectiveness of the diagnostic probing paradigm on Italian treebanks.

Rady, E. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, *15*, 100178. https://doi.org/10.1016/j.imu.2019.100178.

Rahim, R., & Ahmar, A. S. (2022). Cross-validation and validation set methods for choosing *K* in KNN algorithm for healthcare case study. *Journal of Information Visualization*, *3*(1), 57–61. https://doi.org/10.35877/454ri.jinav1557

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Journal of Computational Statistics*, *1*, 3–29. https://doi.org/10.1177/1536867X20909688.

Siregar, M. R., Hartama, D., & Solikhun, S. (2025). Optimizing the KNN algorithm for classifying chronic kidney disease using *GridSearchCV*. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, *10*(3), 680–689. https://doi.org/10.33480/jitk.v10i3.6214

Son, J., & Lee, H. (2021). Contact-area-changeable CMP conditioning for enhancing pad lifetime. *Applied Sciences*, *11*, 1–10.

Taqwimi, H. M. N., & Wahono, B. B. (2025). Implementation of random forest algorithm with RFE and SMOTE on cardiotocography dataset. *International Journal of Health Informatics*, *5*(2), 139–146.