Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds

Evaluasi *Random Forest* dan SMOTE dalam Menangani Ketidakseimbangan Kelas pada Klasifikasi *Chronic Kidney Disease*

Evaluation of Random Forest and SMOTE for Addressing Class Imbalance in Chronic Kidney Disease Classification

Rahmat Dwi Cahyono[1,*], Herman Yuliasnyah[2], Sunardi[3]

[1] Program Studi Magister Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, 55164
[2] Program Studi Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia 55191
[3] Program Studi Doktor Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia 55191
Email: [1,*]rahmatdwicahyono15@gmail.com, [2]herman.yuliansyah@uad.ac.id, [3]sunardi@uad.ac.id
Email Penulis Korespondensi: rahmatdwicahyono15@gmail.com

Received: **25/06/2026**, Revised: **19/10/2026**, Accepted: **28/10/2026**, Published: **31/10/2026**

Abstrak−Chronic Kidney Disease (CKD) merupakan penyakit kronis yang ditandai dengan penurunan fungsi ginjal secara progresif sehingga memerlukan deteksi dini untuk mengurangi risiko komplikasi. Penerapan algoritma *machine learning* seperti Random Forest dapat mendukung klasifikasi CKD, tetapi ketidakseimbangan kelas pada dataset medis berpotensi menurunkan kemampuan model dalam mengenali pasien yang menderita CKD. Penelitian ini bertujuan untuk mengevaluasi pengaruh penerapan *Synthetic Minority Over-sampling Technique* (SMOTE) terhadap kinerja algoritma Random Forest dalam klasifikasi CKD. Dataset yang digunakan terdiri atas 4.000 data pasien dengan 26 atribut yang mencakup karakteristik demografis, riwayat kesehatan, dan faktor gaya hidup. Tahapan penelitian meliputi *Exploratory Data Analysis* (EDA), pembagian data dengan rasio 80:20, penerapan SMOTE pada data latih, pemodelan Random Forest, serta evaluasi menggunakan *confusion matrix*, *classification report*, dan *10-Fold Cross-Validation*. Hasil penelitian menunjukkan bahwa sebelum penerapan SMOTE, model memperoleh *accuracy* sebesar 96%, *precision* 100%, *recall* 59%, dan *F1-score* 74% pada kelas CKD. Setelah penerapan SMOTE, *accuracy* meningkat menjadi sekitar 99%, sedangkan *precision*, *recall*, dan *F1-score* kelas CKD masing-masing mencapai 97%. Hasil *confusion matrix* menunjukkan penurunan jumlah *false negative* dari 16 menjadi 1 kasus, sementara pengujian *10-Fold Cross-Validation* memperlihatkan peningkatan akurasi dan kestabilan performa model. Hasil tersebut menunjukkan bahwa penerapan SMOTE mampu meningkatkan sensitivitas Random Forest terhadap kelas minoritas CKD serta berpotensi mendukung pengembangan sistem deteksi dini penyakit ginjal kronis. Penelitian ini mengevaluasi efektivitas SMOTE dalam meningkatkan kinerja Random Forest, khususnya dalam mendeteksi kelas minoritas dan mengurangi *false negative* pada klasifikasi penyakit ginjal kronis (CKD).

Kata Kunci*: Chronic Kidney Disease (CKD); Random Forest;* Ketidakseimbangan Kelas*; SMOTE*

Abstract−Chronic Kidney Disease (CKD) is a chronic condition characterized by the progressive decline of kidney function, requiring early detection to reduce the risk of complications. Machine learning algorithms such as Random Forest can support CKD classification; however, class imbalance in medical datasets may reduce the model's ability to identify patients with CKD. This study aims to evaluate the effect of the Synthetic Minority Over-sampling Technique (SMOTE) on the performance of the Random Forest algorithm in CKD classification. The dataset consists of 4,000 patient records with 26 attributes covering demographic characteristics, medical history, and lifestyle factors. The research stages include Exploratory Data Analysis (EDA), data splitting with an 80:20 ratio, SMOTE implementation on the training data, Random Forest modeling, and performance evaluation using a confusion matrix, classification report, and 10-Fold Cross-Validation. The results show that before SMOTE implementation, the model achieved an accuracy of 96%, precision of 100%, recall of 59%, and an F1-score of 74% for the CKD class. After SMOTE implementation, accuracy increased to approximately 99%, while precision, recall, and F1-score for the CKD class each reached 97%. Confusion matrix analysis revealed a reduction in false negatives from 16 to 1 case, while 10-Fold Cross-Validation demonstrated improved accuracy and more stable model performance. These findings indicate that SMOTE improves the sensitivity of Random Forest in identifying minority-class CKD cases and has potential applications in developing early detection systems for chronic kidney disease. This study evaluates the effectiveness of SMOTE in improving Random Forest performance, particularly in detecting minority classes and reducing false negatives in Chronic Kidney Disease (CKD) classification.

Keywords: Chronic Kidney Disease (CKD); Classification; Class Imbalance; Random Forest; SMOTE

1. PENDAHULUAN

Chronic Kidney Disease (CKD) merupakan kondisi penurunan fungsi ginjal yang bersifat progresif dan tidak dapat dipulihkan, berlangsung lebih dari tiga bulan. Sekitar 10% populasi dunia diperkirakan menderita CKD, dan penyakit ini diprediksi menjadi salah satu penyebab kematian tertinggi secara global pada tahun 2050 [1] . Di Indonesia, penyakit ginjal menempati posisi kedua setelah penyakit jantung dalam hal pembiayaan kesehatan yang ditanggung oleh BPJS [2] dengan pengeluaran mencapai triliunan rupiah setiap tahunnya. Tingginya beban biaya tersebut disebabkan oleh lamanya perawatan yang harus dijalani pasien, baik melalui dialisis maupun transplantasi ginjal. Teknik klasifikasi berbasis *Machine Learning* telah banyak digunakan dalam mendiagnosis berbagai penyakit maupun menentukan tindakan

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 64
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds

perawatan [3]. Kemajuan *Machine Learning* memungkinkan pengolahan data medis yang besar dan kompleks secara cepat dan akurat untuk mengklasifikasikan kondisi pasien, seperti mendeteksi CKD berdasarkan fitur klinis dan demografis, sehingga mendukung pengambilan keputusan klinis dan membantu tenaga medis dalam pengambilan keputusan [4][5]. Di antara berbagai algoritma ML: *Random Forest*, *Decision Tree*, *Naive Bayes* merupakan metode yang sering digunakan pada studi klasifikasi penyakit, *Random Forest* juga disebut sebagai induk dari *Decision Tree*, karena pada dasarnya mengambil beberapa pohon yang telah tersusun pada tahapan *Decision Tree*. Kumpulan pohon tersebut digunakan untuk mengklasifikasi data ke suatu kelas [6][7].

Penelitian terdahulu yang dilakukan oleh Bhirawa dan Sanjaya pada tahun 2025 melakukan penelitian optimasi klasifikasi penyakit ginjal kronis CKD menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi ketidakseimbangan data. Penelitian tersebut menggunakan 400 data pasien dengan membandingkan algoritma Random Forest dan Gradient Boosting. Temuan tersebut menunjukkan bahwa penanganan ketidakseimbangan kelas memiliki peran penting dalam meningkatkan ketepatan klasifikasi penyakit ginjal kronis. Penelitian ini juga menjadi dasar bahwa pemilihan algoritma dan teknik pengolahan data yang tepat dapat mendukung pengembangan sistem deteksi dini penyakit ginjal kronis[3].

Penelitian berikutnya dilakukan oleh oleh Riansyah et al. pada tahun 2025 dengan judul Perbandingan Akurasi Algoritma Decision Tree dengan Random Forest dalam Diagnosa Penyakit Hepatitis menggunakan 155 data dengan 20 atribut. Penelitian tersebut menerapkan pembagian data sebesar 80% untuk pelatihan dan 20% untuk pengujian. Hasil tersebut menunjukkan bahwa Random Forest memiliki kemampuan klasifikasi yang lebih baik dibandingkan Decision Tree pada dataset yang digunakan. Temuan ini memperkuat potensi penerapan algoritma Random Forest dalam klasifikasi penyakit berbasis data medis, termasuk sebagai pendekatan yang dapat dikaji lebih lanjut dalam deteksi penyakit ginjal kronis[8].

Penelitian yang dilakukan oleh Desmita et al. pada tahun 2025 dengan topik penelitian membandingkan algoritma Decision Tree dan Random Forest dalam klasifikasi penyakit diabetes melitus dengan menerapkan metode *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi ketidakseimbangan data. Penelitian tersebut menjadi referensi dalam penerapan algoritma *machine learning* untuk klasifikasi penyakit berbasis data medis[9].

Selanjutnya, Penelitian yang dilakukan oleh Mahing et al. pada tahun 2024 membahas penerapan algoritma *Support Vector Machine* (SVM) dan *Random Forest* dalam klasifikasi tingkat stres berdasarkan data teks yang diperoleh dari media sosial Twitter. Penelitian tersebut menggunakan pendekatan *machine learning* dengan membandingkan beberapa metode transformasi fitur, yaitu TF-IDF, CountVectorizer, NRCLex, dan *Word Affect Intensities*. Proses klasifikasi dilakukan melalui tahapan *preprocessing*, transformasi data, pembentukan model, serta evaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score. Penelitian ini menunjukkan pentingnya pemilihan algoritma dan metode pengolahan fitur dalam membangun model klasifikasi yang optimal. Pendekatan tersebut dapat menjadi referensi dalam pengembangan model klasifikasi berbasis *machine learning*, khususnya dalam menentukan metode pengolahan data dan evaluasi kinerja algoritma

Berdasarkan penelitian-penelitian terdahulu, penelitian ini menyusun eksperimen dan evaluasi untuk evaluasi kinerja *Random Forest* dalam mengklasifikasikan CKD dengan SMOTE, dengan harapan penelitian ini dapat memberikan keputusan klinis dalam menentukan klasifikasi penyakit pada sistem deteksi dini penyakit ginjal kronis.

Kesenjangan penelitian ini terletak pada kebutuhan untuk mengevaluasi secara kuantitatif pengaruh penerapan SMOTE terhadap sensitivitas *Random Forest* dalam mengidentifikasi pasien CKD pada dataset yang digunakan. Ketidakseimbangan distribusi kelas berpotensi memengaruhi proses pembelajaran model, tetapi tingkat pengaruhnya perlu dibuktikan melalui perbandingan performa sebelum dan sesudah penyeimbangan data. SMOTE dipilih karena mampu membentuk sampel sintetis pada kelas minoritas sehingga dapat meningkatkan representasi kelas tersebut selama pelatihan. Kombinasi *Random Forest* dan SMOTE diharapkan dapat meningkatkan kemampuan deteksi kasus positif tanpa mengabaikan ketepatan klasifikasi pada kelas lainnya. Keunggulan pendekatan ini dievaluasi melalui perubahan nilai *accuracy*, *precision*, *recall*, *F1-score*, serta jumlah kesalahan *false positive* dan *false negative*.

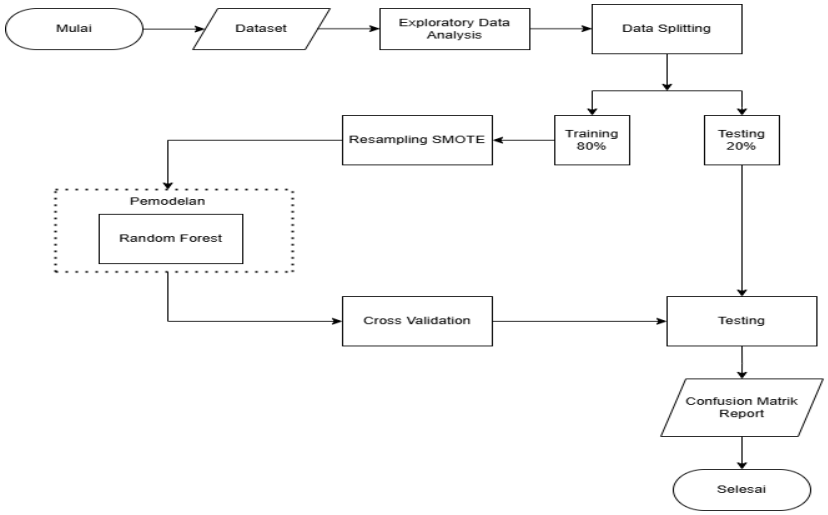
Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengevaluasi kinerja algoritma *Random Forest* sebelum dan sesudah penerapan SMOTE dalam klasifikasi penyakit CKD berdasarkan faktor demografis, riwayat kesehatan, dan gaya hidup pasien. Penelitian dilakukan melalui tahapan *Exploratory Data Analysis* (EDA), pembagian data latih dan data uji dengan rasio 80:20, penerapan SMOTE pada data latih, serta evaluasi menggunakan *confusion matrix*, *classification report*, dan *10-Fold Cross-Validation*. Dua skenario eksperimen digunakan, yaitu Random Forest tanpa SMOTE sebagai model dasar (*baseline*) dan Random Forest dengan SMOTE sebagai model perlakuan. Kontribusi penelitian ini adalah memberikan evaluasi empiris mengenai pengaruh penyeimbangan data terhadap sensitivitas Random Forest, khususnya dalam meningkatkan kemampuan mendeteksi pasien CKD dan mengurangi jumlah *false negative* sebagai salah satu pertimbangan pengembangan sistem pendukung deteksi dini penyakit ginjal kronis.

2. METODOLOGI PENELITIAN

Tahapan penelitian meliputi akuisisi data, *exploratory data analysis*, pembagian data, pemodelan menggunakan algoritma *Random Forest*, *Cross Validation*, serta evaluasi performa model dilakukan menggunakan *confusion matrix* dan *classification report* yang mencakup metrik *accuracy, precision, recall, dan F1-score*, seperti pada Gambar 1.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 65
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 1. Tahapan Penelitian

2.1 Dataset

Penelitian ini menggunakan pendekatan kuantitatif berbasis data *Chronic Kidney Disease* (CKD) menggunakan dataset (*A Dataset of Demographic and Lifestyle Risk Factors for Assessing Chronic Kidney Disease Development in Diabetic Patients*) yang tersedia publik di https://data.mendeley.com/datasets/hjkzgbxgv5/2 dan dipublikasikan oleh Anan et al., [10]. Dataset ini, yang dikumpulkan dari Rumah Sakit Umum BIRDEM, Dhaka, menyajikan kompilasi komprehensif faktor demografis dan gaya hidup untuk memprediksi perkembangan CKD pada pasien diabetes [10]. Tahap ini mencakup identifikasi masalah dengan memahami substansi dalam data serta mencari aspek menarik yang dapat digunakan untuk merumuskan hipotesis awal [11].

2.2 *Exploratory Data Analysis* (EDA)

EDA dataset dilakukan untuk memperoleh pemahaman mengenai struktur data, karakteristik distribusi variabel, serta adanya untuk mengidentifikasi potensi ketidakseimbangan (*imbalance*) kelas target. Dataset yang digunakan berjumlah 4.000 data pasien penyakit CKD. Variabel target adalah diabetes yang berisi label 0 (tidak menderita CKD) dan 1 (menderita CKD) [9]. Proses ini memiliki peran penting karena membantu dalam memahami pernyataan masalah serta hubungan antara berbagai fitur data sebelum data digunakan dalam pemodelan. EDA sendiri merupakan bagian dari tahapan dalam *data science* [12][13].

2.3 Spliting data

Setelah data siap digunakan, dilakukan pembagian dataset menjadi data latih dan data uji. Pembagian ini bertujuan untuk menguji kemampuan generalisasi model terhadap data baru yang belum pernah dilihat sebelumnya. Umumnya digunakan rasio 80% data latih dan 20% data uji agar model memiliki cukup data untuk belajar sekaligus cukup data untuk evaluasi, menghasilkan data set 800 dan untuk data training menghasilkan 3200 data [14].

2.4 *Resampling* SMOTE

Penanganan *class imbalance* menggunakan metode *Synthetic Minority OverSampling Technique* (SMOTE) untuk mengatasi ketidakseimbangan antara kelas target, sehingga model dapat mempelajari pola dengan lebih baik. Dengan melakukan tahap ini, data siap digunakan untuk tahap selanjutnya dalam analisis dan pemodelan [15][16].

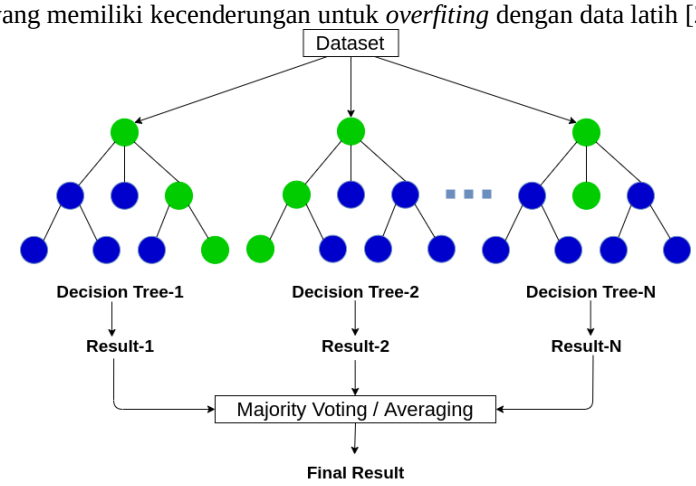
2.5 Pemodelan

Setelah melalui tahap pre-processing, langkah berikutnya adalah membangun model prediksi menggunakan algoritma *Random Forest*. Model dilatih menggunakan data latih yang telah disiapkan pada masing masing skenario, dan diuji menggunakan data uji untuk mengukur performa klasifikasinya [17]. *Random forest* adalah teknik klasifikasi yang dikembangkan dengan memanfaatkan metode *Decision Tree* [18]. Dalam metode ini, pemilihan atribut pada setiap node dilakukan secara acak untuk menentukan klasifikasi [19][20] . Proses klasifikasi ini mengandalkan suara terbanyak dari hasil yang dikembalikan oleh pohon-pohon keputusan. *Random forest* dapat dibangun dengan memanfaatkan teknik Bagging (*Bootstrap Aggregating*) dan pemilihan atribut secara acak untuk membangun sejumlah model yang serupa namun berbeda dalam struktur dan data latihnya. Setelah semua pohon selesai dilatih, hasil prediksi dari masing-masing pohon akan digabungkan menggunakan teknik *votting* (untuk klasifikasi) atau rata-rata (*Averaging*) untuk Regresi. Proses pertumbuhan pohon keputusan dilakukan menggunakan metode *Classification and Regression Tree* (CART). Pohon-pohon ini dibiarkan tumbuh hingga mencapai ukuran maksimal tanpa dilakukan pemangkasan. Hasilnya adalah

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 66
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds

sekumpulan pohon keputusan yang kemudian disebut sebagai hutan (*Forest*). *Random forest* secara umum memperbaiki karakter dari *Decision Tree* yang memiliki kecenderungan untuk *overfiting* dengan data latih [21].



Gambar 2. *Random Forest*

2.6 *Cross-validation*

Cross validation merupakan teknik statistik yang digunakan dalam *Machine Learning* dan pemodelan prediktif lainnya untuk menilai kinerja dan kemampuan generalisasi suatu model. Teknik ini memberikan estimasi performa model yang lebih akurat pada data yang tidak terlihat [22]. Manfaat penting lainnya dari *cross validation*, yakni membantu *data analyst* mengatasi masalah *overfitting* atau kondisi saat model terlalu spesifik pada data pelatihan sehingga kurang baik dalam menganalisis data baru [23]. Skema *fold cross-validation* digunakan untuk mengevaluasi performa model.

2.7 Evaluasi

Evaluasi model dalam penelitian ini dilakukan untuk mengukur kinerja dari algoritma *Machine Learning* dalam melakukan klasifikasi terhadap penyakit CKD. Beberapa metrik evaluasi yang digunakan dalam penelitian ini adalah *accuracy, precision, recall, dan F1-score* [24]. Keempat metrik tersebut dipilih karena dapat memberikan gambaran yang komprehensif terhadap kinerja model, khususnya dalam konteks prediksi medis yang membutuhkan tingkat ketelitian tinggi, terutama dalam mengenali kasus positif. Proses evaluasi dilakukan pada data uji yang telah dipisahkan sebelumnya dari data latih, sehingga hasil yang diperoleh dapat mencerminkan kemampuan model dalam menghadapi data baru secara objektif [25]. Akurasi mengukur proporsi prediksi yang benar terhadap total jumlah data. Rumus akurasi adalah:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

dimana TP (*True Positive*) adalah pasien (CKD) yang berhasil dideteksi sebagai penderita CKD, TN (*True Negative*) adalah pasien non-CKD yang berhasil dideteksi sebagai bukan penderita CKD, FP (*False Positive*) adalah pasien non-CKD yang salah dideteksi sebagai penderita CKD dan FN (*False Negative*) adalah pasein CKD yang salah dideteksi sebagai bukan penderita CKD. Akurasi memberikan gambaran umum kinerja model, tetapi pada dataset tidak seimbang, metrik ini sering menyesatkan karena model dapat mencapai akurasi tinggi hanya dengan memprediksi kelas mayoritas. *Precision* mengukur akurasi prediksi positif. Rumusnya adalah:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Precision penting untuk meminimalkan *false positives*, terutama dalam konteks diagnosis medis. *Precision* tinggi menunjukkan bahwa sebagian besar pasien yang diprediksi menderita penyakit CKD. *Recall* mengukur kemampuan model untuk mendeteksi semua kasus yang sebenarnya positif, yaitu pasien yang benar-benar menderita penyakit CKD. Rumusnya adalah:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Recall sangat penting dalam konteks deteksi dini CKD untuk memastikan tidak ada pasien CKD yang terlewatkan dan tidak terdiagnosis (*false negatives*). *Recall* tinggi menunjukkan bahwa model berhasil mendeteksi sebagian besar pasien yang benar-benar menderita CKD yang sebenarnya. *F1-score* adalah rata - rata harmonik dari *precision* dan *recall*. Rumusnya adalah:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 67
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds

F1- score digunakan untuk menyeimbangkan *precision* dan *recall*, terutama ketika keduanya memiliki nilai yang sangat berbeda. Nilai *F1- score* yang tinggi menunjukkan model yang efektif dalam mengklasifikasikan penyakit CKD tanpa terlalu banyak menghasilkan *false positives* atau *false negatives*.

3. HASIL DAN PEMBAHASAN

Bagian ini melampirkan model klasifikasi berbasis *Machine Learning* yang mampu mendeteksi CKD berdasarkan data klinis pasien. Disini penelitian menguji seberapa optimal kinerja algoritma *Random Forest* dalam klasifikasi penyakit pada data pasien.

3.1 Data Eksperimen

Dataset yang digunakan berjumlah 4.000 data dengan 26 atribut data penyakit CKD. data pasien penyakit CKD yang mencakup fitur demografis, antropometri, riwayat diabetes, gaya hidup, penyakit penyerta (komorbid), dan CKD (1 = yes, 0 = no) sebagai label target. Data tersebut menyajikan kompilasi komprehensif faktor demografis dan gaya hidup untuk memprediksi perkembangan CKD pada pasien diabetes. Rincian atribut yang digunakan dalam penelitian ini disajikan pada Tabel 1. Tabel 1 menyajikan variabel-variabel yang digunakan dalam penelitian untuk membangun model klasifikasi *Chronic Kidney Disease* (CKD). Dataset terdiri atas 10 variabel yang mencakup karakteristik demografis, riwayat kesehatan, gaya hidup, dan kondisi klinis pasien. Variabel demografis meliputi *gender*, *job*, dan *age*, yang menggambarkan jenis kelamin, jenis pekerjaan, serta usia pasien. Variabel riwayat kesehatan terdiri atas *family background of diabetes*, *diabetic year*, dan *hypertension*, yang digunakan untuk merepresentasikan faktor risiko yang berkaitan dengan penyakit diabetes dan hipertensi sebagai kondisi yang berpotensi meningkatkan risiko CKD. Selain itu, dataset juga memuat variabel yang berkaitan dengan gaya hidup dan kepatuhan pengobatan, yaitu *Follow Suggested Diet* dan *Take Medicine for Diabetes*. Kedua variabel tersebut menggambarkan tingkat kepatuhan pasien terhadap anjuran medis dalam pengelolaan diabetes. Variabel *Body Mass Index (BMI*) digunakan untuk merepresentasikan kondisi antropometri pasien yang berkaitan dengan status berat badan dan risiko penyakit metabolik. Variabel target dalam penelitian ini adalah *CKD*, yang memiliki nilai biner dengan label 1 untuk pasien yang terdiagnosis CKD dan 0 untuk pasien yang tidak terdiagnosis CKD. Berdasarkan Tabel 1 jenis datanya, dataset terdiri atas variabel kategorikal, numerik, dan biner sehingga memerlukan tahapan pra-pemrosesan sebelum digunakan dalam proses pemodelan menggunakan algoritma *Random Forest*. Kombinasi variabel-variabel tersebut diharapkan mampu memberikan informasi yang cukup untuk mengidentifikasi pola dan faktor yang berpengaruh terhadap kejadian CKD pada pasien.

Tabel 1. Variabel Penelitian

No	Nama Variabel	Jenis Variabel	Keterangan
1	*Patient ID*	Numerik	Nomor identitas pasien
2	*Gender*	Kategorikal	Jenis kelamin pasien (M = *Male*, F = *Female*)
3	*Job*	Numerik	Jenis pekerjaan (1 = *normal*, 2 = *intermediate*, 3 = *heavy*)
4	*Family Background of Diabetes*	Biner	Riwayat keluarga diabetes (1 = ya, 0 = tidak)
5	*Height (cm)*	Numerik	Tinggi badan pasien dalam sentimeter
6	*Diabetic Year*	Numerik	Lama menderita diabetes (tahun)
7	*Age*	Numerik	Usia pasien (tahun)
8	*Average Age*	Numerik	Rata-rata usia pasien
9	*Weight (kg)*	Numerik	Berat badan pasien dalam kilogram
10	*Average Weight (kg)*	Numerik	Rata-rata berat badan pasien dalam kilogram
11	*BMI*	Numerik	*Body Mass Index* pasien
12	*Follow Suggested Diet*	Biner	Kepatuhan terhadap diet yang disarankan (1 = ya, 0 = tidak)
13	*Take Medicine for Diabetes*	Biner	Konsumsi obat diabetes (1 = ya, 0 = tidak)
14	*Take Insulin*	Biner	Penggunaan insulin (1 = ya, 0 = tidak)
15	*Hypertension*	Biner	Riwayat hipertensi (1 = ya, 0 = tidak)
16	*Heart Disease*	Biner	Riwayat penyakit jantung (1 = ya, 0 = tidak)
17	*Sleep*	Biner	Kecukupan tidur (1 = *sufficient*, 0 = *insufficient*)
18	*Water Consumption*	Biner	Kecukupan konsumsi air (1 = *sufficient*, 0 = *insufficient*)
19	*Smoke*	Biner	Kebiasaan merokok (1 = ya, 0 = tidak)
20	*Zarda, Betel Leaf*	Biner	Kebiasaan mengonsumsi *zarda*/*betel leaf* (1 = ya, 0 = tidak)
21	*Walk Regularly*	Biner	Kebiasaan berjalan kaki secara rutin (1 = ya, 0 = tidak)
22	*Urination Properly*	Biner	Kondisi buang air kecil normal/lancar (1 = ya, 0 = tidak)
23	*Urinary Infection*	Biner	Riwayat infeksi saluran kemih (1 = ya, 0 = tidak)

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 68
This work is licensed under a Creative Commons Attribution 4.0 International License.

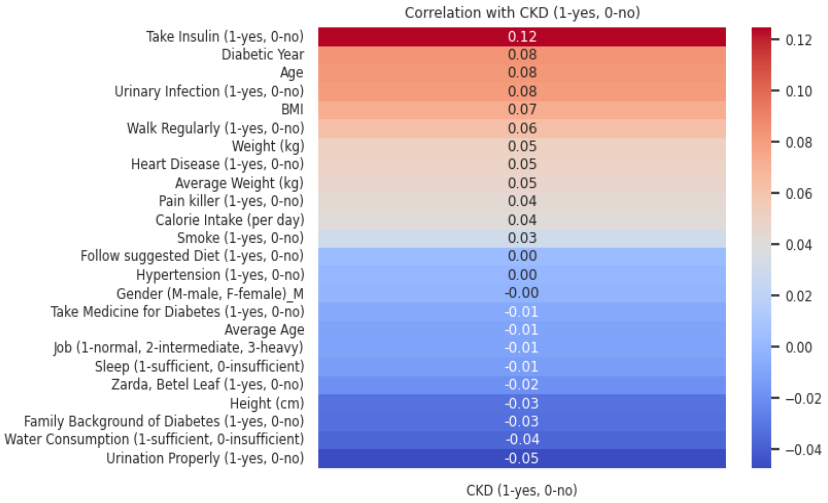
Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds

24	*Pain Killer*	Biner	Konsumsi obat pereda nyeri (1 = ya, 0 = tidak)
25	*Calorie Intake (per day)*	Numerik	Jumlah asupan kalori per hari
26	*CKD*	Biner	Status CKD (1 = ya, 0 = tidak)

3.2 Hasil Eksperimen

3.2.1 *Exploratory Data Analysis* (EDA)

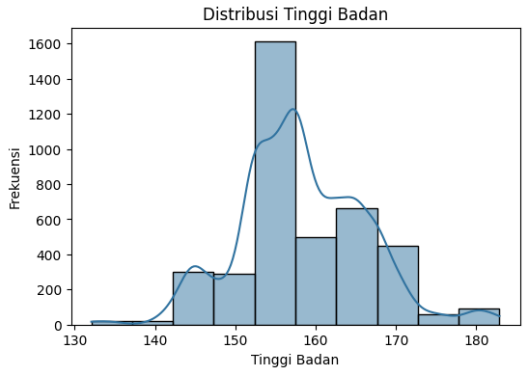
Analisis korelasi antara fitur dan variabel target ditampilkan pada Gambar 3. Tahap *Exploratory Data Analysis* (EDA) dalam penelitian ini dilakukan terhadap 4.000 data pasien untuk memahami struktur, karakteristik distribusi, dan korelasi antar variabel klinis terkait penyakit CKD. Melalui visualisasi matriks korelasi, diketahui bahwa faktor seperti usia, jenis kelamin, BMI, dan Riwayat keluarga diabetes, memiliki korelasi positif terhadap diagnosis CKD, yang memberikan petunjuk awal mengenai variabel paling berpengaruh. Selain itu, analisis distribusi pada variabel tinggi badan dan usia menunjukkan pola yang mendekati distribusi normal, di mana dataset didominasi oleh kelompok usia menengah hingga lanjut yang secara medis memiliki risiko lebih tinggi terhadap penurunan fungsi ginjal. Keseluruhan proses EDA ini berfungsi untuk mengidentifikasi adanya ketidakseimbangan kelas target serta menjadi landasan bagi peneliti dalam menentukan teknik pra-pemrosesan data yang tepat sebelum masuk ke tahap pemodelan algoritma.



Gambar 3. Korelasi Fitur terhadap Diagnosis CKD

3.2.2 Distribusi Tinggi Badan

Pada bagian distribusi tinggi badan, data menunjukkan pola yang mendekati normal dengan frekuensi tertinggi berada pada rentang nilai tengah. Visualisasi pada Gambar 4 memperlihatkan bahwa mayoritas individu dalam dataset memiliki tinggi badan di kisaran 150 cm hingga 160 cm, di mana puncak frekuensi berada pada angka sekitar 155 cm hingga 157 cm dengan jumlah lebih dari 1.600 data. Sebaran data ini menunjukkan adanya konsentrasi massa data pada nilai tertentu yang kemudian melandai ke arah kiri (nilai lebih rendah) dan kanan (nilai lebih tinggi), yang memberikan gambaran karakteristik fisik dari 4.000 sampel pasien yang digunakan dalam penelitian klasifikasi CKD ini.



Gambar 4 Distribusi Tinggi Badan

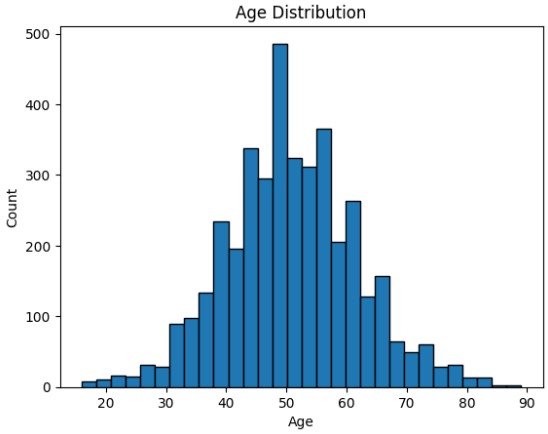
3.2.3 Distribusi Usia

Bagian distribusi usia dalam penelitian ini menunjukkan bahwa data 4.000 pasien memiliki pola yang mendekati distribusi normal dengan dominasi pada kelompok usia menengah hingga lanjut. Berdasarkan visualisasi pada Gambar 5, frekuensi tertinggi berada pada rentang usia 50 tahun, di mana sebaran data mencakup individu dari usia sekitar 20 hingga hampir

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 69
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds

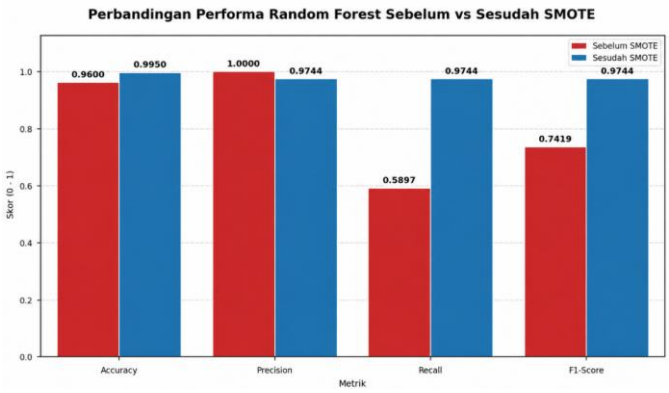
90 tahun. Karakteristik distribusi ini mengindikasikan bahwa faktor usia merupakan variabel yang berpotensi memberikan kontribusi signifikan terhadap prediksi penyakit, mengingat risiko kesehatan yang berkaitan dengan fungsi ginjal umumnya meningkat seiring bertambahnya usia pasien.



Gambar 5 Distribusi Usia

3.2.4 Resampling SMOTE

Berdasarkan Gambar 6, penerapan metode Synthetic Minority Over-sampling Technique (SMOTE) pada algoritma Random Forest menunjukkan peningkatan performa klasifikasi penyakit Chronic Kidney Disease (CKD). Sebelum penerapan SMOTE, model memperoleh nilai accuracy sebesar 96%, precision sebesar 100%, recall sebesar 59%, dan F1-score sebesar 74%. Setelah penerapan SMOTE, nilai accuracy meningkat menjadi 99%, sedangkan precision sedikit menurun menjadi 97%. Peningkatan paling signifikan terlihat pada recall, yaitu dari 58% menjadi 97%, serta F1-score yang meningkat dari 74% menjadi 97%. Peningkatan recall tersebut menunjukkan bahwa model setelah penerapan SMOTE memiliki kemampuan yang lebih baik dalam mengidentifikasi kasus positif CKD sehingga dapat mengurangi potensi kesalahan klasifikasi false negative. Meskipun terjadi sedikit penurunan precision, peningkatan recall dan F1-score mengindikasikan adanya keseimbangan yang lebih baik antara kemampuan model dalam mendeteksi kasus positif dan ketepatan hasil prediksi. Dengan demikian, berdasarkan hasil eksperimen yang ditampilkan, penerapan SMOTE memberikan peningkatan performa klasifikasi Random Forest, khususnya dalam mengenali kelas minoritas CKD.



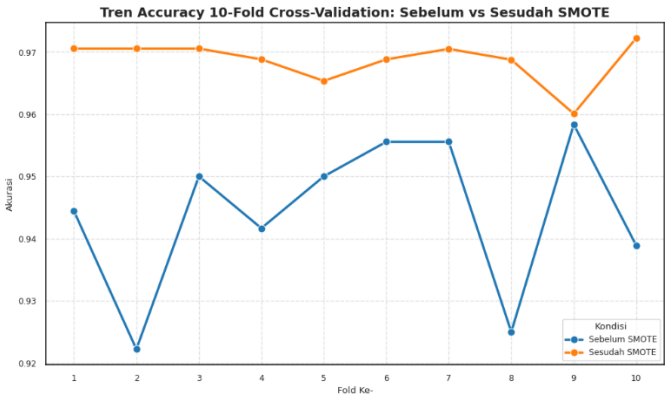
Gambar 6. Perbandingan Performa *Random Forest* Sebelum dan Sesudah SMOTE

3.2.5 *Cross Validation*

Penelitian ini menerapkan *Cross Validation* dengan Stratified KFold untuk masing - masing algoritma dengan k = 10. Berdasarkan gambar 7 Stratified KFold membagi dataset menjadi 10 bagian yang seimbang pada setiap iterasi dengan satu bagian dijadikan data test dan empat bagian lainnya sebagai data train. hasil pengujian menggunakan 10-Fold Cross-Validation menunjukkan bahwa algoritma Random Forest setelah penerapan SMOTE menghasilkan akurasi yang lebih tinggi dan relatif stabil dibandingkan sebelum penerapan SMOTE. Sebelum SMOTE, akurasi berkisar antara 92% hingga 95%, sedangkan setelah SMOTE meningkat menjadi sekitar 96% hingga 97%. Hasil ini menunjukkan bahwa penerapan SMOTE berkaitan dengan peningkatan akurasi klasifikasi CKD pada seluruh fold pengujian. Namun, peningkatan akurasi tersebut tetap perlu dievaluasi bersama nilai recall dan F1-score kelas CKD untuk memastikan kemampuan model dalam mendeteksi kasus positif.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 70
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 7. Perbandingan Akurasi *Random Forest* Sebelum dan Sesudah Penerapan SMOTE Menggunakan 10-*Fold Cross-Validation*

Dari hasil pengujian menggunakan metode 10-Fold *Cross-Validation* menunjukkan adanya perbedaan performa algoritma *Random Forest* sebelum dan sesudah penerapan SMOTE dalam klasifikasi penyakit CKD. Sebelum penerapan SMOTE, nilai akurasi pada setiap fold menunjukkan fluktuasi dengan rentang sekitar 92% hingga 96%. Akurasi tertinggi diperoleh pada fold ke-9 sebesar 95%, sedangkan akurasi terendah terdapat pada fold ke-2 sebesar 92%. Setelah penerapan SMOTE, akurasi model cenderung lebih stabil dengan rentang sekitar 96,0% hingga 97%. Fold terbaik setelah penerapan SMOTE terdapat pada fold ke-10 dengan akurasi sekitar 97%, sedangkan nilai terendah diperoleh pada fold ke-9 sebesar 96%.

3.2.6 Evaluasi Performa Model

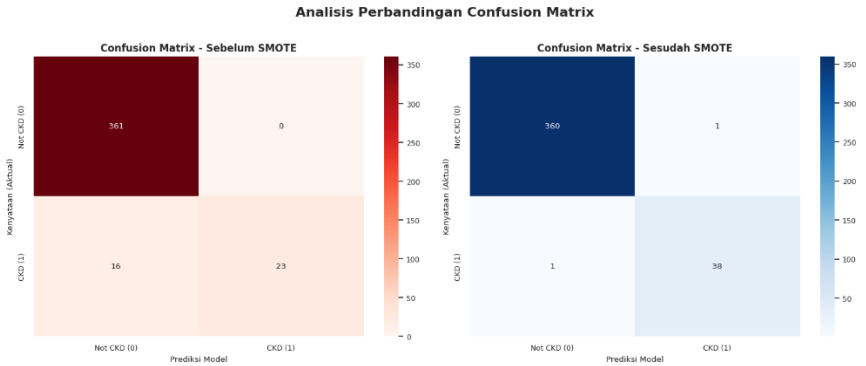
Evaluasi performa dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk membandingkan hasil klasifikasi CKD sebelum dan sesudah SMOTE. Perbandingan ini bertujuan untuk mengetahui pengaruh penerapan SMOTE terhadap kemampuan model dalam mengklasifikasikan kelas non-CKD (0) dan CKD (1), khususnya dalam meningkatkan sensitivitas model terhadap kelas minoritas. Hasil evaluasi menunjukkan adanya perbedaan performa sebelum dan sesudah penerapan SMOTE, sebagaimana disajikan pada Tabel 2.

Tabel 2. Performa Klasifikasi Sesudah & Sebelum *Random Forest*

Model	Not CKD (0)				CKD (1)		
	Acc	**Prec**	**Rec**	**F1-score**	**Prec**	**Rec**	**F1-score**
***Random Forest* Sebelum SMOTE**	96%	96%	100%	98%	100%	59%	74%
***Random Foreset* Sesudah SMOTE**	99%	100%	100%	100%	97%	97%	97%

3.2.7 Analisis *Confusion Matrix*

Hasil analisis menunjukkan adanya perbedaan performa algoritma *Random Forest* dalam klasifikasi penyakit CKD sebelum dan sesudah penerapan SMOTE. Perbandingan *confusion matrix* dilakukan untuk mengevaluasi kemampuan model dalam mengidentifikasi pasien CKD dan non-CKD, serta menganalisis perubahan jumlah kesalahan klasifikasi, khususnya false positive dan false negative, sebagaimana ditunjukkan pada Gambar 8.



(a) *Confusion Matrix* Sebelum SMOTE (b) *Confusion Matrix* Sesudah SMOTE

Gambar 7. Visualisasi Performa Model

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 71
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds

Berdasarkan perbandingan kedua *confusion matrix*, hasil perbandingan *confusion matrix* menunjukkan adanya peningkatan kemampuan algoritma *Random Forest* dalam mengklasifikasikan penyakit CKD setelah penerapan metode SMOTE. Sebelum penerapan SMOTE gambar (a), model berhasil mengklasifikasikan 361 pasien non-CKD dan 23 pasien CKD dengan benar, tetapi masih terdapat 16 kasus CKD yang salah diklasifikasikan sebagai non-CKD *(false negative)*. Setelah penerapan SMOTE gambar (b), jumlah kasus CKD yang berhasil terdeteksi meningkat menjadi 38 kasus, sedangkan *false negative* menurun menjadi 1 kasus. Nilai *recall* kelas CKD meningkat dari 59% menjadi 97%, sementara precision menurun sedikit dari 100% menjadi 97% akibat adanya satu kasus false positive. Hasil tersebut menunjukkan bahwa penerapan SMOTE meningkatkan sensitivitas model terhadap kelas minoritas dan mengurangi jumlah kasus CKD yang tidak terdeteksi. Peningkatan ini mengindikasikan bahwa SMOTE berpotensi membantu mengatasi ketidakseimbangan kelas dalam klasifikasi CKD, terutama dalam meminimalkan kesalahan false negative yang menjadi perhatian penting dalam deteksi dini penyakit CKD.

3.3 Pembahasan

Hasil penelitian menunjukkan bahwa penerapan *Synthetic Minority Over-sampling Technique* (SMOTE) meningkatkan kinerja algoritma *Random Forest* dalam klasifikasi CKD. Sebelum penerapan SMOTE, model memperoleh *accuracy* sebesar 96%, *precision* 100%, *recall* 59%, dan *F1-score* 74% pada kelas CKD. Setelah penerapan SMOTE, *accuracy* meningkat menjadi 99%, sedangkan *precision*, *recall*, dan *F1-score* masing-masing mencapai 97%. Peningkatan tersebut didukung oleh penurunan jumlah *false negative* dari 16 menjadi 1 kasus, yang menunjukkan peningkatan kemampuan model dalam mengidentifikasi pasien CKD.

Temuan ini sejalan dengan penelitian Bhirawa dan Sanjaya (2025) yang menerapkan SMOTE pada klasifikasi penyakit ginjal kronis menggunakan algoritma *Random Forest* dan *Gradient Boosting*. Penelitian tersebut melaporkan bahwa *Random Forest* mencapai *accuracy* 99,30%, sedangkan *Gradient Boosting* memperoleh 95,28% setelah penerapan SMOTE. Hasil penelitian ini dengan *accuracy* sekitar 99% menunjukkan kecenderungan yang serupa, yaitu penerapan SMOTE mampu meningkatkan kinerja klasifikasi CKD. Perbedaannya terletak pada jumlah dataset, yaitu 4.000 data pada penelitian ini dan 400 data pada penelitian Bhirawa dan Sanjaya (2025) [3].

Penelitian lain yang dilakukan oleh Riansyah et al. (2025) membandingkan algoritma *Decision Tree* dan *Random Forest* dalam diagnosis penyakit hepatitis menggunakan 155 data. Hasilnya menunjukkan bahwa *Random Forest* memperoleh *accuracy* 90,32%, *precision* 100%, dan *recall* 89,28%, sedangkan *Decision Tree* menghasilkan *accuracy* 80,65%, *precision* 88%, dan *recall* 88%. Hasil tersebut mendukung penggunaan *Random Forest* dalam klasifikasi penyakit, meskipun penelitian ini memperoleh *accuracy* lebih tinggi, yaitu 99% setelah penerapan SMOTE. Perbedaan kinerja tersebut perlu dipahami berdasarkan karakteristik dataset, jenis penyakit, dan pendekatan pengolahan data yang digunakan [4].

Penelitian selanjutnya Desmita et al. (2025) membandingkan *Random Forest* dan *Decision Tree* dalam klasifikasi diabetes melitus menggunakan 100.000 data dengan penerapan SMOTE. Random Forest memperoleh *accuracy* tertinggi sebesar 96,02% dan *F1-score* 76,13% pada pembagian data 60:40, sedangkan *Decision Tree* mencapai *accuracy* 94,80% dengan *F1-score* 70,93%. Hasil tersebut memperlihatkan bahwa *Random Forest* memberikan kinerja lebih baik dibandingkan *Decision Tree* pada dataset diabetes yang digunakan. Penelitian ini menunjukkan kecenderungan serupa dengan peningkatan *accuracy* menjadi 99% dan *F1-score* kelas CKD mencapai 97% setelah penerapan SMOTE. Perbedaan nilai evaluasi menunjukkan bahwa karakteristik data dan distribusi kelas perlu dipertimbangkan dalam membandingkan kinerja algoritma [9].

Penelitian lain yang dilakukan oleh Mahing et al. (2024) membandingkan algoritma *Support Vector Machine* (SVM) dan *Random Forest* dalam klasifikasi tingkat stres menggunakan 8.439 data teks dari Twitter. Hasil penelitian menunjukkan bahwa SVM dengan metode TF-IDF memperoleh *accuracy* 84%, *precision* 80%, *recall* 89%, dan *F1-score* 84%, sedangkan *Random Forest* dengan *CountVectorizer* menghasilkan *accuracy* sekitar 80%. Hasil tersebut berbeda dengan penelitian ini karena SVM menunjukkan kinerja lebih unggul pada klasifikasi teks, sementara penelitian ini berfokus pada optimalisasi *Random Forest* melalui penerapan SMOTE pada data medis. Perbedaan tersebut memperlihatkan bahwa efektivitas algoritma klasifikasi bergantung pada karakteristik data, metode pengolahan fitur, dan permasalahan yang diselesaikan [25].

Secara keseluruhan, perbandingan dengan penelitian terdahulu menunjukkan bahwa *Random Forest* memiliki potensi yang baik dalam klasifikasi penyakit, sedangkan penerapan SMOTE dapat membantu meningkatkan kemampuan model mengenali kelas minoritas. Kontribusi utama penelitian ini terletak pada peningkatan *recall* kelas CKD dari 59% menjadi 97% dan penurunan *false negative* dari 16 menjadi 1 kasus. Hasil *10-Fold Cross-Validation* juga menunjukkan *accuracy* setelah SMOTE yang relatif stabil pada kisaran 96%-97%. Temuan tersebut memperkuat potensi kombinasi *Random Forest* dan SMOTE dalam mendukung pengembangan sistem deteksi dini penyakit ginjal kronis, meskipun pengujian pada dataset lain masih diperlukan untuk memastikan kemampuan generalisasi model.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 72
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds

4. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan metode SMOTE mampu meningkatkan kinerja algoritma *Random Forest* dalam klasifikasi penyakit CKD, khususnya dalam mengatasi ketidakseimbangan kelas dan meningkatkan kemampuan deteksi pasien CKD. Berdasarkan hasil pengujian, model *Random Forest* sebelum penerapan SMOTE memperoleh *accuracy* sebesar 96%, *precision* 100%, *recall* 59%, dan *F1-score* 74% pada kelas CKD. Setelah penerapan SMOTE, nilai *accuracy* meningkat menjadi sekitar 99%, sedangkan *precision*, *recall*, dan *F1-score* kelas CKD masing-masing mencapai 97%. Peningkatan tersebut didukung oleh hasil *confusion matrix* yang menunjukkan penurunan jumlah *false negative* dari 16 menjadi 1 kasus serta peningkatan jumlah pasien CKD yang berhasil diidentifikasi dari 23 menjadi 38 kasus. Pengujian menggunakan *10-Fold Cross-Validation* juga memperlihatkan peningkatan dan kestabilan akurasi setelah penerapan SMOTE, dengan rentang sekitar 96% hingga 97%. Hasil tersebut menunjukkan bahwa kombinasi *Random Forest* dan SMOTE lebih efektif dibandingkan *Random Forest* tanpa SMOTE dalam mengenali kelas minoritas pada dataset penelitian ini. Kontribusi penelitian ini adalah memberikan bukti empiris mengenai efektivitas SMOTE dalam meningkatkan sensitivitas *Random Forest* terhadap kelas minoritas CKD dan meminimalkan kesalahan *false negative*, sebagai dasar pengembangan model pendukung deteksi dini penyakit ginjal kronis. Meskipun demikian, penelitian masih terbatas pada satu sumber dataset dan belum dilakukan pengujian menggunakan data klinis dari institusi yang berbeda. Penelitian selanjutnya disarankan menerapkan optimasi *hyperparameter*, membandingkan algoritma klasifikasi lainnya, serta melakukan validasi menggunakan dataset multi-institusi untuk meningkatkan kemampuan generalisasi dan keandalan model dalam mendukung pengembangan sistem deteksi dini CKD.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: Rahmat Dwi Cahyono, Herman Yuliansyah, dan Sunardi;
Metodologi: Rahmat Dwi Cahyono, Herman Yuliansyah, dan Sunardi;
Perangkat Lunak: Rahmat Dwi Cahyono;
Validasi: Rahmat Dwi Cahyono, Herman Yuliansyah, dan Sunardi;
Analisis Formal: Rahmat Dwi Cahyono dan Herman Yuliansyah;
Investigasi: Rahmat Dwi Cahyono;
Sumber Daya: Rahmat Dwi Cahyono, Herman Yuliansyah, dan Sunardi;
Kurasi Data: Rahmat Dwi Cahyono;
Penulisan Draf Awal: Rahmat Dwi Cahyono;
Penulisan, Peninjauan, dan Penyuntingan: Rahmat Dwi Cahyono, Herman Yuliansyah, dan Sunardi;
Visualisasi: Rahmat Dwi Cahyono;
Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan ekternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 73
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 64-75
DOI: 10.47065/bulletinds.v6i1.10441
https://ejurnal.seminar-id.com/index.php/bulletinds

UCAPAN TERIMAKASIH

Penelitian ini didukung oleh Program Magister Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan.

REFERENCES

[1] V. Miguel, I. W. Shaw, and R. Kramann, "Metabolism at the crossroads of inflammation and fibrosis in chronic kidney disease," *Nature Reviews Nephrology.*,vol. 21, no. 1, pp,39-56, 2025. doi: 10.1038/s41581-024-00889-z.

[2] W. Maryati, I. O. Rahayuningrum, and Hestiana, "Ketepatan Kode Diagnosis Chronic Kidney Disease dalam Mendukung Kelancaran Klaim BPJS di Rumah Sakit," *Indones. Heal. Inf. Manag. J.*, vol. 11, no. 1, pp. 43–50, 2023, doi: 10.47007/inohim.v11i1.497.

[3] R. M. A. A. Bhirawa and U. P. Sanjaya, "From Data Imbalance to Precision: SMOTE-Driven Machine Learning for Early Detection of Kidney Disease (Optimasi Klasifikasi Data Tidak Seimbang pada Dataset Medis pada Kasus Penyakit Ginjal Kronis dengan Teknik SMOTE)," *J. Inovtek Polbeng - Seri Inform.*, vol. 10, no. 1, pp. 1–12, 2025, doi:10.35314/7jgjmg64.

[4] M. Riansyah, S. Bahri, and H. P. Fiqna, "Perbandingan Akurasi Algoritma Decision Tree dengan Random Forest dalam Diagnosa Penyakit Hepatitis," *Indones. J. Sci. Technol. Humanit.*, vol. 2, no. 3, pp. 261–267, 2025, doi: 10.60076/ijstech.v2i3.1108.

[5] D. Mualfah *et al.*, " Teknik SMOTE untuk mengatasi imbalance data pada deteksi penyakit stroke menggunakan algoritma random forest " Jurnal Computer Science and Information Technology (CoSciTech) algoritma random forest, vol. 3, no. 2, pp. 107–113, 2022, doi: https://doi.org/10.37859/coscitech.v3i2.3912.

[6] A. Fatima, S. Fatima, and A. Kiran, "Big Data and Machine Learning Based Early Chronic Kidney Disease Prediction," *J. Sci. Res. Technol.*, vol. 2, no. 3, pp. 22–28, 2024, https://www.jsrtjournal.com/index.php/JSRT/article/view/88

[7] M. R. Nurul Istiqamah, "Klasifikasi Ulasan Konsumen Menggunakan Random,"Journal of system and computer engineering., vol. 5, no. 1, pp. 66–77, 2024, doi: 10.61628/jsce.v5i1.1061.

[8] K. N. Qodri, M. R. Fikri, and L. Ardi, "Prediksi Analitik untuk Penyakit Ginjal Kronis: Perbandingan Metode Machine Learning," *J. Keilmuan Teknol. Inf.*, vol. 1, no. 1, pp. 15–22, 2025, doi: 10.61902/jkti.v1i1.1686.

[9] N. L. Desmita, S. Lonang, and D. T. Kumoro, "Comparative Analysis of Decision Tree and Random Forest Algorithms for Predicting Diabetes Mellitus," *Sainstech Innov. J.*, vol. 8, no. 1, pp. 507–518, 2025, doi:10.37824/sij.v8i1.2025.783.

[10] R. N. I. A. Anan, U. T. Arshi, S. Karim, M. K. Hasan, M. M. Rahman, T. Taher, "A dataset of demographic and lifestyle risk factors for assessing chronic kidney disease development in diabetic patients," *Data Br.*, vol. 64, p. 112414, Feb. 2026, doi: 10.1016/j.dib.2025.112414.

[11] H. P. F. M. Riansyah, S. Bahri, "Perbandingan Akurasi Algoritma Decision Tree dan Random Forest dalam Diagnosa Penyakit Hepatitis," *Indones. J. Sci. Technol. Humanit.*, vol. 2, no. 3, pp. 261–267, Feb. 2025, doi: 10.60076/ijstech.v2i3.1108.

[12] Y. C. Ridwan1, Hanny Hikmayanti Handayani 2, Santi Arum Puspita Lestari 3, "Evaluasi Kinerja Algoritma Random Forest Dan Gradient Boosting Untuk Klasifikasi Penyakit Jantung," *J. Komtika (Komputasi dan Inform.*, vol. 9, no. 1, pp. 112–124, 2025, doi: 10.31603/komtika.v9i1.13450.

[13] M. U. K. Muhammad Arief, "Klasifikasi Penyakit Mata Menggunakan Metode Support Vector Machine dan Modified Balanced Random Forest," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 11, no. 1, pp. 1596–1607, Jan. 2026, doi: 10.36341/rabit.v11i1.6655.

[14] W. J. Veron and S. Muhammad, "Penerapan Metode Naive Bayes dalam Mengklasifikasi Tingkat Obesitas Pada Pria," *JOISIE J. Inf. Syst. Informatics Eng.*, vol. 6, no. 2, pp. 69–77, 2022, doi: 10.35145/joisie.v6i2.2467.

[15] E. Setiawan, B. Sartono, and K. A. Notodiputro, "SMOTE and Weighted Random Forest for Classification of Areas Based on Health Problems in Java," *J. Appl. Informatics Comput.*, vol. 9, no. 4, pp. 1587–1592, 2025, doi: 10.30871/jaic.v9i4.9933.

[16] B. N. U. H. H. R. Fikry, R. Kurniawan, "Sistem Klasifikasi Penyakit Diabetes Menggunakan Algoritma Random Forest Berdasarkan Data Medis," Journal of Computer Science and Artificial Intelligence (JCSAI)., vol. 3, no. 01, pp. 2–7, 2026, https://ruangjurnal.or.id/index.php/jcsai/article/view/293.

[17] I. Puspasari and W. I. Kusumawati, "Pemodelan Kesehatan Jantung Janin: Studi Komparatif Klasifikasi Menggunakan Metode Machine Learning," *J. Ilm. Scroll Jendela Teknol. Inf.*, vol. 13, no. 1, pp. 1–8, 2025, https://journal.univ45sby.ac.id/index.php/SCROLL/article/view/1.

[18] I. Liana Hulu, N. Mumtazia, S. Shapira Nurulita, A. W. Rakhmawati, and F. I. Komputer, "Penerapan Random Forest untuk Klasifikasi Penyakit Tanaman Padi," *JRIIN J. Ris. Inform. dan Inov.*, vol. 2, no. 8, pp. 1353–1363, 2025, https://jurnalmahasiswa.com/index.php/jriin/article/view/1795.

[19] N. N. Amalina and H. An, "A Multi-Head Attention Soft Random Forest for Interpretable Patient No-Show Prediction," *Systems*, vol. 14, no. 5, p. 576, May 2026, doi: 10.3390/systems14050576

[20] A. Latifah, A. Alrizal, and Y. Fendriani, "Klasifikasi Penyakit Kanker Payudara pada Citra Mammogram Menggunakan Algoritma Convolutional Neural Network (CNN) dan Random Forest," *J. ONLINE Phys.*, vol. 10, no. 3, pp. 95–103, Jul. 2025, doi: 10.22437/jop.v10i3.43794.

[21] A. Sunyoto and H. Al Fatta, 'Hidayat, A. Sunyoto, dan H. A. Fatta, 'Klasifikasi Penyakit Jantung Menggunakan Random Forest Classifier,' Jurnal Sistem Komputer dan Kecerdasan Buatan, vol. VII, no. 1, pp. 31–37, Sep. 2023.," vol. VII, no. September, pp. 31–40, 2023, doi:10.47970/siskom-kb.v7i1.464.

[22] M. Habib Aulia Raihan, Herman Yuliansyah, "Model Klasifikasi Emosi Berbasis Teks dengan Algoritma Decision Tree dan Support Vector Machine," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 7, no. 2, pp. 195–203, 2025.

[23] S. Sarmini, S. Sunardi, and A. Fadlil, "Performa Random Forest dan XGBoost pada Deteksi Penipuan E-Commerce Menggunakan Augmentasi Data CGAN," *Build. Informatics, Technol. Sci.*, vol. 6, no. 3, pp. 1919–1931, 2024, doi: 10.47065/bits.v6i3.6430.

[24] S. A. Hidayah and Z. Abidin, "Prediksi Penyakit Asma Menggunakan Naïve Bayes dan Random Forest Berbasis Data Klinis Multivariat," *J. Algoritma.*, vol. 23, no. 1 pp. 1756–1767, 2026, doi: 10.33364/algoritma/v.23-1.3521.

[25] N. Fathirachman Mahing, A. Lazuardi Gunawan, A. Foresta Azhar Zen, F. Abdurrachman Bachtiar, and S. Agung Wicaksono, "Klasifikasi Tingkat Stress dari Data Berbentuk Teks dengan Menggunakan Algoritma Support Vector Machine (SVM) dan Random Forest," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, pp. 1067–1076, 2024, doi: 10.25126/jtiik.2024118010.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 74
This work is licensed under a Creative Commons Attribution 4.0 International License.