

IMPLEMENTATION OF THE RANDOM FOREST ALGORITHM FOR CHRONIC KIDNEY DISEASE CLASSIFICATION

Stevani Tuhuleruw¹, Lea Sida Christmas A. Lesbassa², Emanuella M. C. Wattimena³, Citra Fathia Palembang^{4*}

^{1,2,3,4}Computer Science Study Program, Faculty of Science and Technology, Universitas Pattimura
Jl. Ir. M. Putuhena, Ambon City, 97233, Maluku Province, Indonesia

Corresponding author's e-mail: *citra.palembang@lecturer.unpatti.ac.id

ARTICLE INFO	ABSTRACT
Article History Received : February 7th, 2026 Reviewed : March 4th, 2026 Accepted : March 12th, 2026 Published : March 17th, 2026 Keywords Chronic Kidney Disease; Random Forest; Machine Learning; Classification; Clinical Data.	Chronic Kidney Disease (CKD) is a serious health problem characterized by a gradual decline in kidney function and often detected at an advanced stage due to the absence of early clinical symptoms. Early detection is therefore essential to reduce complications and improve patient outcomes. This study aims to implement the Random Forest algorithm for the classification of chronic kidney disease based on clinical data. The dataset used consists of 400 patient records with 26 clinical attributes, including blood pressure, creatinine, haemoglobin, sodium, potassium, albumin, and other medical indicators. Prior to model development, several preprocessing steps were performed, including handling missing values using median and mode imputation, encoding categorical variables, data normalization using Min-Max Scaling, and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE). The dataset was then divided into training data (80%) and testing data (20%). The Random Forest classifier was implemented using 100 decision trees with the Gini criterion. The evaluation results show that the proposed model achieved high performance with an accuracy of 0.975, precision of 0.97, recall of 0.98, and an F1-score of 0.975. The confusion matrix analysis indicates that the model can effectively classify CKD and non-CKD cases with minimal classification errors. In addition, feature importance analysis reveals that clinical parameters such as creatinine, haemoglobin, albumin, diabetes, and hypertension play significant roles in predicting chronic kidney disease. These findings demonstrate that the Random Forest algorithm has strong potential to be utilized as a decision-support tool for early detection of chronic kidney disease based on clinical data.



This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/). Editor of *Beta-Barekeng: J. Math. & Comp. Sci.*, Published by Beta Math Archipelago

¹How to cite this article:

S. Tuhuleruw, L. S. C. A. Lesbassa, E. M. C. Wattimena, and C. F. Palembang, "IMPLEMENTATION OF THE RANDOM FOREST ALGORITHM FOR CHRONIC KIDNEY DISEASE CLASSIFICATION", *Beta-Barekeng: J. Math. & Comp. Sci.*, vol. 1, iss. 1, pp. 29-34, April, 2026.

© 2026 by the Author(s)

e-mail: betabarekeng@gmail.com
Homepage <http://ojs.beta-math.id/index.php/beta-barekeng>
DOI: <https://doi.org/10.33333/beta-barekengv1.i1.p29-34>

1. Introduction

Chronic Kidney Disease (CKD) is one of the serious global health problems due to its progressive nature and the fact that it cannot be completely cured. This condition often requires long-term renal replacement therapy, such as dialysis or kidney transplantation, particularly in the advanced stages of the disease [1]. The kidneys play a vital role in maintaining the balance of fluids and electrolytes in the body, filtering metabolic waste, and regulating blood pressure and red blood cell production. Chronic and progressive kidney damage can be triggered by various risk factors, including diabetes mellitus, hypertension, cardiovascular diseases, and unhealthy lifestyle habits such as poor dietary patterns and lack of physical activity [2]. Therefore, early detection of CKD is very important to prevent disease progression that may lead to complete kidney failure and death.

With advances in data science and artificial intelligence, medical data analysis using machine learning algorithms has become a promising alternative to assist healthcare professionals in performing early diagnosis [3],[4]. One of the algorithms widely used for tabular data classification is Random Forest due to its ability to handle both numerical and categorical variables, as well as its capability to minimize the risk of overfitting [5]. Random Forest is one of the ensemble learning methods used for classification and regression tasks and is widely applied in machine learning due to its robustness against overfitting and its ability to handle data with various types of variables [6].

Previous research by Wibowo applied the Random Forest algorithm to classify chronic kidney disease, the results showed an accuracy of 98.5%, which was higher compared to the Decision Tree and SVM methods [7]. Another study using the Random Forest algorithm was able to model the prediction of Chronic Kidney Disease (CKD) with very high performance on the dataset and testing scheme used. The evaluation results showed that the accuracy, precision, recall, F1-score, and AUC values reached their maximum levels, indicating the model's capability to recognize clinical patterns that distinguish CKD and non-CKD patients after undergoing structured and comprehensive preprocessing stages [8]. Meanwhile, Windi and Majid (2025), in their study comparing the performance of three machine learning algorithms—Random Forest, Support Vector Machine (SVM), and Naïve Bayes—in predicting the risk of cardiovascular disease using the Mendeley Cardiovascular Disease Dataset, reported that Random Forest achieved the best performance with an accuracy of 99%, recall of 100%, precision of 98%, and an F1-score of 99%, followed by SVM and Naive Bayes [9]. Fandi and Viry in 2021, also stated that the Random Forest method is superior to the Decision Tree method. The Decision Tree method still produced some incorrect predictions with an accuracy rate of 84.4%, while the Random Forest method provided more accurate predictions with an accuracy rate of 85.5% [10].

Based on the discussion of previous studies, this research applies the Random Forest algorithm to classify chronic kidney disease using a medical dataset obtained from an open-source repository (Kaggle). The dataset contains medical records of 400 patients with 26 clinical attributes and one target attribute that classifies patients into two categories: “CKD” (chronic kidney disease) and “non-CKD” (non-chronic kidney disease). The attributes include several important clinical indicators related to kidney function, such as blood pressure, creatinine level, hemoglobin level, sodium level, potassium level, albumin level, the presence of pus cells in urine, and comorbid conditions such as diabetes and hypertension. These clinical features provide comprehensive information regarding the health condition of the patients. By utilizing these attributes, the Random Forest model is expected to effectively learn patterns from the data and produce accurate classifications in detecting the likelihood of chronic kidney disease.

2. Research Methods

This study was conducted in several systematically designed stages to develop a machine-learning-based classification model for chronic kidney disease. These stages include data collection, data preprocessing, Random Forest model development, model training and validation, and model performance evaluation using several evaluation metrics.

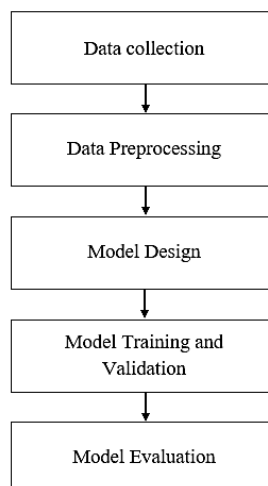


Figure 1. Research Methods

2.1. Data Collection

In the initial stages of the research, a search for datasets relevant to the research topic was conducted. The dataset used in this study is a public dataset obtained from Kaggle and used in the classification process.

2.2. Data Preprocessing

Before the model is trained, the data undergoes several preprocessing stages:

1. Handling Missing Values

Handling missing values through the imputation process aims to complete incomplete data so that all attributes can be utilized during the model training process [11].

2. Encoding of Categorical Variables

In this research, the encoding process was performed using label encoding and one-hot encoding techniques according to the characteristics of each variable. Label encoding was applied to categorical attributes with two categories (binary), such as the presence of diseases like hypertension or diabetes, where the category values were converted into numerical values such as 0 and 1. Meanwhile, one-hot encoding was used for attributes with more than two categories, where each category was represented as a new binary variable. This process aims to ensure that all attributes in the dataset are represented in numerical form so that they can be processed by machine learning algorithms during the model training stage [12].

3. Normalization of Numeric Attributes

Numerical attributes such as blood pressure, creatinine, haemoglobin, sodium, potassium, and albumin were normalized to the range of [0,1] using the Min–Max Scaling method to improve training efficiency and ensure balanced contribution of each variable during model training. The Min–Max normalization formula is presented in Eq. (1), where x_i represents the original data value, x' denotes the normalized value, $\min(x)$ is the minimum value of a feature, and $\max(x)$ is the maximum value of a feature.

$$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

4. Class Imbalance

The class distribution in the initial dataset indicates a class imbalance, where one class is underrepresented. Therefore, oversampling using the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to reduce model bias and ensure that the model can learn patterns from both classes proportionally [13].

After the preprocessing stage, the dataset was randomly split into training data (80%) and testing data (20%) to allow the model to be evaluated on unseen data.

2.3. Model Design

The model used in this study is the Random Forest Classifier, implemented using the scikit-learn library in the Python programming language. The main parameters configured in the model include:

1. Number of decision trees (n_estimators): 100
2. Splitting criterion (criterion): Gini
3. Maximum tree depth (max_depth): Automatically optimized (None)
4. Random state: 42 (to ensure reproducibility)

2.4. Model Training and Validation

Model training was conducted in two stages:

1. Cross-Validation (5-Fold)

The training data were divided into five folds. In each iteration, four folds were used to train the model, while the remaining fold was used as validation data. This process was repeated five times so that each fold served as validation data once. The average performance across the five iterations was then calculated to provide an initial assessment of the model's performance on the training dataset. The model performance during cross-validation was evaluated using several classification metrics derived from the confusion matrix, namely accuracy, precision, recall, and F1-score. These metrics are commonly used to measure the effectiveness of classification models.

Accuracy measures the overall correctness of the model in classifying instances and is calculated using the following formula:

$$\text{Accuracy}\% = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (2)$$

Precision indicates the proportion of correctly predicted positive instances among all instances predicted as positive:

$$Precision\% = \frac{TP}{TP+FP} \times 100\% \quad (3)$$

Recall (also known as sensitivity) measures the ability of the model to correctly identify all actual positive instances:

$$Recall\% = \frac{TP}{TP+FN} \times 100\% \quad (4)$$

F1-score represents the harmonic mean of precision and recall and provides a balanced measure between the two metrics:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

2. Final Training

After the parameters were determined to be optimal based on the cross-validation results, the model was retrained (fit) using the entire training dataset (100%).

2.5. Model Evaluation

After the model training process was completed, the next stage was model evaluation to measure how well the model performs in classification tasks. The evaluation was conducted using testing data that had not been used during the model training process. This was intended to assess the model's ability to generalize to new, unseen data.

In this study, the performance of the Random Forest Classifier was evaluated using several commonly used classification metrics, namely accuracy, precision, recall, and F1-score. In addition, the evaluation was also presented using a confusion matrix to analyze the distribution of the model's predictions compared to the actual class labels [14].

3. Results And Discussion

3.1. Data Processing

Before the model was trained, the data first underwent several preprocessing steps, including:

1. Handling missing values using simple imputation methods such as median and mode.
2. Converting categorical variables into numerical form using one-hot encoding or label encoding techniques.
3. Normalizing numerical attributes using the Min-Max Scaling approach to accelerate the model training process.

The class distribution in the dataset shows that the CKD class dominates the non-CKD class. Therefore, to avoid imbalance that may cause bias in the model, data balancing was performed using oversampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique)

3.2. Data Splitting

This study uses the Random Forest Classifier algorithm as the main model, which is known to be effective in handling tabular data consisting of various numerical and categorical variables. Several key parameters applied to this model include:

1. Number of trees (estimators): 100
2. Splitting criterion: Gini
3. Maximum tree depth: Unlimited (automatically optimized)
4. Random state: 42 (to ensure reproducibility)

The model was trained using 80% of the total dataset as training data, while the remaining 20% was used as testing data.

3.3. Training dan Evaluasi Model

The model was trained using a 5-fold cross-validation technique on the training dataset. The evaluation results on the testing data show that the model achieved excellent performance, with the following metrics:

Table 1. Evaluation Metrics of the Random Forest Model

Evaluation Metrics	Value
Accuracy	0,975
Precision	0,97
Recall	0,98
F1-Score	0,975

Based on **Table 1**, The high accuracy and F1-score values indicate that the model demonstrates excellent performance in distinguishing between patients with chronic kidney disease and those without it. The balance between precision and recall also shows that the model is not only accurate in identifying positive cases but is also reliable in detecting the presence of the disease overall.

3.4. Evaluasi dengan Confusion Matrix

To evaluate the performance of the developed classification model, a Confusion Matrix was used as a visualization tool as well as a measure of prediction accuracy. The Confusion Matrix provides detailed information regarding the number of correct and incorrect predictions for each analyzed class [15].

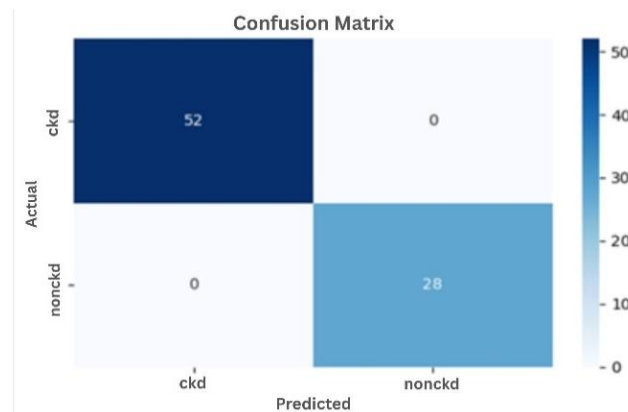


Figure 2. Confusion Matrix

In this study, the model predicts two class categories: CKD, representing patients diagnosed with chronic kidney disease, and non-CKD, representing patients without chronic kidney disease.

The classification results for predicting chronic kidney disease (CKD) and non-CKD patients are summarized using a Confusion Matrix, as presented in **Table 2**.

Table 2. Confusion Matrix of the Random Forest Classification Model

	Predicted CKD	Predicted non-CKD
Actual CKD	52	0
Actual non-CKD	0	28

Based on the evaluation results, the following explanations can be provided:

True Positive (TP) = 52 → The model correctly classified 52 patients who truly suffer from the disease.

True Negative (TN) = 28 → The model also successfully identified 28 healthy patients correctly.

False Positive (FP) = 0 → No cases were found where healthy patients were incorrectly classified as having the disease.

False Negative (FN) = 0 → There were no patients who actually had the disease but were predicted as healthy by the model.

Since no misclassification occurred, the model achieved perfect accuracy (100%) on the testing data. This indicates that the Random Forest algorithm is highly effective in distinguishing between patients with chronic kidney disease and those without it, based on the analyzed clinical attributes. The use of a confusion matrix is very helpful in providing a clear overview of the model's performance. It not only shows the overall accuracy but also reveals the types of errors that may occur, such as false positives or false negatives, which can have significant implications in a medical context.

4. Conclusions

This study successfully implemented the Random Forest algorithm to classify chronic kidney disease based on patients' clinical data. Through several stages, including data preprocessing, class balancing using the SMOTE technique, and model training and evaluation, the model achieved excellent performance with an accuracy of 0.975, precision of 0.97, recall of 0.98, and an F1-score of 0.975. The evaluation results using the confusion matrix indicate that the model was able to classify all test data with very high accuracy without misclassification. Furthermore, the feature importance analysis revealed that several clinical parameters, such as creatinine, hemoglobin, albumin, and comorbid conditions including diabetes and hypertension, have significant contributions in detecting chronic kidney disease. These findings demonstrate that the Random Forest algorithm has great potential to be utilized as a decision-support tool for the early

detection of chronic kidney disease based on clinical data, thereby supporting faster and more accurate decision-making in the healthcare sector. For future research, it is recommended to test the model using larger and more diverse datasets to improve the model's generalization capability. In addition, future studies may compare the performance of Random Forest with other machine learning algorithms such as Support Vector Machine, Gradient Boosting, or deep learning methods to obtain a more optimal prediction model. The development of the system in the form of an application or a more interactive dashboard can also be considered so that the prediction results can be directly utilized by medical professionals as a supporting tool in the early diagnosis of chronic kidney disease.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgement

We sincerely thank all reviewers for their time, expertise, and constructive comments, which have greatly contributed to improving the quality of the manuscripts and maintaining the academic standards of the journal.

Conflict of Interest

The authors declare that he/she has no conflicts of interest to report study.

Declaration of Generative AI and AI-assisted technologies

Generative AI tools were used solely for language refinement (grammar, spelling, and clarity). The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. The authors reviewed and approved all final text.

References

- [1] A. C. Narsa, V. Maulidya, D. Reggina, W. Andriani, and H. R. Rijai, "Case Study: Chronic Kidney Disease (Stage V) Patient with Pulmonary Edema and Fluid Electrolyte Imbalance," *JSK J. Sains dan Kesehat.*, vol. 4, no. 1, pp. 17–22, 2022.
- [2] S. N. Rudang, R. E. Rambe, and S. Annisa, "Indonesian Journal of Pharmaceutical and Identifikasi Drug Related Problems Pada Pasien Penyakit Ginjal Kronis Di Rumah Sakit Prof. dr. Chairuddin Panusunan Lubis Tahun 2023," *Indones. J. Pharm. Clin. Res.*, vol. 07, no. 01, pp. 34–45, 2024.
- [3] K. Magda, O. Y. Inonu, and E. R. Susanto, "Komparasi Metode Machine Learning Untuk Diagnosis Penyakit Kanker Payudara," *Expert J. Manaj. Sist. Inf. dan Teknol.*, vol. 15, no. 1, pp. 15–21, 2025.
- [4] R. G. Wardhana, G. Wang, and F. Sibuea, "PENERAPAN MACHINE LEARNING DALAM PREDIKSI TINGKAT KASUS PENYAKIT DI INDONESIA Master of Information Systems Management Bina Nusantara University Abstraksi Keywords : Pendahuluan Tinjauan Pustaka Metode Penelitian Proses implementasi machine learning dapat," *J. Inf. Syst. Manag.*, vol. 5, no. 1, pp. 40–45, 2023.
- [5] N. A. P. Indaryono, R. R. Saedudin, and F. Hamami, "Analisa perbandingan algoritma random forest dan naïve bayes untuk klasifikasi curah hujan berdasarkan iklim di indonesia," *JIPi (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 1, pp. 158–167, 2024.
- [6] H. Azis and N. Rismayanti, "PREDIKSI ANEMIA DARI PIXEL GAMBAR DAN LEVEL HEMOGLOBIN MENGGUNAKAN RANDOM FOREST PREDICTING ANAEMIA FROM IMAGE PIXELS AND," *J. Ilm. NERO*, vol. 9, no. 1, pp. 21–34, 2024.
- [7] H. S. Wibowo, "Penerapan Algoritma Random Forest untuk Klasifikasi Penyakit Ginjal Kronik," *J. Teknol. dan Sist. Komput.*, vol. 9, no. 2, pp. 123–130, 2021.
- [8] T. Prasetyo, "Prediksi penyakit ginjal kronis menggunakan algoritma random forest Chronic kidney disease prediction using the random forest algorithm," *INFOTECH J. Inform. Teknol.*, vol. 7, no. 1, pp. 1–8, 2026.
- [9] W. A. Rayadhani and M. Rahardi, "Comparative Analysis of Random Forest , SVM , and Naive Bayes for Cardiovascular Disease Prediction," *J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 3234–3243, 2025.
- [10] F. Y. Pamuji and V. P. Ramadhan, "Jurnal Teknologi dan Manajemen Informatika Komparasi Algoritma Random Forest Dan Decision Tree Untuk Memprediksi Keberhasilan Immunotherapy," *J. Teknol. dan Manaj. Inform.*, vol. 7, no. 1, pp. 46–50, 2021.
- [11] A. Sya'adah, A. A. Fadhilah, and I. Daulay, "PREDIKSI PENYAKIT JANTUNG DENGAN RANDOM FOREST MELALUI TEKNIK IMPUTASI DAN FEATURE SCALING PADA," *J. SIKOM (Sistem Inf. Komputer)*, vol. 1, no. 1, pp. 73–92, 2024.
- [12] A. Syaputra, "Implementation of Machine Learning Algorithms for Predicting Student Final GPA Using Multiclass Classification Models," *J. Artif. Intell. Softw. Eng.*, vol. 5, no. 2, pp. 660–675, 2025, doi: 10.30811/jaise.v5i2.7015.
- [13] Z. Fatah and I. Addewiyah, "Penerapan Algoritma Random Forest dan Teknik SMOTE untuk Prediksi Kematian Akibat Gagal Jantung Menggunakan RapidMiner," *JAMASTIKA*, vol. 4, no. 2, pp. 82–89, 2025.
- [14] O. Y. Inonu, K. Magda, and Amarudin, "Analisis Kinerja Algoritma Random Forest Dengan Model Machine Learning Pada Dataset Penyakit Diabetes," *Expert J. Manaj. Sist. Inf. dan Teknol.*, vol. 15, no. 1, pp. 1–7, 2025.
- [15] P. L. Romadloni, B. A. Kusuma, and W. M. Baihaqi, "Komparasi metode pembelajaran mesin untuk implementasi pengambilan keputusan dalam menentukan promosi jabatan karyawan," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 622–628, 2022.