Performance Comparison of SMOTE-Based Random Forest and Decision Tree Models for Chronic Kidney Disease Classification

Manuscript received August 4, 2026; revised September 23, 2026

Rahmat Dwi Cahyono
Department of Informatics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
2507048003@webmail.uad.ac.id

Herman Yuliansyah*
Department of Informatics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
herman.yuliansyah@uad.ac.id

Sunardi
Department of Informatics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
sunardi@uad.ac.id

Abstract*— *Chronic Kidney Disease (CKD) is a chronic condition characterized by a gradual decline in kidney function and is often detected at an advanced stage. Previous studies have employed machine learning methods-such as Random Forest and Decision Trees-for CKD classification; however, there is no consensus regarding which algorithm is most optimal for CKD datasets. The large number of clinical variables in CKD data increases computational complexity, necessitating high-performance machine learning methods to improve classification accuracy and support early disease detection. This study aims to develop a CKD classification model and compare the performance of Decision Tree and Random Forest algorithms. The research utilized a health dataset obtained from Mendeley Data. The methodology involved data acquisition, exploratory data analysis, data preprocessing (including addressing class imbalance via the Synthetic Minority Over-sampling Technique, or SMOTE), and model development using Random Forest and Decision Trees, with evaluation performed via 10-fold stratified cross-validation. Model performance was assessed using confusion matrices and classification reports, covering metrics such as accuracy, precision, recall, and F1-score. Experimental results indicate that Random Forest yielded the best classification performance, achieving an accuracy of 0.9437. It is concluded that comparing Random Forest and Decision Tree methods-combined with SMOTE-is an effective approach for identifying the optimal algorithm to enhance accuracy and support decision-making systems for the early detection of CKD.

Keywords—Chronic Kidney Disease, Decision Tree, Machine Learning, Random Forest, SMOTE.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a chronic disease characterized by a progressive decline in kidney function and lasts more than three months. CKD is a major global health problem because its prevalence continues to increase and is associated with substantial morbidity and mortality. It is estimated that about 10% of the world's population suffers from CKD, and the disease is projected to be one of the world's leading causes of death [1]. In Indonesia, CKD also has a huge economic impact. Based on BPJS Kesehatan data, kidney disease occupies the second position as the disease with the highest health financing after heart disease [2]. The high cost is mainly due to the need for long-term therapy, such as hemodialysis and kidney transplantation. This condition shows the importance of early detection to slow the progression of the disease and improve the quality of life of patients.

Recent developments in machine learning have made significant contributions to healthcare, particularly in disease diagnosis and clinical decision-making. Various classification algorithms can process complex medical data to identify disease patterns more efficiently than conventional approaches [3]. For CKD, machine learning can support early disease identification using patients' clinical, demographic, and lifestyle data, thereby assisting healthcare professionals in making more informed decisions [4].

Among various classification algorithms, Decision Tree and Random Forest are widely used because they offer complementary characteristics. Decision Tree provides a simple and interpretable model structure, making it suitable for applications that require transparent decision-making. In contrast, Random Forest is an ensemble learning method that combines multiple Decision Trees to improve predictive performance and reduce the risk of overfitting. It is also generally robust when handling complex and noisy datasets [5],[6]. Previous studies have reported that Random Forest can achieve higher classification performance than Decision Tree, particularly when working with complex or imbalanced medical datasets [7],[8]. However, these findings are not necessarily consistent across different CKD datasets because dataset characteristics, class distributions, and selected features may vary.

In addition, CKD classification studies have reported variations in performance because of differences in dataset characteristics, preprocessing techniques, and evaluation methods. In imbalanced datasets, classification models tend to favor the majority class, which may reduce their ability to correctly identify CKD-positive patients. Therefore, appropriate class-balancing techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), are needed to improve the representation of the minority class and support more reliable model evaluation [9],[10].

Based on these considerations, this study aims to develop a CKD classification model by comparing the performance of Decision Tree and Random Forest algorithms using a dataset containing clinical, demographic, and lifestyle characteristics. SMOTE was applied to address class imbalance, and model performance was evaluated using stratified 10-fold cross-validation and accuracy, precision, recall, and F1-score metrics. The study provides a comparative evaluation of the two algorithms and examines their potential for supporting CKD classification and future decision-support applications.

 EISSN: 2961-872X●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002553

II. METHODS

This study employed a quantitative approach using a CKD dataset obtained from Mendeley Data and published by Anan et al. [11]. The dataset was collected from BIRDEM General Hospital in Dhaka and contains demographic and lifestyle factors associated with CKD development in diabetic patients. The research procedure consisted of data acquisition, exploratory data analysis, data splitting, class balancing using SMOTE, model development using Random Forest and Decision Tree, cross-validation, and model evaluation. Model performance was assessed using confusion matrices and classification metrics, including accuracy, precision, recall, and F1-score. Furthermore, evaluate the comparative performance of the Algorithm. The following is a diagram of the stages that will be carried out, as shown in Fig. 1.

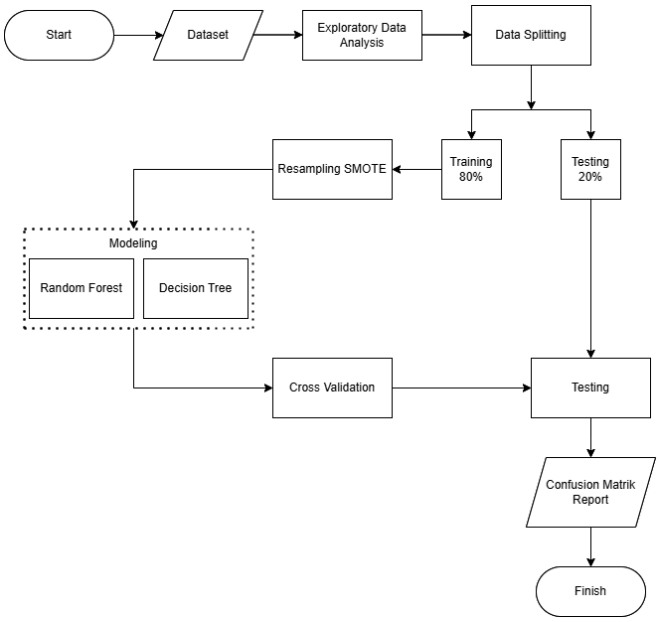


Fig. 1.Research Stages

A. Dataset

The next stage is data collection using a dataset (A Dataset of Demographic and Lifestyle Risk Factors for Assessing Chronic Kidney Disease Development in Diabetic Patients) which is publicly available at https://data.mendeley.com/datasets/hjkzgbxgv5/2. This dataset, collected from BIRDEM General Hospital, Dhaka, presents a comprehensive compilation of demographic and lifestyle factors to predict the progression of CKD Disease in diabetic patients.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the structure of the dataset, examine the distribution of the variables, and identify potential class imbalance in the target variable. The dataset contained 4,000 observations. The target variable was CKD status, with label 0 representing patients without CKD and label 1 representing patients with CKD [12]. EDA was used to examine the characteristics and relationships among the variables before model development [13],[14].

C. Splitting Data

After the data is ready to be used, the dataset is divided into training data and test data. This division aims to test the model's generalization ability against new data that has never been seen before. The dataset was divided into training and test sets using an 80:20 ratio. Thus, 3,200 observations were allocated to the training set and 800 observations to the test set. This separation allowed the model to be trained on one subset and evaluated on previously unseen data [15].

D. Resampling Using SMOTE

Handling Class Imbalance using the Synthetic Minority Over-Sampling Technique (SMOTE) to address imbalances between target classes, so that the model can better learn patterns. By doing this stage, the data is ready to be used for the next stage in analysis and modeling [16],[17]. The dataset exhibited class imbalance, which may cause classification models to favor the majority class and reduce their ability to identify CKD-positive cases. Therefore, the Synthetic Minority Over-sampling Technique (SMOTE) was selected to address the class imbalance by generating synthetic samples of the minority class based on existing minority-class observations. SMOTE was applied only to the training data after the data-splitting stage, while the test data were kept unchanged for unbiased performance evaluation. The resampled training data were subsequently used to train the Random Forest and Decision Tree models, while the original test data were retained for model evaluation. During the 10-fold cross-validation procedure, SMOTE was applied only to the training portion of each fold, while the corresponding validation fold was kept unchanged. This procedure prevented information from the validation data from being introduced during resampling and model training.

E. Modeling

After going through the SMOTE Resampling stage, the next step is to build a prediction model using the Decision Tree algorithm, and Random Forest. Both models were trained using the training data that had been prepared for each scenario, and tested using test data to measure the performance of the classification.

1) Random Forest

Random Forest is an ensemble classification method that combines multiple Decision Trees. During model construction, each tree is trained using a random subset of observations and features. The final classification is determined by aggregating the predictions of the individual trees, typically through majority voting [18]. Random Forest commonly uses bagging and random feature selection to improve model robustness. In this study, the Decision Trees were constructed using the CART (Classification and Regression Tree) approach [19]. The ensemble structure can reduce the tendency of individual Decision Trees to overfit the training data [20]. Fig. 2 shows an illustration of the algorithm's working process Random Forest.

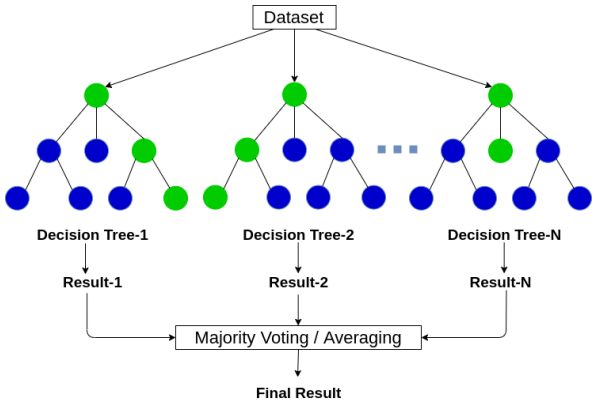


Fig. 2. Schematic representation of the Random Forest classification process

2) Decision Tree

Decision Tree is a tree-based classification method that recursively partitions data according to selected features [21],[22]. The process begins at a root node and proceeds through a series of decision nodes until reaching leaf nodes that represent the predicted classes. Because Decision Tree determines splits based on feature thresholds, feature normalization is generally not required. In this study, the Decision Tree model used impurity criteria such as Gini impurity and entropy to determine the quality of candidate splits [23],[24]. Gini Impurity is defined by Equation (1).

$$I\,gini(x) = \sum x \;\; p(x|y)(1 - p(x\backslash y) \quad (1)$$

where the total is always be extended to all existing classes. This is common to be used as a measure and is also used as a default value in Scikit-Learn. If a sample is given, Gini Impurity measures the likelihood of misclassification of a label if a label is randomly selected with the probability of branch distribution. Cross-entropy of Impurity is defined by Equation (2).

$$Icross - entropy\,(x) = -\sum x\, p(x|y)\, log\, p(x|y) \quad (2)$$

This measure is based on information theory and assumes that there is no value only when a sample belonging to an individual class is present at the separation but will be of maximum value when there is a uniform distribution between classes. Similar to Gini Impurity, it has a difference, namely cross-entropy allows the user to choose a separation that minimizes uncertainty in the classification, while Gini impurity directly minimize the possibility of errors in classification. Fig. 3 shows an illustration of the working process of the DT algorithm.

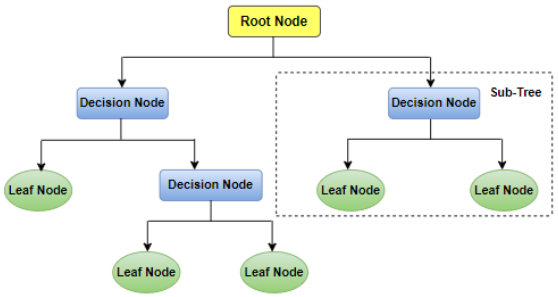


Fig. 3. Decision Tree

F. Cross-Validation

Cross-validation was used to evaluate the performance and generalization capability of the machine learning models. In this study, 10-fold stratified cross-validation was applied. The dataset was divided into 10 folds while maintaining a similar class distribution across the folds. In each iteration, nine folds (90% of the data) were used for training and one-fold (10% of the data) was used for validation. The procedure was repeated for 10 iterations, with each fold serving as the validation fold exactly once. The performance obtained across the 10 iterations was then used to assess the classification performance of the Random Forest and Decision Tree models. This technique provides more accurate estimation of model performance on invisible data [25]. Other important benefits of cross-validation are to assist data analysts in overcoming problems overfitting or conditions when the model is too specific on training data so that it is not good at analyzing new data. Scheme fold cross-validation used to evaluate model performance [20].

G. Evaluation

The model evaluation in this study was carried out to measure the performance of machine learning algorithms in classifying CKD disease. Some of the evaluation metrics used in this study are Accuracy, Precision, Recall, and F1 Score [26]. The four metrics were chosen because they can provide a comprehensive picture of the model's performance, especially in the context of medical predictions that require a high level of accuracy, especially in recognizing positive cases. The evaluation process is carried out on test data that has been previously separated from the training data, so that the results obtained can reflect the model's ability to deal with new data objectively [27],[28]. Accuracy measures the correct proportion of a prediction to the total amount of data with a formula as in Equation (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

where TP (True Positive) is a CKD patient who is successfully detected as a CKD patient, TN (True Negative) is a non-CKD patient who is successfully detected as a non-CKD patient, FP (False Positive) is a non-CKD patient who is falsely detected as a CKD patient and FN (False Negative) is a CKD patient who is wrongly detected as not a CKD patient. Accuracy provides an overview of the model's performance, but on an unbalanced dataset, this metric is often misleading because the model can achieve high accuracy just by predicting the majority class. Precision measures the accuracy of positive predictions with formulas such as in Equation (4).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Precision is important to minimize FP especially in the context of medical diagnosis. The high precision showed that most of the predicted patients did indeed suffer from CKD disease. Recall measures the model's ability to detect all cases that are actually positive (fraudulent) with a formula like in Equation (5).

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (5)$$

Recall is essential in the context of early detection of CKD to ensure that no CKD patient is missed and undiagnosed (False Negative). High recall showed that the model successfully detected most of the patients who actually had actual CKD. F1-score is the harmonic average of precision and recall with a formula such as Equation (6).

EISSN: 2961-872X●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002553

$$F1\text{-}Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

F1-score is used to balance precision and recall, especially when the two have very different values. A high F1-score value indicates an effective model in classifying CKD disease without producing too much False Positives or False Negatives.

III. Results and Discussion

This section presents a machine learning-based classification model that is capable of accurately detecting CKD based on the patient's clinical data. Through the dataset (A Dataset of Demographic and Lifestyle Risk Factors for Assessing Chronic Kidney Disease Development in Diabetic Patients) which is publicly available at Mendeley Data. Here, the study compares several popular algorithms, namely Radom Forest and Decision Tree.

A. Experimental Data

The dataset used amounted to 4,000 data with 26 attributes of CKD disease data. CKD disease patient data which included demographic features, anthropometry, history of diabetes, lifestyle, comorbidities, and CKD (1 = yes, 0 = no) as target labels. The data presents a comprehensive compilation of demographic and lifestyle factors to predict the development of CKD in diabetic patients. Details of the attributes used in this study are presented in Table 1.

Table I. Research Variables

No	Variable Name	Variable Type	Remarks
1	Gender	Categorical	Patient's gender (M = Male, F = Female)
2	Job	Numerical	Job type (1 = normal, 2 = intermediate, 3 = heavy)
3	Family Background of Diabetes	Binary	Family history of diabetes (1 = yes, 0 = no)
4	Diabetic Year	Numerical	Diabetes for a long time (years)
5	Age	Numerical	Patient's age (years)
6	BMI	Numerical	Patient's Body Mass Index
7	Follow Suggested Diet	Binary	Diet adherence (1 = yes, 0 = no)
8	Take Medicine for Diabetes	Binary	Consumption of diabetes medication (1 = yes, 0 = no)
9	Hypertension	Binary	History of hypertension (1 = yes, 0 = no)
10	CKD	Binary	Chronic Kidney Disease Status (1 = yes, 0 = no)

B. Experimental Results

1) Exploratory Data Analysis (EDA)

An analysis of the correlation between features and target variables is shown in Fig. 4. The Exploratory Data Analysis (EDA) stage in this study was conducted on 4,000 patient data to understand the structure, distribution characteristics, and correlations between clinical variables related to CKD disease. Through the visualization of the correlation matrix, it is known that factors such as age, gender, BMI, and family history of diabetes, have a positive correlation with the diagnosis of CKD, which provides an early clue regarding the most influential variables. In addition, distribution analysis on height and age variables showed a pattern close to the normal distribution, where the dataset was dominated by the middle to advanced age group who medically had a higher risk of decreased kidney function. The entire EDA process serves to identify the imbalance of the target class and serves as a basis for researchers to determine the right pre-processing techniques before entering the algorithm modeling stage.

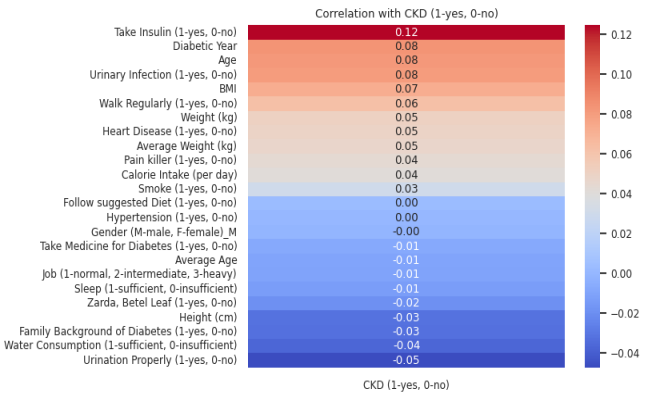


Fig. 4. Feature Correlation to DIAGNOSIS_CKD

2) Height Distribution

In the height distribution section, the data showed a pattern that was close to normal with the highest frequency being in the mid-value range.

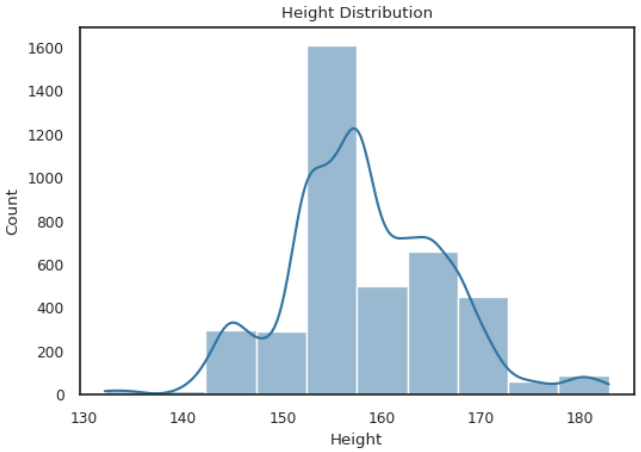


Fig. 5. Height Distribution

The visualization in Fig. 5 shows that the majority of individuals in the dataset have heights in the range of 150 cm to 160 cm, where the peak frequency is around 155 cm to 157 cm with a total of more than 1,600 data. This distribution of data shows that there is a concentration of data mass at a certain value that then slopes to the left (lower value) and right (higher value), which provides an overview of the physical characteristics of the 4,000 patient samples used in this CKD classification study.

3) Age Distribution

The age distribution section in this study shows that the data of 4,000 patients has a pattern that is close to the normal distribution with a dominance in the middle to advanced age group.

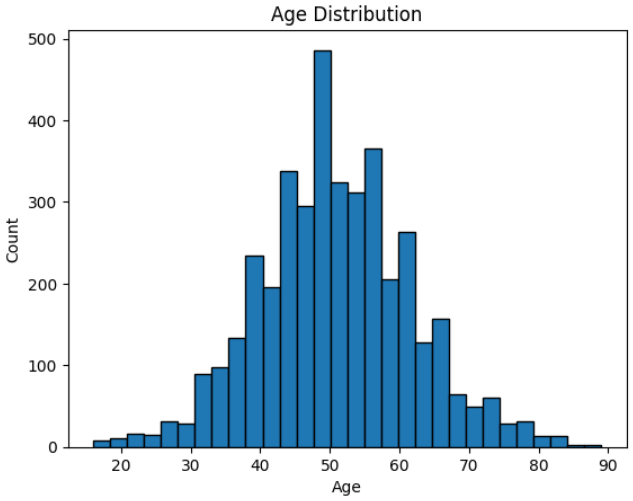


Fig. 6. Distribution of patient age in the CKD dataset

Based on the visualization in Fig. 6, the highest frequency is in the 50-year age range, where the distribution of data includes individuals from about 20 to nearly 90 years of age. This distribution characteristic indicates that the age factor is a variable that has the potential to make a significant contribution to disease prediction, considering that health risks related to kidney function generally increase with the age of the patient.

4) Cross-Validation

This study applied cross-validation with KFold and Stratified KFold for each algorithm with k = 10. K-Fold divides the dataset into 10 balanced parts in each iteration, with one part being used as a test data and the other four parts as a data train. Furthermore, Stratified K-Fold is used to address potential class imbalances. This technique ensures that the proportion of each CKD label in each fold resembles its original distribution, so that each class is represented on both the training data and the test data.

Based on Fig. 7, the best Fold overall was generated by the Random Forest algorithm on the Fold 5 with an accuracy of 0.9896. These results show that in the 10th iteration, the combination of the training data and the test data provides the optimal composition for Random Forest to accurately recognize the patterns of CKD patients. Further evaluation of the best-performing Random Forest (Fold 10) is presented in Fig. 8. From the results of the evaluation of the Random Forest model on the 10th fold, which consists of a confusion matrix and feature importance. The matrix on the left shows a fairly good classification performance with 361 correct predictions in class 0 and 23 correct predictions in class 1, although there are 16 class 1 data that are incorrectly predicted as class 0 (false negative). Meanwhile, the graph on the right identifies that age, Diabetic Year and Weight (Kg) are the three features that have the most significant effect on the model's classification results.

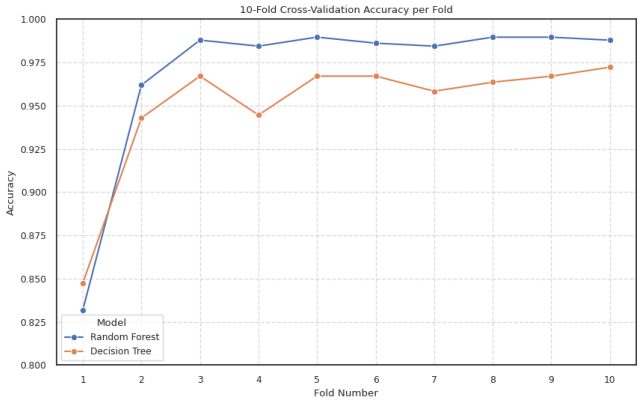


Fig. 7. Accuracy distribution across the 10-fold cross-validation process

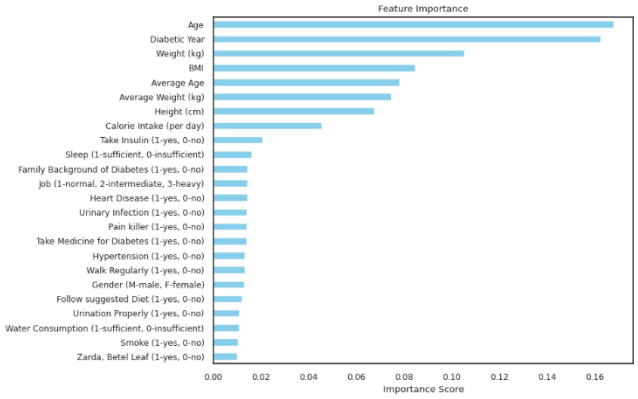


Fig. 8. Confusion matrix and feature importance of the best-performing Random Forest fold

5) Model Performance Evaluation

Performance evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The results of the performance comparison of each machine learning algorithm are presented in Table 2.

TABLE II. COMPARISON OF AVERAGE MACHINE LEARNING MODEL PERFORMANCE

Models	Not CKD (0)				CKD (1)		
	Acc	Prec	Rec	F1	Prec	Rec	F1
Model 1	0.9437	0.9573	0.9822	0.9696	0.7400	0.5362	0.6218
Model 2	0.9237	0.9589	0.9576	0.9582	0.5571	0.5652	0.5612

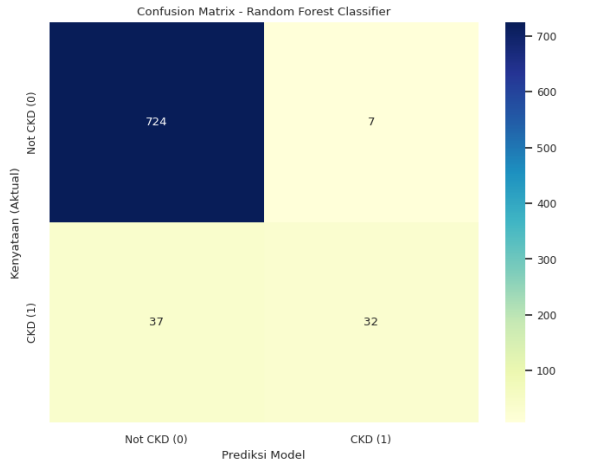
Remark: Model 1 (Random Forest and SMOTE), Model 2 (Decision Tree and SMOTE)

The results of the evaluation showed that machine learning algorithms had different performance in predicting CKD classification. Some algorithms show higher accuracy values, while others excel in recall metrics and F1-scores, which are more relevant for medical prediction cases.

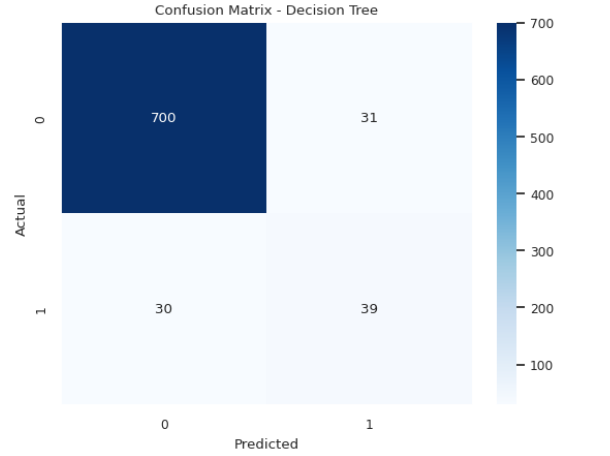
6) Confusion Matrix Analysis

The confusion matrix analysis for each boosting algorithm is shown in Fig. 6. The results of the analysis showed differences in the pattern of misclassification between algorithms, especially in detecting minority classes. This difference indicates that each boosting algorithm has different characteristics in handling naturally unbalanced class distributions.

EISSN: 2961-872X●PISSN: 2961-8924●DOI: 10.56741/IISTR.esl.002553



(a) Random Forest and SMOTE



(b) Decision Tree and SMOTE

Fig. 9. Confusion matrices of Random Forest and Decision Tree with SMOTE

Based on the visualization of the Confusion Matrix in Fig. 9, the Random Forest Classifier model shows an overall superior performance compared to Decision Tree with an accuracy value of 0,9437 and a precision of 0,9573, mainly thanks to its ability to reduce the False Positive number from 31 to 7 cases through an ensemble learning approach. Although Random Forest produced a higher F1-Score of 0,7400, both models were still hampered by the high False Negative number (30–37 cases) due to the dominance of the Not CKD (class imbalance) class, which resulted in a relatively low Recall rate (46%–56%). This low recall is risky in the medical diagnosis of CKD because it can pass patients who are actually sick, so further research is recommended to apply SMOTE or threshold tuning techniques to optimize the detection of positive cases.

C. Discussion

The discussion of the results of this study shows that the Random Forest algorithm provides the most optimal performance in the classification of CKD disease with an accuracy rate of 0.9437, and outperforms Decision Tree which obtained an accuracy of 0.9237. Although Random Forest excels in overall prediction accuracy, Decision Tree has a higher recall value (0.5652), which indicates a better ability to identify patients who are actually suffering from pain to minimize medical misdiagnosis. The significant differences in results in these three methods are influenced by the characteristics of each algorithm in dealing with complex clinical variables, so the study concludes that the selection of the model should consider the balance between total accuracy and sensitivity of early detection for patients.

IV. Conclusion

This study compared Random Forest and Decision Tree models combined with SMOTE for Chronic Kidney Disease (CKD) classification using demographic, lifestyle, and health-related variables. The experimental results showed different performance characteristics between the two models across accuracy, precision, recall, and F1-score. The comparison indicates that model performance should be assessed using multiple evaluation metrics, particularly when the dataset contains class imbalance.

From a practical perspective, the findings provide a basis for selecting and further developing machine learning models for CKD classification and decision-support applications. However, the results should be interpreted within the context of the dataset used in this study. Future research may investigate hyperparameter optimization, alternative ensemble methods, threshold adjustment, and validation using datasets from multiple institutions to improve model generalizability.

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.39-26, 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: https://doi.org/10.35314/7jgjmg64.

[4] 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.

[5] 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, doi: 10.60076/ijstech.v2i3.1108.

[6] H. Y. S Winiarti, F I Indikawati, A Oktaviana, "Consumable Fish Classification Using k-Nearest Neighbor Consumable Fish Classification Using k-Nearest Neighbor," vol. 821, 2020, doi: 10.1088/1757-899X/821/1/012039.

[7] 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.

[8] N. A. Fitri, K. Syaban, and N. Miftachurohmah, "Analisis Kinerja Algoritma Random Forest untuk Klasifikasi Penyakit Diabetes di Puskesmas Wundulako," *JISTech J. Inf. Syst. Technol.*, vol. 2, no. 2, pp. 106–111, Dec. 2025, doi: 10.71234/jistech.v2i2.104.

[9] R. F. N. Iskandar, D. H. Gutama, D. P. Wijaya, and D. Danianti, "Klasifikasi Menggunakan Metode Random Forest untuk Awal Deteksi Diabetes Melitus Tipe 2," *J. Tek. Ind. Terintegrasi*, vol. 7, no. 3, pp. 1620–1626, 2024, doi: 10.31004/jutin.v7i3.26916.

[10] A. Pinandito, S. A. Wicaksono, and S. H. Wijoyo, "Implementasi Machine Learning dalam Deteksi Risiko Tinggi Diabetes Melitus pada Kehamilan," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 4, pp. 739–746, 2023, doi: 10.25126/jtiik.2023107005.

[11] 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.

[12] 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: 10.31764/jtam.v6i4.8808.

[13] 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.

[14] 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.

[15] 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: 10.31004/obsesi.v6i4.1570.

[16] 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.

[17] U. H. Hafist Rezallul Fikry, Rudi Kurniawan, Bani Nurhakim, "Sistem Klasifikasi Penyakit Diabetes Menggunakan Algoritma Random Forest Berdasarkan Data Medis," vol. 3, no. 01, pp. 2–7, 2026, doi: https://ruangjurnal.or.id/index.php/jcsai/issue/view/7.

[18] 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, doi: 10.13026/42eg-8e59.

[19] 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.

[20] 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.

[21] M. H. A. Raihan, H. 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, doi: 10.31942/jinrpl.v7i2.12727.

[22] Z. M. J. Nafis, R. Nazilla, R. Nugraha, and S. 'Uyun Shofwatul 'Uyun, "Perbandingan Algoritma Decision Tree dan K-Nearest Neighbor untuk Klasifikasi Serangan Jaringan IoT," *Komputika J. Sist. Komput.*, vol. 13, no. 2, pp. 245–252, 2024, doi: 10.34010/komputika.v13i2.12609.

[23] F. Y. Pamuji and V. P. Ramadhan, "Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotheraphy," *J. Teknol. dan Manaj. Inform.*, vol. 7, no. 1, pp. 46–50, 2021, doi: 10.26905/jtmi.v7i1.5982.

[24] R. B. Aldwin, R. F. Orslid, P. M. Artin, G. O. Ttaviano, and F. R. O. Icoud, "Perbandingan Kinerja Model Klasifikasi dalam Memprediksi Intensi Pembelian Pengunjung E-Commerce," *J.Dinamik.*, vol. 30, no. 1999, pp. 1–26, 2002, doi: https://doi.org/10.35315/dinamik.v30i2.10593.

[25] H. Yuliansyah, R. Ardiansyah, and A. Yudhana "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, doi: 10.31942/jinrpl.v7i2.12727.

[26] S. A. Hidayah and Z. Abidin, "Prediksi Penyakit Asma Menggunakan Naïve Bayes dan Random Forest Berbasis Data Klinis Multivariat," *Jurnal Algoritma.*, pp. 1756–1767, 2026, doi: 10.33364/algoritma/v.23-1.3521.

[27] H. Yuliansyah, R. Ardiansyah, and A. Yudhana, "A Machine Learning Approach for Multi-Label Classification in Candidate Election Social Media Analysis," *ECTI Trans. Comput. Inf. Technol.*, vol. 20, no. 1, pp. 50–62, 2026, doi: 10.37936/ecti-cit.2026201.263596.

[28] H. Razak, F. R., Biddinika, M. K., and Yuliansyah, "Radial basis function model for obesity classification based on lifestyle and physical condition," *J. Tek. Elektro, Teknol. Inf. dan Komput.*, vol. 8, no. 2, pp. 192–200, 2024, doi: http://doi.org/10.31961/eltikom.v8i2.1347.