MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Open Access Journal of Data Science & Artificial Intelligence
ISSN: 2996-671X

Predictive Analysis of Chronic Kidney Disease Using Patient Electronic Health Records

Kottaisamy K[1]* and Chidambaranathan S[2]

[1]Assistant Professor, PG Department of Computer Applications (MCA & MCA GEN AI), FLABS, SRM Institute of Science and Technology (Deemed to be University), India

[2]Associate Professor, Department of Computer Applications, St. Xavier's College (Autonomous), India

Research Article

Volume 4 Issue 1

Received Date: October 26, 2026

Published Date: November 10, 2026

DOI: 10.23880/oajda-16000169

***Corresponding author:** Kottaisamy K, Assistant Professor, PG Department of Computer Applications (MCA & MCA GEN AI), FLABS, SRM Institute of Science and Technology (Deemed to be University), Ramapuram Campus, Chennai-600 089, India

Abstract

In emerging and underdeveloped nations, machine learning-based medical diagnostics can aid in the early detection of chronic kidney disease. Kidney illness is detected using machine learning algorithms on a real-time dataset. Chronic kidney disease, also known as chronic renal disease or chronic kidney failure, is the incapacity of the kidneys to carry out their normal functions. One potentially lethal condition is chronic renal illness. The glomerular filtration rate steadily declines for at least three months due to a common medical condition. The study trained and assessed the proposed model using the renal disease dataset. The projected degree of prediction accuracy of the proposed model seems to be confirmed by experimental data. The patient's electronic medical information and machine learning may identify patterns associated with certain diseases and alert physicians to any anomalies. Numerous feature selection and classification methods are employed to forecast renal illness. You can avoid early renal failure by anticipating kidney illness.

Keywords: Glomerular Filtration Rate; Machine Learning; Chronic Kidney Disease; Renal Failure; Medical Diagnostics

Abbreviations

CKD: Chronic Kidney Disease; GFR: Glomerular Filtration Rate; SVM: Support Vector Machines.

Introduction

The development of machine learning algorithms in domains relevant to illness detection emphasizes the technology's application in the medical industry. Machine learning can detect patterns associated with specific diseases in inpatient electronic medical data and alert doctors to any irregularities. Early renal failure can be prevented by using a range of feature selection and classification techniques to predict renal disease. One of the body's most crucial organs, the kidney, filters the body's water and waste products to produce urine. Chronic kidney disease (CKD), also known as chronic renal disease or chronic kidney failure, is a condition where the kidneys are unable to carry out their usual functions. CKD is a potentially lethal disorder. Glomerular Filtration Rate (GFR) declines steadily for three months or more due to a common health condition. Hypertension, irregular foamy urine, vomiting, itching, shortness of breath, and cramping are among the disease's typical symptoms [1],

CC BY

and its primary causes are diabetes and high blood pressure.

Dialysis or kidney transplant patients frequently have CKD identified in its later stages. When dialysis treatment or kidney transplantation are the only alternatives left to preserve the patient's life, CKD is frequently discovered in its latter stages. On the other hand, a timely diagnosis can help prevent renal failure [2]. Monitoring Glomerular Filtration Rate (GFR) on a regular basis is the most effective way to assess kidney function or to determine the stages of renal disease [3].

CKD symptoms are independent of any one illness. Certain people may not even detect any signs at all as the conditions steadily worsen. As a result, it could be challenging to detect the illness in its beginning stages. Recent advances in machine learning have made it possible to identify diseases by simply analyzing current patients' medical histories and creating a model to predict behavior.

This paper's primary goal is to precisely detect chronic renal illness and assist mankind in fighting this dangerous disease, but we defined the objectives of this paper in a specific manner, namely,

- To extract useful information with classification accuracy for chronic renal disease prediction.
- On the required datasets, use the binary classification of the Sklearn Random Forest classifier and Ensemble Random Forest methods.
- To forecast which of the two methods employed in this paper will perform best in identifying renal illness.

Numerous studies in Part II use Machine Learning Classifier techniques to predict CKD. Section III describes the dataset and explores two distinct approaches: preprocessing and model development. Section IV provides the experimental investigation of CKD prediction. Section V of this book contains the conclusion

Related Works

Numerous studies have examined the difficulty of utilizing a machine-learning model to diagnose renal illnesses. The challenge of using a model based on machine learning to diagnose renal diseases has been the subject of numerous investigations. Qin J, et al. [4] provided data assertions and a sample diagnosis to help in the diagnosis of CKD. KNN is used in data assertion. Six classifier techniques are used to assess the precision of a diagnosis: the logistic regression method, the random forest, K-NN, a Naive Bayes algorithm, Feed-forward Neural Network, and SVM. With a 99.75% accuracy rate, the random forest offers higher performance in these classifiers. Vasquez-Morales GR, et al. [5] developed a typical using neural networks with an accuracy rating of 95% for predicting the risk of the onset of chronic renal illness using a dataset of 40000 occurrences.

As treating CKD is expensive and intrusive, many people do not receive therapy until it is too late. Therefore, it is still crucial to find this disease early. Additionally, SVM machine learning classifier method testing results with an accuracy of 93% were provided by Amirgaliyev Y, et al. [6]. Machine learning may be used to identify patterns linked to certain diseases and notify doctors of any irregularities in inpatient electronic medical data. The results of the study show that improving the efficiency of feature selection in classification algorithms can result in an increase in the accuracy of machine learning techniques used in the diagnosis and prediction of chronic renal disease, such as Decision Tree, Random Forest Classifier, and Extra Tree. The trial's findings may aid in the early detection of issues and the development of treatment plans for kidney disease, ultimately saving lives. The suggested model's experimental findings seem to support the anticipated level of prediction accuracy.

De Almeida LK, et al. [7] study utilized RBF functions, Support Vector Machines (SVM) using polynomial, linear, and sigmoid in addition to Decision Trees, Random Forests, and SVM. Aimed at their investigation, they used the MIMIC-II record. With respective prediction accuracy of 80% and 87%, they came to the conclusion that random forests and decision trees generated the finest outcomes. Almasoud M and Ward TE [8] used the CKD dataset's 400 occurrences and 25 characteristics. Albumin, Hemoglobin, and specific gravity were discovered to be the feature characteristics of the CKD dataset when they used the filter feature evaluation approach to attributes. Following the selection of the features, they analyzed the dataset and performed 10-fold cross-validation. The approach, gradient boosting, provided the greatest accuracy of 99.1%. To assess whether a patient has CKD or not, machine learning methods may be applied. 400 Indian patient records with 25 variables were collected from the Apollo Hospital in Tamil Nadu for the dataset. A number of machine learning algorithms are used to split the dataset into trained and tested portions. Early detection of renal issues lowers the risk of CKD. The Decision Tree, Logistic Regression, Random Forest, K-Nearest Neighbor, Support Vector Classifier, and are used as classification approaches. The best and most effective method for detecting renal disease is this one. When diagnosing kidney disease, the Random Forest Classifier performs the best [9]. Xiao J, et al. [10] utilized nine machine learning techniques to the data of 551 patients, including Logistic regression, XGBoost, Lasso regression, SVM, Ridge regression, Neural Network, Random Forest, K-Nearest Neighbor, and Elastic Net. The linear model had the highest accuracy, according to their analysis of precision, recall, ROC curve, and accuracy.

Copyright© Kottaisamy K and Chidambaranathan S.

Materials and Methods

Data Source & Description

This study's dataset is based on chronic renal disease patient information received from the Kidney Care Centre in Tamil Nadu, India. There are 1100 observations in the dataset. The dataset contains 821 CKD patients and 279 non-CKD observations, as well as 13 numerical, 14 nominal, and 1 feature that reflects the severity of kidney disease. Id, Age, Blood pressure, Pedal edema, Potassium, Albumin, Gender, Red blood cell count, Specific gravity, Coronary artery disease, Sodium, Appetite, Serum creatinine, Hemoglobin, Pus cell clumps, White blood cell count, Hypertension, Diabetes mellitus, Anemia, Pus cell, Sugar, Red blood cells, Blood urea, Bacteria, Blood glucose random, Packed cell volume, and CKD or Not CKD data columns the data fields in the dataset.

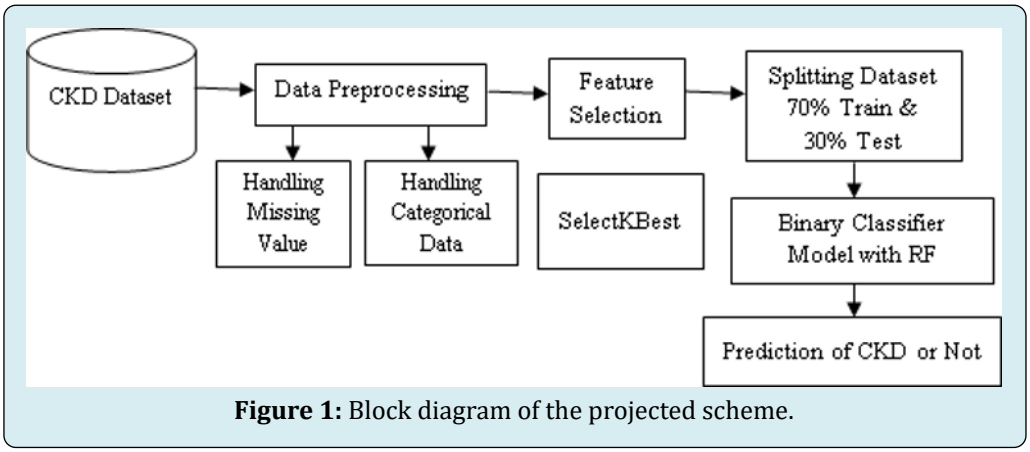

Figure 1: Block diagram of the projected scheme.

In the figure 1 depicts the algorithmic technique proposed in this work. After feeding the data set into the prediction algorithm, the evaluation model, which includes a confusion matrix, is used to assess the method's classification accuracy. Finally, the most reliable method for determining whether or not someone has a chronic renal illness is revealed.

Preprocessing of Data

Before preprocessing, we assess the missing data to determine the fraction of missing data for each variable in the data. Cleaning the raw data is an important step in developing a prediction model because it improves the performance of machine learning algorithms. To replace missing values with the most frequently occurring value, use the "most frequently" approach. To convert category variables to numerical numbers, label encoding is used.

Before modeling could begin, the data set was divided into two sections: the Train set (70%), which was used for selecting and validating models, and the Test set (30%), which included data used to determine how well models generalized to new data.

Selection of Features

In this stage SelectKBest for classification, Choose the SelectKBest class to extract the best features from a certain dataset. We can apply the method classification data by shifting the "score_func" input. When preparing a large dataset for training, it is crucial to find the best features. The SelectKBest methodology chooses attributes built on the given k maximum score. It allows us to discard less important facts and minimise training time. Using the SelectKBest class, define the model. As a scoring function for categorization, we'll use the 'chi2' approach. The k parameter specifies the desired number of characteristics. The fit and transform approach will then be used to training x and y data. k is the number of features you wish to choose from, with options of 20, 15, and 10.

Model of Machine Learning

Model development, which involves forecasting whether or not CKD would manifest, is the most important stage. In this work, we employed an Ensemble Random Forest of machine learning techniques to predict CKD.

For supervised learning, the classification method Random Forest is employed. It develops a substantial forest of trees to analyze accuracy effectively. The precision of this classifier increases with the number of trees. Even when no hyper parameters are adjusted, Random Forest yields more dependable results because to its flexibility. It is easy to use and efficient, particularly when working with big data sets. It maintains accuracy by identifying anomalies and outliers. Nevertheless, it is challenging to implement and expensive to compute. In order to build the Decision Tree, to build the

Copyright© Kottaisamy K and Chidambaranathan S.

Decision Tree using entropy and information gain. Consider a dataset that has N classifications. The following formula can be used to compute entropy:

$$Entropy(S)=\sum_{i=1}^{N}-P_i \log_2 P_i$$

The likelihood of picking a random example from class i is given by p_i. Information Gain represents the quantity of information gained by a feature for the target variable.

Information gain= $Entropy_{parent} - Entropy_{children}$

The parent node's entropy is signified by $Entropy_{parent}$. The average entropy of the child nodes that originate after this variable is denoted by $Entropy_{children}$

To evaluate the models' effectiveness, we must first identify the relevant metrics. We have four essential features (numbers) from which we may generate measurement metrics to describe the classifiers' performance in the digital diagnostics challenge. The Ensemble Random Forest (ERF) algorithm, which was used in this study, generates output in the following steps:

Ensemble Random Forest (ERF) Algorithm:
Input: Training sets N, Randomized subset of Features F,
Output: Ensemble Random Foresight's Binary classifier
Step 1: Draw samples at random from the original chronic kidney disease dataset. The bootstrapped dataset is a collection of such 'N' randomly selected samples.
Step 2: Preprocessing data with the Simple Imputer method
Step 3: For the bootstrapped data set, create a decision tree by accounting for a random subset of variables is f ⊆ F.
Step 4: Splitting the tree repeatedly until the tree reaches one leaf node and is still complete and to the greatest extent possible, repeat the previous steps 100 times.
Step 5: Run each decision tree on a fresh data point to forecast its result.
Step 6: The majority of votes cast determines the anticipated class.
Step 7: Use the dataset's out-of-bag instances to assess the model and establish the final class.
Step 8: Using test data, the most accurate models will be used to predict whether or not a person has CKD.
Step 9: Print the performance metrics for the model.
End

Prediction Model: The main indicators must first be defined in order to assess the models' efficacy. The four fundamental characteristics (numbers) from which measurement metrics to characterize the classifiers' performance in the digital diagnostics task may be derived. The four numbers are as follows: False Positive - The number of patients who were classified as healthy when they were not. False Negative - number of patients misdiagnosed with the illness. The proportion of patients who were correctly acknowledged as having the disease. TN (True Negative) - number of correctly classified healthy patients. By measuring sensitivity, accuracy, specificity, confusion matrix, and F-measure using the related scientific relationships provided below, classification performance is evaluated.

Accuracy is one of the most well-liked categorization performance measures. It is the percentage of accurately categorized samples to all samples. The investigation's accuracy assessment procedure is as follows.

$$Accuracy=\frac{(True\ positive+True\ negative)}{(True\ positive+False\ positive+True\ negative+False\ negative)}$$

The proportion of affirmative identifications that were in fact accurate is known as precision. The study's formula for precision calculation is

$$Precision=\frac{(True\ positive)}{(True\ positive+False\ positive)}$$

F-Measure is computed using the weighted average of the precision and sensitivity measurements. The following formulation was utilized in this investigation to determine f-measure. Using the area of information recovery, F-Measure gauges categorization performance.

$$F1-Score=\frac{2}{\frac{1}{Recall}+\frac{1}{Precision}}$$

A recall is the proportion of all data samples for a class that a model developed using machine learning correctly identifies as being a member of an interesting class (the "positive class"), among every sample for that class.

$$Recall=\frac{(True\ positive)}{(True\ positive+False\ negative)}$$

Results and Discussion

After applying the random forest using sklearn method and ensemble random forest machine-learning algorithms to the dataset, to obtain the accuracy values listed below. Each model can use training and testing datasets with features such as 20 features, 15 features, and 10 features.

The Training and Validation Model Score is shown in Figure 2.

Copyright© Kottaisamy K and Chidambaranathan S.

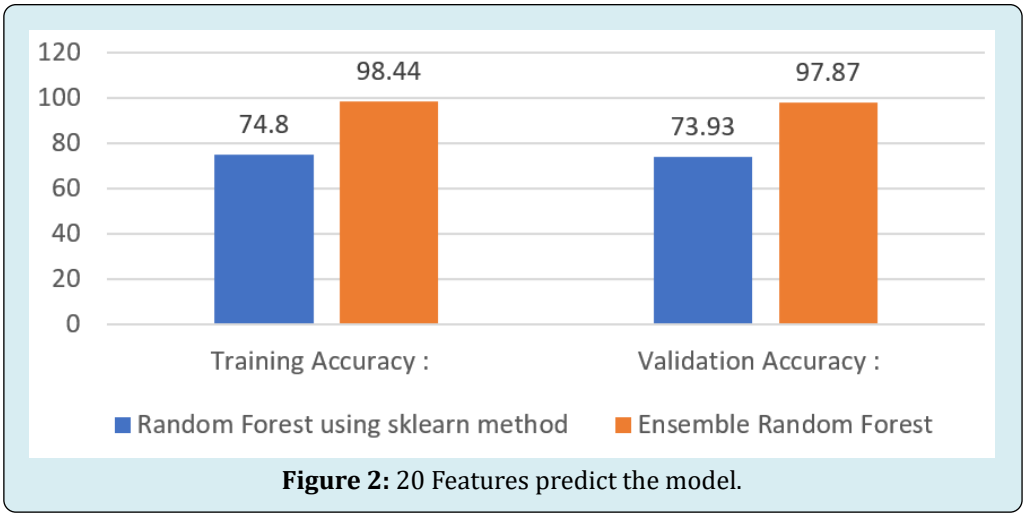

Figure 2: 20 Features predict the model.

The Training and Validation Model Score is shown in Figure 3.

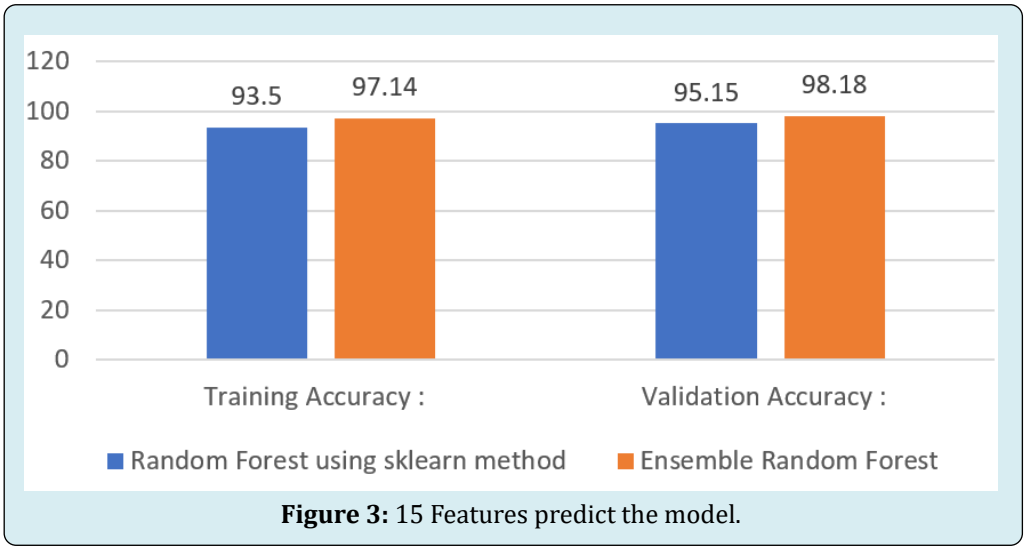

Figure 3: 15 Features predict the model.

The Training and Validation Model Score is shown in Figure 4.

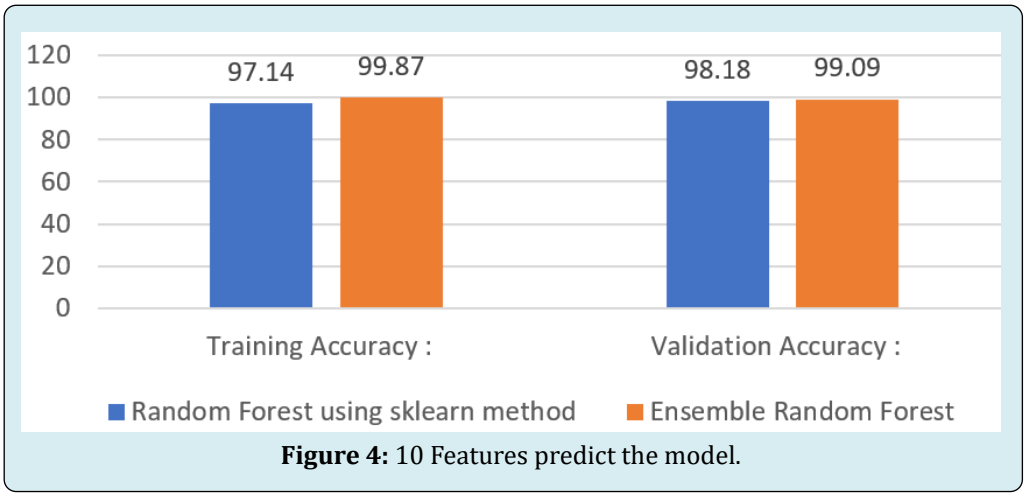

Figure 4: 10 Features predict the model.

Copyright© Kottaisamy K and Chidambaranathan S.

Figures 2, 3, 4, and tables 1, 2, and 3 indicate that the training and validation model accuracy for the features are 20, 15, and 10.

	Recall	Precision	F1-Score	Support
CKD	0.66	1	0.79	250
Not CKD	1	0.48	0.65	80
Accuracy			0.74	330
Macro avg	0.83	0.74	0.72	330
Weighted avg	0.74	0.87	0.76	330

Table 4: Displays the classification performance of 20 features in the confusion matrix using the Random Forest using sklearn method.

	Recall	Precision	F1-Score	Support
CKD	0.94	1	0.97	250
Not CKD	1	0.83	0.91	80
Accuracy			0.95	330
Macro avg	0.97	0.92	0.94	330
Weighted avg	0.95	0.96	0.95	330

Table 5: Displays the classification performance of 20 features in the confusion matrix using the Ensemble Random Forest.

	Recall	Precision	F1-Score	Support
CKD	0.94	1	0.97	250
Not CKD	1	0.83	0.91	80
Accuracy			0.95	330
Macro avg	0.97	0.92	0.94	330
Weighted avg	0.95	0.96	0.95	330

Table 6: Displays the classification report of 15 features in the confusion matrix using the Random Forest using sklearn method.

	Recall	Precision	F1-Score	Support
CKD	0.99	1	0.99	250
Not CKD	0.99	0.96	0.98	80
Accuracy			0.99	330
Macro avg	0.99	0.98	0.98	330
Weighted avg	0.99	0.99	0.99	330

Table 7: Displays the classification performance of 15 features in the confusion matrix using the Ensemble Random Forest.

	Recall	Precision	F1-Score	Support
CKD	0.99	0.99	0.99	250
Not CKD	0.97	0.96	0.97	80
Accuracy			0.98	330
Macro avg	0.98	0.98	0.98	330
Weighted avg	0.98	0.98	0.98	330

Table 8: Displays the classification performance of 10 features in the confusion matrix using the Random Forest using sklearn method.

Copyright© Kottaisamy K and Chidambaranathan S.

	Recall	Precision	F1-Score	Support
CKD	0.99	0.99	0.99	250
Not CKD	0.99	0.99	0.99	80
Accuracy			0.99	330
Macro avg	0.99	0.99	0.99	330
Weighted avg	0.99	0.99	0.99	330

Table 9: Displays the classification performance of 10 features in the confusion matrix using the Ensemble Random Forest.

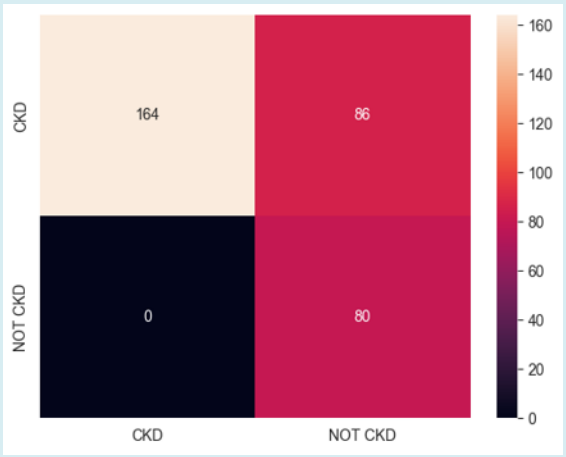

Figure 5: RF using sklearn with 20 features.

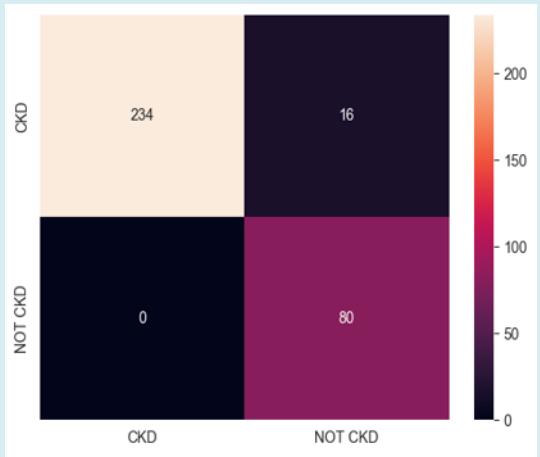

Figure 7: RF using sklearn with 15 features.

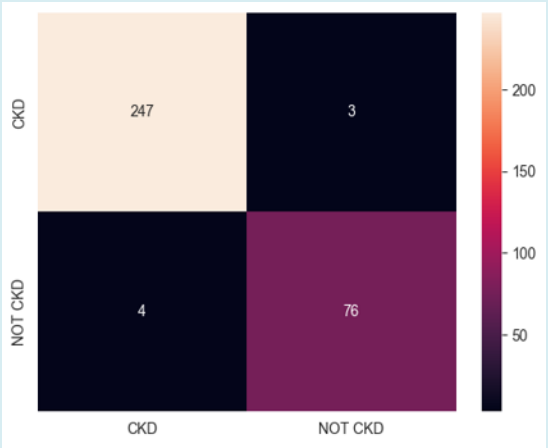

Figure 6: ERF with 20 features.

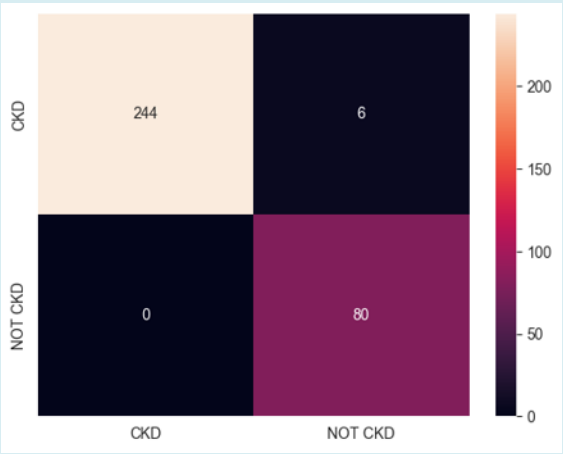

Figure 8: ERF with 15 features.

Figure 5 Random Forest classification report with 20 characteristics using sklearn. The CKD accurately predicted by sklearn Random Forest classifier is 164, and the Not CKD is 80. 86 predicted a terrible outcome, which was an inaccurate report with 20 features. Figure 6 Report on Ensemble Random Forest classification with 20 features. The CKD accurately predicted by the Random Forest classifier is 247, and the Not CKD is 76. 4 predicted a negative outcome, which was an inaccurate report with 20 features.

Figure 7 Random Forest classification report with 15 characteristics using sklearn. The CKD accurately predicted by sklearn Random Forest classifier is 234 and the Not CKD is 80. 16 predicted a terrible outcome, which was an inaccurate report with 15 features. Figure 8 Report on Ensemble Random Forest classification with 15 features. The CKD accurately predicted by the Random Forest classifier is 244, and the Not CKD is 80. 6 predicted a negative outcome, which was an inaccurate report with 15 features.

Copyright© Kottaisamy K and Chidambaranathan S.

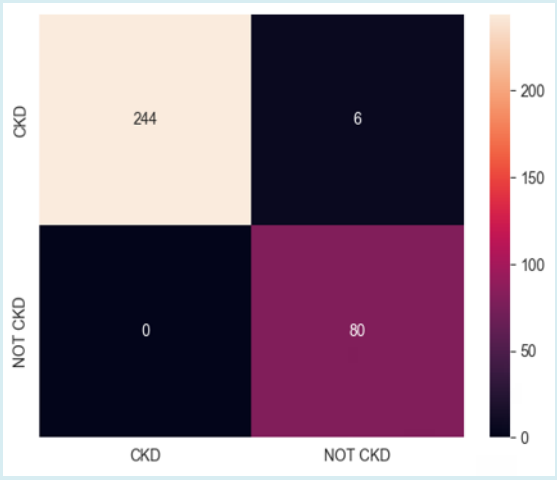

Figure 9: RF using sklearn with 10 features.

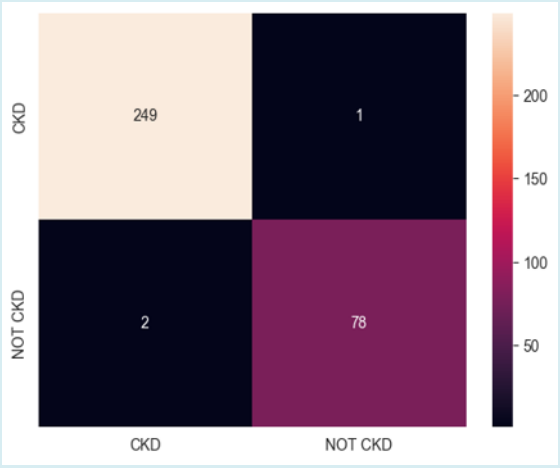

Figure 10: ERF with 10 features.

Figure 9 Random Forest classification report with 10 characteristics using sklearn. The CKD accurately predicted by the sklearn Random Forest classifier is 244, and the Not CKD is 80. 6 predicted a disastrous outcome, an erroneous report with 10 characteristics. Figure 10 Report on Ensemble Random Forest classification with 10 features. The Random Forest classifier successfully predicted the CKD to be 249, and the Not CKD to be 78. 3 predicted a negative outcome, which was an erroneous report with 10 characteristics.

Conclusion

To examining the medical histories of present patients and developing a model to forecast behavior, illnesses may now be identified thanks to recent developments in machine learning. The dataset under investigation was subjected to a number of machine learning approaches in addition to the random forest classifier. Our study revealed that those with both diabetes mellitus and hypertension had little knowledge of chronic renal disease and its risk factors. The best feature selection techniques for medical datasets were found using the classification model. F-measure, Precision, Recall, and Accuracy are the four metrics used to quantify performance. The trial's findings may help the medical community make early predictions and decisions about renal disease therapy and life-saving interventions.

References

1. Webster AC, Nagler EV, Morton RL, Masson P (2016) Chronic kidney disease. Lancet 6736(16): 1–15.
2. Serpen AA (2016) Diagnosis rule extraction from patient data for chronic kidney disease using machine learning. Int J Biomed Clin Eng 5(2): 64–72.
3. Tekale S, Shingavi P, Wandhekar S (2018) Prediction of chronic kidney disease using machine learning algorithm. IJARCCE 7(10): 92–96.
4. Qin J, Chen L, Liu Y, Liu C, Feng C, et al. (2020) A machine learning methodology for diagnosing chronic kidney disease. IEEE Access 8: 20991–21002.
5. Vasquez-Morales GR, Martinez-Monterrubio SM, Moreno-Ger P, Recio-Garcia JA (2019) Explainable prediction of chronic renal disease in the Colombian population using neural networks and case-based reasoning. IEEE Access 7: 152900–152910.
6. Amirgaliyev Y, Shamiluulu S, Serek A (2018) Analysis of chronic kidney disease dataset by applying machine learning methods.
7. De Almeida LK, Lessa L, Peixoto A, Gomes R, Celestino J (2020) Kidney failure detection using machine learning techniques.
8. Almasoud M, Ward TE (2019) Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int J Adv Comput Sci Appl 10(8).
9. Kottaisamy K, Chidambaranathan S (2020) Comparative study of kidney disease prediction using classification algorithms. Studies in Indian Place Names 40(71): 923–926.
10. Xiao J, Ding R, Xu X, Guan H, Feng X, et al. (2019) Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med 17(1): 119.

Copyright© Kottaisamy K and Chidambaranathan S.

YOUR RESEARCH DESERVES GLOBAL VISIBILITY

• Publish • Discover • Share • Connect

MEDICAL	CLINICAL	LIFESCIENCES	PHARMA
90+ OPEN ACCESS JOURNALS		400+ RESEARCH TOPICS	

SUBMIT YOUR RESEARCH
Submit your manuscript through our online submission system.

REPRINTS
If you require a reprint of any article, please contact us.
info@medwinpublishers.com

BECOME A REVIEWER
Share your expertise and contribute to scholarly publishing.

SCAN TO EXPLORE MEDWIN JOURNALS

• Journals • Submit Manuscript • Reviewer Program • Article Tracking

Our Journals indexed in

Embase® ELSEVIER | PubMed National Library of Medicine NLM | Crossref | Google Scholar | INDEX COPERNICUS INTERNATIONAL | ISSN INTERNATIONAL STANDARD SERIAL NUMBER INTERNATIONAL CENTRE | ROAD DIRECTORY OF OPEN ACCESS SCHOLARLY RESOURCES | EuroPub | Scilit | WorldCat® | Academic Resource Index ResearchBib | ADVANCED SCIENCE INDEX | INTERNATIONAL Scientific Indexing | ICMJE INTERNATIONAL COMMITTEE of MEDICAL JOURNAL EDITORS

This article published by **Medwin Publishers,** 28475 Witherspoon Dr, Novi, MI 48377, United States, +1-248-247-6042, WhatsApp +1 (248) 275-1133, info@medwinpublishers.com

© 2026, by the Authors. The articles published from this journal are distributed to the public under "**Creative Commons Attribution License**" (https://creativecommons.org/licenses/by/4.0/). Therefore, upon proper citation of the original work, all the articles can be used without any restriction or can be distributed in any medium in any form.