A Machine Learning based Technique to Detect Kidney Disease

Dr. Farooque Azam, Bangalore ,India
Akshay Jha, Bangalore ,India
Arbaazkhan R Soudagar, Bangalore ,India
Bijay B K, Bangalore ,India
Yogesh S Pawar, Bangalore,India

***Abstract*— Chronic kidney disease is the term used to describe the state of the kidneys as a result of conditions like diabetes, glomerulonephritis, or high blood pressure. These problems could creep up on you slowly over a long time, frequently without any symptoms. Renal failure may eventually set in, requiring dialysis or a kidney transplant to extend life. Therefore, with early discovery and treatment, many repercussions can be avoided or postponed. This endeavour aims to improve diagnosis precision while speeding up diagnosis through the use of classification algorithms. The proposed study classifies the various stages of chronic renal illness using machine learning techniques. Results from tests utilising a variety of techniques, including Naive Bayes, Decision Trees, K-Nearest Neighbor, and Support Vector Machines.**

Keywords—Chronic Kidney Disease, Machine Learning, Prediction, PCA, Co-relation Metrics, Random Forest.

I. INTRODUCTION

One of the most crucial bodily organs, the kidney filters all of the waste products and water from the body to produce urine. Chronic Kidney Disease (CKD), usually referred to as chronic renal disease or chronic kidney failure, is a potentially fatal condition caused by the kidneys' failure to carry out their normal functions. It is a widespread health issue that causes Glomerular Filtration Rate (GFR) to continuously decline for three months or longer. Common signs of the condition include high blood pressure, frothy urine that isn't regular, vomiting, shortness of breath, itching, and cramps. [1], but diabetes and high blood pressure are the primary causes of this condition. When dialysis or a kidney transplant are the only treatments left to save the patient's life, CKD is frequently discovered in its later stages. Conversely, renal failure can be avoided with an early diagnosis [2]. Monitoring the Glomerular Filtration Rate (GFR) on a regular basis is the best technique to assess kidney function or determine the stages of renal disease [3]. GFR is computed using blood creatinine, age, gender, and race. the worth of a person. Instead of performing numerous expensive tests, a machine learning algorithm may predict this with a lot more aid for the clinician. We only need to input a few of the patient's acquired facts into the computer, and we can quickly determine whether a patient has CKD.

II. RELEATED WORK

A comparison research on CKD prediction using SVM and K-NN classifiers was conducted by Sinha et al. in 2015 [4]. According to the experimental findings, K-NN classifier outperformed SVM with an accuracy of 78.75% compared to SVM's accuracy of roughly 73.75%. Accuracy, execution time, and precision were used to calculate the algorithms' performance.K. A. Padmanaban and G. Parthiban conducted research employing classifiers such as Naive Bayes and Decision tree approaches in the WEKA tool for the early prediction of CKD. They found that decision trees were 91% more accurate than Nave Bayes [5]. Medical professionals must identify the precise remedies in order to save lives. They require the aid of machine learning techniques to accomplish this. For this, Charleonnan and her colleagues looked into a variety of machine learning techniques. For CKD identification, they used the classifiers DT, LR, SVM, and KNN. Their findings support the SVM methodology as the best method for detecting this illness [6]. An increase in albumin discharged through urine can be caused by CKD. Utilizing a dataset made up of 250 individuals with CKD and 150 healthy patients, Celik et al. sought to diagnose and predict CKD [7]. They have made use of classifiers like decision trees and support vector machines. They used the J48 programme of the WEKA tool and the sequential minimum optimization (SMO) approach to create these classifiers. Wibawa and squad in 2017 [8] utilised correlation-based feature selection and AdaBoost as an ensemble learning technique (CFS). They compared AdaBoost and CFS against Naive Bayes, KNN, and SVM for the goal of detecting CKD and found them to be the

most reliable classifiers. Devika et al. suggested using classifiers as an analogy for predicting CKD [9]. Their work is based on classifiers from Naive Bayes, K-NN, and Random Forest. The results of the experiment demonstrate that the RF classifier is significantly superior. Machine learning has been extremely important in the field of disease diagnosis over the past year, and it has elevated medical diagnosis to a whole new level. Similarly, Sivaranjani. documentation, they discussed ML techniques like SVM and RF [10]. Forward and backward selection were used to choose the features, and Principal Component Analysis was used to reduce the dimensionality of the data. As compared to SVM, the results demonstrate that RF provided the superior accuracy. For the objective of predicting cancer, Shaikh F.J. et al. conducted research on decision trees (DT), support vector machines (SVM), and artificial neural networks (ANN) [11]. In their study publication from 2019 [12], Dahiwade D et al. used K-NN and convolutional neural network for precise disease prediction. Amritavarshini and her colleagues developed a paper about multimodal systems that reduce traffic congestion in the year 2020. Amritavarshini and her colleagues developed a study on multimodal systems that improves the performance of authentication systems that combine a person's physical or behavioural characteristics [13]. All areas of the medical background, including emotional recognition, have been impacted by machine learning. A work on a multimodal system for emotion recognition was just released by Veni S and Thushara S [14]. The work of Saiharsha B et al. to quantify the effectiveness of deep learning structures in the context of picture categorization is noteworthy [15].

III. PROPOSED METHODOLOGY

Table 1 : CKD Stages According to GFR Measurement Values

Stage	GFR	Description
1	90–100 mL/min	Normal kidney function
2	60–89 mL/min	Mildly reduced function
3A	45–59 mL/min	Moderately function
3B	30–44 mL/min	Moderately function
4	15–29 mL/min	Severely reduced function
5	<15mL/min or dialysis	End stage kidney failure

As demonstrated in Table 1, CKD can be divided into six stages based on the value of GFR. CKD symptoms do not depend on a particular illness. Some patients may not experience any symptoms at all as the symptoms appear gradually. As a result, it is quite challenging to identify the disease in its early stages. By simply evaluating the patient records of current patients and training a model to predict the behaviour of new patients, machine learning (ML) has recently played a key role in the diagnosis of diseases [3]. ML is a subfield of artificial intelligence in which the computer learns autonomously and improves its predictions over time. Supervised learning is a subset of machine learning that can be applied to datasets for classification or regression. In many different fields, especially in biomedicine for the identification and categorization of various disorders, machine learning is utilised quite successfully. Each ML method has its own strengths and weaknesses, and they can all be used to forecast diseases. Among them, decision-tree offers more accurate classified reports for illnesses related to the kidney [3]. As a result, it appears to be an excellent tool for developing a prediction system to identify kidney illnesses at an early stage.

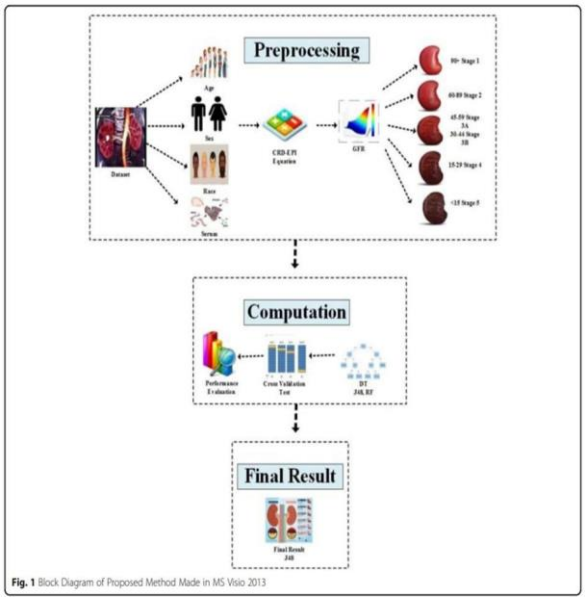

Fig. 1 Block Diagram of Proposed Method Made in MS Visio 2013

A. Methods

In order to forecast the stages of chronic renal disease, this study presents the results in three steps, namely preprocessing, calculation, and final results. The authors created the block diagram for the suggested method in the MS Visio 2013 programme, which is displayed in Fig. 1. The procedures were developed in accordance with the necessary standards and laws.

B. Preprocessing

The collection of the patient dataset for this phase is the first step. Age, sex, race, and serum creatinine are the four parameters chosen from the dataset to be used as inputs in the calculation of GFR. The Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) Equation [9] is one of many mathematical equations that have been used to estimate GFR in the literature. Comparing this equation to Modification of Diet in Renal Disease, it is more accurate for calculating all stages of CKD (MDRD) Equation that solely considers age, gender, and ethnicity as well as serum creatinine is known to be accurate when GFR is greater than 60, which is the situation for CKD that is in its later stages.

C. Dataset

The dataset for the proposed system has been selected from the University of California Irvine (UCI).

Table 2 : Variable Description Used in Analysis

Attribute Symbols and Description	Type	Class
age (Age)	Numerical	Predictor
bp (Blood Pressure)	Numerical	Predictor
sg (Specific Gravity	Nominal	Predictor
al (Albumin)	Nominal	Predictor
su (Sugar)	Nominal	Predictor
rbc (Red Blood Cells)	Nominal	Predictor
pc (pus Cell)	Nominal	Predictor
pcc (Pus Cell Clumps)	Nominal	Predictor
rc (Race)	Nominal	Predictor
bgr (Blood Glucose Random)	Numerical	Predictor
bu (Blood Urea)	Numerical	Predictor
sc (Serum Creatinine)	Numerical	Predictor
sod (Sodium)	Numerical	Predictor
pot (Potassium)	Numerical	Predictor
hemo (Hemoglobin)	Numerical	Predictor
pcv (Packed Cell Volume)	Numerical	Predictor
sex (Sex)	Nominal	Predictor
rc (Red Blood Cell Count)	Numerical	Predictor
htn (Hypertension)	Nominal	Predictor
dm (Diabetes Mellitus)	Nominal	Predictor
appet (Appetite)	Nominal	Predictor
ane (Anemia)	Nominal	Predictor
class (Class)	Nominal	Target

Table 2 lists the 400 instances and 25 attributes of the Machine Learning Repository along with their descriptions, types, and classes. There are just two classifications in this dataset: those with chronic kidney disease (CKD) and those without it (NOTCKD). The proposed system further divides the CKD class into various stages, with Stage 1 denoting normal kidney function, Stage 2 denoting mildly reduced kidney function, Stage 3A denoting moderately reduced kidney function, Stage 3B denoting moderately reduced kidney function, Stage 4 denoting severely reduced kidney function, and Stage 5 denoting end-stage kidney failure of CKD, as shown in Table 1. All retrieved attributes are represented by symbols and descriptions in Table 2; The type column of Table 2 displays the datatype of the attributes, while the third column, class, categorises the attributes of the dataset into two categories, predictor and target. Target will be predicted using predictor attributes. The class/stage of chronic renal disease will be predicted using all predictor variables.

Hardware Requirements

The hardware used for this study is consisted of intel® core™ i5, CPU 2.40GHz, RAM 4 GB, 64-bit operating system (x-64 based processor).

Glomerular filtration rate (GFR)

A substance's rate of clearance from plasma is estimated to determine the GFR, which is defined as the volume of plasma filtered by glomeruli per unit of time. It is regarded as one of the finest characteristics to gauge renal function and gauge the severity of CKD [3]. Filtration markers, a kidney-excreted material, are used to compute the GFR value. The GFR is then calculated using a formula that incorporates the clearance of filtration marker. There are many mathematical formulas that can be used to estimate GFR, but the following are the most often used ones:

a. Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) Equation.
b. Modification of Diet in Renal Disease (MDRD) Equation.

CKD-EPI equation The equation for CKD-EPI is written as follow [9]:

$$GFR = 141 \times \min(Scr/\kappa, 1)\alpha \times \max(Scr/\kappa, 1)\text{-}1.209 \times 0.993Age \times 1.018 \text{ [if female]} \_ 1.159 \text{ [if black]}$$

SCr in eq. 1, represents the serum creatinine and k is constant, it stands for Kappa. There are different values of k for male and female, i.e. k = 0.7 for female and k = 0.9 is for male.

MDRD equation The equation for MDRD is written as follow:

$$GFR = 175 * SCr^{-1.154} * age^{-0.203} * 0.742 \text{ (if female)}$$

The modification of diet in renal disease (MDRD) is thought to be less accurate than the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) for the calculation of the glomerular filtration rate (eGFR) [10]. As a result, in the proposed work, we will calculate GFR using the CKD-EPI equation. To determine the matching person's GFR, the equation (CKD-EPI) requires four inputs: sex, race, serum creatinine, and age.

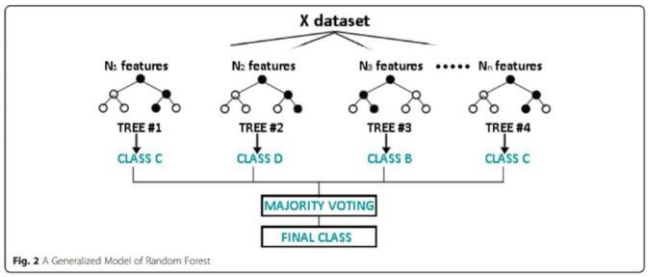

Fig. 2 A Generalized Model of Random Forest

Computation

Our work has incorporated a computational engine using the WEKA data mining tool [11]. Execution time and classification accuracy are used as performance indicators when comparing classification algorithms. The 15-fold cross validation technique has been used to test and validate the model. Finally, the classification's performance evaluation is completed.

Classification of algorithms

Binary/ binomial classification

The challenge for this kind of categorization consists of two values for the class variable. The algorithms foretell one of these from the two classes that are provided. i.e., whether a sickness exists or not, whether a match can be found, etc.

Multiclass/ multinomial classification

When there are [0 to K-1] classes or labels, this type of categorization is employed to solve the problem. The classifier makes a prediction from the provided K-1 classes for one of these. Multiclass J48 and Random Forest classifiers are employed in this study to categorise CKD into various phases. The next subsections provide a description of both algorithms and how they relate to one another.

J48 algorithm

The most popular decision tree method, J48 (C4.5), which is an evolution of Quinlan's previous ID3 Algorithm and is known to have a respectable accuracy rate in bio-medical applications, is a decision tree algorithm. Both numerical and categorical data can be handled by it [14]. The statistical classifier is another name for it [15]. It handles both noise and missing values and is simple to implement [16]. Additionally, J48's performance is subpar for a tiny training set [16]. The J48 algorithm, which was employed in this investigation, generates output based on the following steps.

1. Choose the dataset as an input to the rule for process. To split categorical attributes, J48 works just as the ID3 algorithm.
2. Calculate the Normalized information gain for each feature.
3. The feature with the maximum information gain is chosen as the best attribute. An attribute with the maximum information gain is selected as the root node to create a decision tree.
4. Repeat the above-mentioned step until some stop criterion, to compute the information gain for each attribute and add that attribute as children node.

Random Forest algorithm

An method used for supervised classification is called Random Forest. To efficiently calculate the accuracy, a vast forest of trees is created [18]. The amount of trees used in this classifier directly affects how accurate it is. Because of its adaptability, Random Forest produces findings that are more dependable even without hyper-parameter adjustment. It is easy to use and effective, especially when dealing with big data sets. The accuracy rate is kept by identifying outliers and abnormalities. However, it requires expensive calculation and is not particularly simple to implement.

The working of Random Forest algorithm, used in this study, is based on the following steps to generate output:

1. Select samples randomly from the original dataset. Such kind of randomly selected samples are usually referred to as the bootstrapped data set.
2. Build a decision tree for the bootstrapped data set by considering a random subset of variables.
3. Repeat the above process 100 times (to the largest extent possible).
4. Predict the outcome for new data point by running the new data down all decision trees that are made.
5. The predicted class is judged based on the majority of votes.
6. Finally, evaluate the model by using the out of bag instances of the dataset to derive final class.

Cross validation

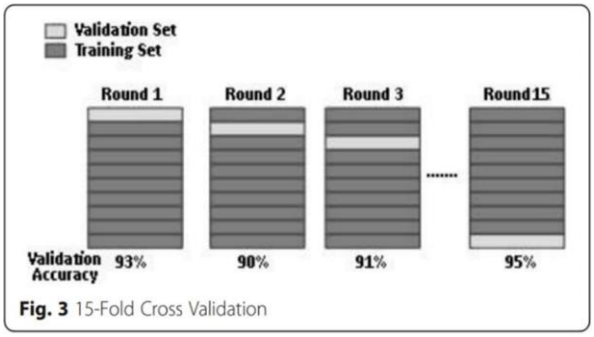

Fig. 3 15-Fold Cross Validation

This technique divides the data set into a number of k-folds for model validation (one test other training). The model constructed from other elements is tested using a one-fold. Building and testing the model are performed for each fold. The average of all k-test errors is then determined. The performance of the model on the dataset is estimated in this study using 15-fold cross validation. Fig. 3 depicts the overall process of 15-fold cross validation. In Figure 3, the entire dataset is first randomly mixed before being divided into 15 groups. One group is used as the test dataset for each group, while the remaining groups are used as the training dataset. On the training set, the model is fitted, and evaluated on the test set. Evaluation scores are retained as 93% in Round 1, 90% in Round 2 and till 95% in round 15.

Performance evaluation of classification

The following mathematical connections are used to calculate the accuracy, sensitivity, specificity, F-Measure, and confusion matrix in order to assess the performance of classification.

Accuracy

Accuracy is one of the most widely used classification performance metrics. It is the proportion of samples that were correctly categorised to all samples. The study's formula for calculating accuracy reads as follows:

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

Where, TP represents true positive values, TN represents true negative values, FP represents false positive values and FN represents false negative values.

Sensitivity

It is also known as hit rate, recall, and true positive rate (TPR). It displays the proportion of accurately categorised positive instances to all positive instances. In this investigation, sensitivity was calculated using the following formula.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity

It is also called True Negative Rate (TNR) or inverse recall. It measures the percentage of correctly classified negative instances to the total number of negative instances. The formula to calculate specificity, used in this study, is written as follows.

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

F-measure

F-Measure is calculated by taking the weighted average of sensitivity and precision values. The formula to calculate F Measure, used in this study, is written as follows.

$$F\text{−Measure} = \frac{2*\text{sensitivity}*\text{precision}}{\text{sensitivity} + \text{precision}}$$

F-Measure uses the field of information retrieval for the estimation of classification performance.

Precision

Precision is defined as what proportion of positive identifications was actually correct. The formula to calculate precision, used in this study, is written as follows.

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

	True Class			
Predictive Class		A	B	C
	A	TPA	EBA	ECA
	B	EAB	TPB	ECB
	C	EAC	EBC	TPC

Table 3 Confusion Matrix for Multi-Class Classification

Precision

Precision is defined as what proportion of positive identifications was actually correct. The formula to calculate precision, used in this study, is written as follows.

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

Confusion matrix

A model's predictions are tabulated and shown in the confusion matrix. Numerous both accurate and inaccurate forecasts are displayed. These are determined by comparing the n-test data with the classification findings. The matrix is represented as an x-by-x matrix, where x is the number of classes in the dataset. Confusion matrix is a powerful technique for determining a classifier's accuracy [10]. TPA is shown as the true positive values in Table 3, which indicates that they anticipated values successfully as real positive values in class A. According to TPB, the anticipated values for class B were successfully identified as real positive values. TPC stands for "true positive values," which signifies that in class C, projected values were identified as real positive values. EAB stands for class A samples that were mistakenly labelled as class B. EAC stands for the class A samples that were mistakenly categorized as C. EBA samples are class B specimens that were mistakenly categorized as A. EBC stands for the class B samples that were mistakenly categorized as C. ECA are class C samples that were mistakenly labelled as class A. ECB refers to class C samples that were mistakenly categorized as B.

Discussion

Chronic diseases linked to kidney failure are referred to as chronic kidney disease (CKD). Blood and urine tests have historically been used to assess how well the kidneys are working. To detect CKD in its early stages and its symptoms, it is crucial to create a CKD screening system. so that the disease can be treated at an early stage and complications avoided by taking preventive measures. When relevant data is provided, machine learning (ML) algorithms can be utilised to create conclusions that are rational and correct. Studies have been done to identify CKD using a variety of factors, such as age, sex, estimated GFR, serum calcium, etc. In their study, S. Ramya et al. employed the R language's radial basis function to predict CKD [6]. They made advantage of patient medical record and obtained as an input dataset from several laboratories.

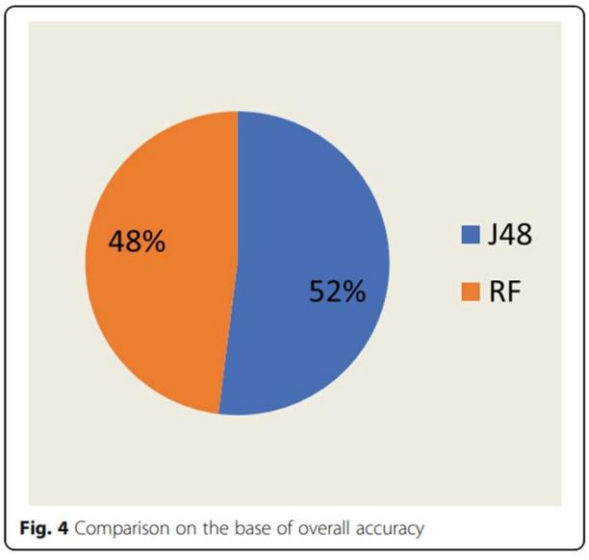

Fig. 4 Comparison on the base of overall accuracy

Table 4 : Detailed Information of Various Studies

Resources of Data Set	Disease	Tool	Accuracy	Execution Time in seconds
Medical reports of patients collected from different laboratories	Chronic Kidney Disease	R	85.3%.	N/A
Medical record of patients in Shanghai Huadong Hospital	Chronic Kidney Disease	online tool	82%	N/A
University of California Irvine (UCI) Machine Learning Repository	Chronic Kidney Disease	DTREG Predictive Modeling System	96.7%	12

Their research found 85.3% accuracy in CKD detection. Jing Xiao carried out a study in 2019 to identify different CKD phases [7]. The model in this study was trained using the logistic regression machine learning technique, and predictions were made using an online application. The authors also used the Shanghai Huadong Hospital's patient medical records as an input dataset. This study has an 85% success rate in identifying CKD. Later, in 2019, El-Houssainy et al. [8] trained the model with the DTREG predictive modelling system utilising data from the UCI repository. Using a probabilistic neural network, they revealed the results with 96.7% accuracy in just 12 seconds. Table 13 provides further information on the aforementioned experiments, and Fig. 5 displays a graph of accuracy. Within 0.03 seconds, this study's accuracy was 85.5%. Even though the PNN in Fig. 5 has a lower performance efficiency, its time efficiency is higher. The accuracy of ML algorithms typically increases when a big amount of data is provided. Although we employed a limited dataset for this study, the sample size was sufficient for the analysis, which led us to the conclusion that the J48 algorithm outperformed the random forest algorithm. It is anticipated that J48 will perform better than PNN if a large dataset is used. Our research demonstrates that CKD phases may be predicted and classified using ML classification approaches with reasonable accuracy and in a shorter amount of time.

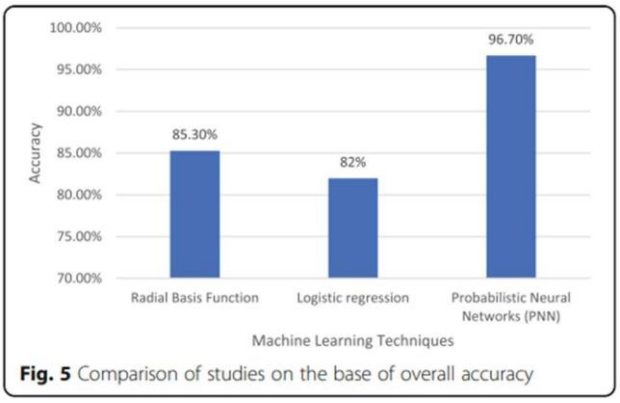

Fig. 5 Comparison of studies on the base of overall accuracy

comparable to the research displayed in Table 4. Results from Tables 1, 2, and 3 demonstrate that J48 is superior to Random Forest in terms of accuracy rate, precision, and F-Measure for grading the severity of CKD into stages.

Conclusion

In order to predict the different stages of CKD, we developed and compared two algorithms, namely J48 and random forest. The ratio of correctly categorized examples using J48 is observed to be 85.5%, compared to 78.25% for Random Forest. In contrast, J48 takes 0.03 s and Random forest 0.28 s to complete the task. Since J48 delivers results with greater accuracy and in less time than Random Forest, it can be claimed that it is accurate and efficient in terms of execution time. Because J48 handles both categorical and continuous values, but Random forest favors characteristics with categorical values, it outperforms Random forest in terms of performance. Multiple decision trees are constructed by Random Forest, which then combines them to create a reliable prediction model. However, this method makes the system sluggish and useless for real-time prediction. J48 is simple to implement, but Random forest is challenging due to the vast amount of trees. Therefore, based on our findings, we advise doctors to create an automated decision support system for detecting CKD utilizing j48.

References

1. Webster AC, Nagler EV, Morton RL, Masson P. Chronic kidney disease.Lancet. 2016;6736(16):1–15.
2. Serpen AA. Diagnosis rule extraction from patient data for chronic kidneydisease using machine learning. Int J Biomed Clin Eng. 2016;5(2):64–72. https://doi.org/10.4018/IJBCE.2016070105.
3. Tekale S, Shingavi P, Wandhekar S. Prediction of chronic kidney disease using machine learning algorithm. Ijarcce. 2018;7(10):92–6. https://doi.org/10.17148/IJARCCE.2018.71021.
4. Ponum M, Hasan O, Khan S. EasyDetectDisease: an android app for earlysymptom detection and prevention of childhood infectious diseases. Interact J Med Res. 2019;8(2):e12664. https://doi.org/10.2196/12664.
5. Hill NR, Fatoba ST, Oke JL, Hirst JA, O'Callaghan CA, Lasserson DS et. al. (2016) Global prevalence of chronic kidney disease—a systematic review and meta-analysis. PLoS One 11:e0158765, 7, DOI: https://doi.org/10.1371/ journal.pone.0158765.
6. Ramya S, Radha N. Diagnosis of chronic kidney disease using machine learning algorithms. Int J Innovative Res Comput Commun Eng. 2016;4(1):812–20.
7. Xiao J, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med. 2019;17(1):1–13.
8. E. H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," Inform Med. Unlocked, vol. 15, no. April, p. 100178,2019.
9. Teo BW, Xu H, Wang D, Li J, Sinha AK, Shuter B, et al. GFR estimating equations in a multiethnic asian population. Am J Kidney Dis. 2011;58(1):56–63. https://doi.org/10.1053/j.ajkd.2011.02.393.
10. Stevens LA, Claybon MA, Schmid CH, Chen J, Horio M, Imai E, et al. Evaluation of the chronic kidney disease epidemiology collaboration equation for estimating the glomerular filtration rate in multiple ethnicities.Kidney Int. 2011;79(5):555–62. https://doi.org/10.1038/ki.2010.462.
11. Swathi Baby P, Panduranga Vital T. Statistical analysis and predicting kidney diseases using machine learning algorithms. Int J Eng Res. 2015;V4(07):206–10.
12. Ani R, Sasi G, Sankar UR, Deepa OS. "Decision support system for diagnosisand prediction of chronic renal failure using random subspace classification," 2016. Int Conf Adv Comput Commun Inform. 2016;2016: 1287–92.
13. C4.5 Algorithm. Available at: https://en.wikipedia.org/wiki/C4.5_algorithm.
14. Saad Y, Awad A, Alakel W, Doss W, Awad T, Mabrouk M. Data mining of routine laboratory tests can predict liver disease progression in Egyptian diabetic patients with hepatitis C virus (G4) infection: a cohort study of 71 806 patients. Eur J Gastroenterol Hepatol. 2018;30(2):201–6. https://doi.org/10.1097/MEG.0000000000001008.
15. V. Kumar and L. Velide, "A data mining approach for prediction and treatmentSupervised machine learning algorithm:" vol. 3, no. 1, pp. 73–79, 2014.
16. B. Gupta, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining," vol. 163, no. 8, pp. 15–19, 2017.
17. Tabassum BG, Mamatha B, Majumdar J. "Analysis and Prediction of ChronicKidney Disease using Data Mining Techniques". Int J Eng Res Comput Sci Eng. 2017. https://doi.org/10.13140/RG.2.2.26856.72965.
18. Gupta DL, Malviya AK, Singh S. Performance analysis of classification tree learning algorithms. Int J Comput Appl. 2012;55(6):39–44. https://doi.org/10.5120/8762-2680.
19. Beeravalli V. "Comparison of Machine Learning Classification Models for Credit Card Default Data", Medium.com. 2018. Available at: https://medium.com/@vijaya.beeravalli/comparison-of-machine-learning-classification- models-for-credit-card-default-data-c3cf805c9a5a.
20. Lateef Z. "A Comprehensive Guide to Random Forest in R", Edureka.co. 2020. Available at: https://www.edureka.co/blog/random-forest-classifier/.
21. Jena L, Kamila NK. Distributed data mining classification algorithms for prediction of chronic-kidney-disease. Int J Emerg Res Manag Technol. 2015;9359(11):110–8.