International Journal of Informatics, Information System and Computer Engineering

Risks of Chronic Kidney Disease Prediction using various Data Mining Algorithms

Akalya Devi C, Fatima Abdul Jabbar**, Kavi Varshini S***, Kriti S Rithanya****, Miruthubashini M*****, Naveena K S*******

*Assistant Professor, 2UG Scholar, Department of Information Technology, PSG College of Technology, Coimbatore, India.

*Corresponding Email: [1]akalya.jk@gmail.com

ABSTRACTS

Twenty million people have chronic kidney disease where patients experience a gradual deterioration of kidney function, the result of which is kidney failure. Early detection of chronic renal disease can help to slow its progression, avert complications, and reduce the risk of cardiovascular complications. Data mining has been broadly used in order to support medical professionals and physicians in the prediction and examination. Here, in this paper, multiple data mining algorithms are used to solve a problem in the field of medical diagnosis and examine how effective they were at predicting the consequences. The study's focus was on the diagnosis of chronic renal disease. This dataset used for this study consists 400 instances & 25 attributes. Preprocessing of the large amount of raw data is carried out to impute any missing data and determine which of the variables should be taken into account in the prediction models. The accuracy of the prediction is used to compare and contrast the various predictive analytic models.

ARTICLE INFO

Article History:
Received 18 Dec 2021
Revised 20 Dec 2021
Accepted 25 Dec 2021
Available online 26 Dec 2021

Keywords:
Chronic kidney disease, K-Nearest Neighbor Classification, predictive analytics, Decision Tree, data mining, Support Vector Machine, Random Forest.

1. INTRODUCTION

Chronic Kidney Disease (CKD) or Chronic Kidney Failure is the increasing impairment of the kidney's ability to function normally. Chronic Kidney Disease is induced primarily due to high blood pressure, diabetes, hypertension, and several other factors in particular smoking, obesity, heart disease, heredity, consumption of alcohol, usage of drugs, age, race, ethnicity, etc. In India and other developing nations, chronic diseases remain a leading cause of mortality. The number of casualties in India owing to chronic disease was anticipated to be 5.21 million in 2008 and is likely to increase from over 7.63 million by 2020. There are five distinct stages of disease development in which each stage increases in severity while as it advances between stage 1 and stage 5. Stage 1 is when a person's kidney function falls below normal. As the affected individuals it goes ahead into step 2, they may experience a mild to moderate loss of renal functions. The worsening condition escalates in level 3, where there is a moderate to average deterioration in the nephrological operation followed by acute damage in the functioning of the excretory system in stage 4. Stage 5 is the absolute collapsing of the urinary organs. (Almustafa, 2021).

The massive increase in the amount of medical data available to predict the disease has raised the question of being effectively classified, managed, and transferred. To extract useful insight and knowledge from this raw data, effective ways are required. Data mining techniques are a dependable and pragmatic way of accomplishing this. Data mining is the process for processing massive amounts of data and extracting knowledge from all of this. In addition to the medical sector, the data are sequentially organized and are exploited in multiple number of real-time applications such as social networking sites, online websites, and so on. Data mining is categorized in many other domains including graphic data extraction, web data mining, textual data mining, image data extraction domain. These data mining sectors facilitate in decision-making and the extraction of useful information from the dataset undergoing investigation.

Prediction of the risk of chronic kidney disease is based on several health parameters including random blood glucose level, blood pressure, serum creatinine level, and others. Supervised classification algorithms which are used to predict the risk of chronic kidney disease are Decision Tree Classification, Support Vector Machine Classification (SVM), Random Forest Classification, and K-Nearest Neighbor Classification (KNN) (Aqlan, et al., 2017). From experimenting, Random Forest Classification and KNN were shown to be the best classifiers for classification. Random Forest and KNN classifications have maximum reliability than Decision Tree and SVM classifiers.

2. LITERATUR REVIEW

In this research paper, recent data mining procedures were used to classify and forecast chronic kidney disease which considers various influencing factors such as blood pressure, red blood cells count, haemoglobin, etc. The techniques used in this paper provide more accuracy than the techniques used in other existing works.

Kaur G et al. applied two data mining classifiers to predict chronic kidney disease: KNN and SVM, which gave the exactitude and error percentage (Arasu, D., & Thirumalaiselvi, R. 2017).

Bhatla N et al. Has analysed most of the dangerous diseases among which breast cancer, heart disease, and diabetes are the predominant ones (Bhatla, N., & Jyoti, K. 2012). On investigating 168 articles the techniques for implementing the diagnosis of various diseases have been performed. All techniques, data mining approaches, and evaluation methodologies are carefully investigated and properly considered.

Kunwar V et al. Using categorization approaches such as Naive Bayes and Artificial Neural Networks (ANN), authors hypothesize chronic renal disease. According to the RapidMiner tool's trial results, Naive Bayes generates further accurate outcomes than ANN. (Gharibdousti, M. S et al., 2017). Decision Tree, Linear Regression, Super Vector Machine, Naive Bayesian, and Artificial Neural Networks (ANNs) were one of the classification strategies utilized (Ilyas, et al., 2021).

The correlation matrix was used to investigate the features' correlation. As a result, they observed the influence of properties on classification findings.

Padmanaban K. A et al. On the incurable renal disease dataset, researchers implemented data extraction algorithms such as Naive Bayes and the Decision tree algorithm (Padmanaban, K. A., & Parthiban, G. (2016). On comparing and contrasting several categorization algorithms, they recommended decision tree classification to reach substantial results with suitable accuracy by estimating its performance to its specificity and sensitivity (Kunwar, et al., 2021).

Sharma S et al. Evaluated 12 data mining clustering techniques by implementing them to the CKD dataset (sharma, et al., 2016). To determine efficiency, the findings of the prediction were contrasted with the factual medical outcomes. A few of the metrics used to evaluate performance comprise predictive accuracy, precision, sensitivity, and specificity (Kunwar, et al., 2016). With an accuracy at about 98.6%, a sensitivity of 0.9720, a precision of 1, and a specificity of 1, the decision tree showed the best performance.

Arasu D et al. Employed significant data extraction methods in particular clustering, classification, association analysis, and regression to predict renal diseases (Milley, A. 2000). These techniques had insignificant shortcomings in the picturality of preprocessing or at any other stages. Various data mining techniques are evaluated and the major problems are briefly explained.

Vijayarani S et al has focused on using a novel machine learning classification strategy to predict chronic renal illness employing SVM on a data sample of 400 observations and 24 attributes (Vijayarani, et al., 2015).

3. PROPOSED WORK

1. Due to CKD millions of individuals pass on each year since they don't experience legitimate treatment. CKD risk factors fall under four main categories: Susceptibility components which lead to a rise in renal damage susceptibility,

2. The terminology "initiation factors" refers to the elements that play a key role in renal damage.

3. Progression Factors leads to more regrettable reality of kidney harm and fast decay functionalities once the harm gets begun.

4. Kidney failure occurs as a result of end-stage conditions, culminating in morbidity and mortality.

Kidney illnesses are anticipated and compared utilizing SVM and ANN algorithm stationed on the exactness and performance time. SVM, KNN, and some other algorithms have been used to assess the performance of the CKD dataset from the UCI repository and the raw data which have been taken was cleaned and processed by various steps which have been explained in the figure 1.

Four different classifiers have been analyzed majorly established on the succeeding approaches: Decision-Tree, Support Vector Machine (SVM), Random Forest, K-nearest neighbor (KNN) in Section 3.

These technics were picked for the examination and review for the reason that of their ubiquity within the later important writing. A concise portrayal about the chosen strategies has been given underneath.

3.1. Data Mining Algorithms & Technuques

An algorithmic data mining program can be a well-specified plan of action that takes data as in and out. It includes designs in the shape of models. It comprises a small number of algorithms and strategies namely classification, grouping, prediction, association rules, neuronal networks, etc., to perform knowledge revelation from data banks. Table 1 shows the evaluation plans employed here.

Table 1. Classification of CKD & Evaluation plans

S. No	Phases of CKD	GFR (Glomerular Filtration Rate)	Evaluation Strategies
1.	Nephrological damage with common GFR	90 or beyond	Treating the coexisting conditions, reduction of hazard variables for cardiac and vascular illness
2.	Renal impairment with moderate reduction	60-89	Approximation of ailment advancement
3.	Reasonable reduction	30-59	Assessment and medication of sickness intricacies

4.	Rigorous diminution	15-29	Formulation of excretory organs switching remedy (dialysis, granting)
5.	Renal failure	Less than 15	Nephrological organs grafting therapy

The prediction analytics conducted is based upon the typically picked data columns of data, which comprises of the age, blood pressure, number of red blood cells, and appetite fields. These above mentioned four entries incorporate the numeric data in the case of blood pressure and age, while categorical data for the number RBGs and appetite. The nominal data has indeed been converted into numeric types so as to –make classification techniques suitable to string-based categorical attributes, which cannot be handled using statistical models. The proposed framework for the study is illustrated in Figure 1.

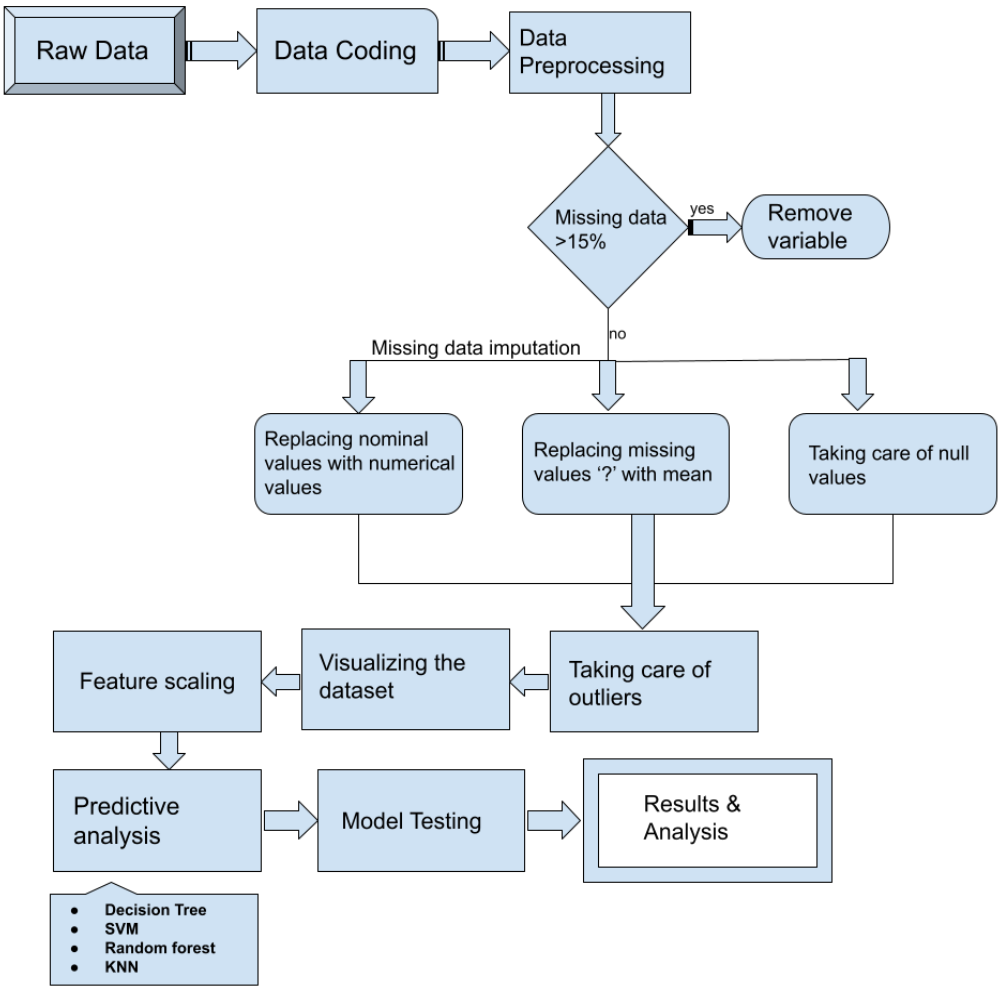

Figure 1. Proposed framework for CKD analysis and prediction

Here are the basic steps which were performed initially;

1. Acquire the data from the local disk.
2. With the help of the column identifier IDs, manually choose the columns.
3. To make all the nominal values numerical, the conversion is made.
4. After the categorical transformation, make the last data matrix.
5. Inside the last data matrix, search for the missing values.
6. Compute the average of every column that constitutes the variable.
7. Load in the missing values with the appropriate average value from the mean values.
8. To make a non-uniform feature matrix, shuffle the data matrix.
9. Divide the training and testing data matrices.
10. Make the observation vectors ready for training and testing.

3.2. Classification

The best and most common data mining approach is classification. Where entities are classified into different categories called classes and assigned to them. Each and every thing needs to be distributed precisely to one class and not more than one and never to no classes at all. Decision tree, SVM, KNN, and Random Forest were the classification algorithms included in this model.

3.2.1. Decision Tree Classification

This method is especially beneficial for deciphering classification problems in which a tree is formed to depict the categorization process. The tree is linked to every tuple in the database to yield classification as long as it is established. Classification tree analysis and regression tree analysis are the two forms of decision trees used in data mining, and they have been employed for a spectrum of potential results such as belonging to a specific statistical class or an actual number.

1. Fitting Decision Tree to the training set.
2. Predicting the test result.
3. Calculating the accuracy.
4. Displaying the confusion matrix.

3.2.2 SVM Classification

SVM is a set of rules for supervised machine learning that can be used to resolve classification and regression problems. It uses a strategy called the kernel trick to convert your data and after that based on these changes it finds an ideal boundary between the possible outputs (Sinha, P., & Sinha, P. 2015). The following steps are the ones performed;

1. Support Vector Machine (a classification technique) is applied on the available data for the purpose of predictive analysis.
2. Using the training data matrix and the training observation vector the classifier is trained.
3. The testing data matrix with unseen data is utilized to examine the classifier

4. The predictions (Observations predicted by SVM classifier) are returned as output.

The entire performance is computed by comparing and contrasting the outcomes of support vector machine classifier and the actual perceptions.

3.2.3 Random – Forest Classifier

Random Forest is an analyzer that equips the average of a number of decision trees on discrete subsections of a given set of data to advance the dataset's predicted accuracy. The following steps are the ones performed (Sinha, P., & Sinha, P. 2015).

1. Fitting Random Forest classification to the training set.
2. Predicting the test result.
3. Calculating the accuracy.
4. Displaying the confusion matrix.

3.2.4 KNN Classifier

It's a type of distance-based technique that's typically used while the values of each and every attribute is uninterrupted and continual, but it can also be used with nominal features (Subasi, et al., 2017). To compute the categorization of an unknown sample data based on the classification of the closest instance or instances. More occurrences inside the preparation set use the same way to group the k-nearest neighbors (also known as k-nearest Neighbor), (Vijayarani, et al., 2015). The steps that were taken were as follows:

1. K-Nearest Neighbor (one of the classification technics) is employed over the given data for the purpose of predictive analytics.
2. Before initiating the entire process, the value of k should be initialized which will be symbolizing the number of neighbors that has to be considered.
3. The k-nearest neighbor classifier needs to be trained with the specified k value over the training data matrix and the training observation vector.
4. With the help of the test data matrix, which contains the unseen data the classifier is tested and evaluated for the required metrics.
5. The forecasts (observations predicted by the KNN classifier) made by the KNN analyzer should be returned.
6. The entire accuracy and performance of the KNN classifier is estimated by comparing the predictions made by KNN and the actual observations.

5. RESULT AND DISSCUSION

The chronic kidney disease (CKD) dataset was acquired based on the UCI machine learning repository and is employed in this study for prediction and validation. Both numerical and nominal attributes were included in CKD dataset. There are 25 attributes and 400 instances. This dataset also contains missing values. There are 24 attributes and one class attribute (i.e.) CKD, NOT-CKD. Table 2 gives the attribute description of the dataset.

Table 2. CKD Dataset Attributes description

S. No	Attribute Name	Expansion	S. No	Attribute Name	Expansion
1	age	Age of the patient	13	pot	Potassium
2	bp	Blood pressure	14	hemo	Hemoglobin
3	sg	Specific gravity	15	pcv	Packed cell volume
4	al	Albumin	16	wc	White blood cell count
5	su	Sugar	17	rc	Red blood cell count
6	rbc	Red blood cells	18	htn	Hypertension
7	pc	Pus cell	19	dm	Diabetes mellitus
8	pcc	Pus cell clumps	20	cad	Coronary artery disease
9	ba	Bacteria	21	appet	Appetite
10	bgr	Blood glucose random	22	pe	Pedal edema
11	bu	Blood urea	23	ane	Anemia
12	sc	Serum creatinine	24	sod	Sodium

Data Cleaning and data pre-processing is the most critical point in the data mining procedure as it influences the rate of success drastically. The categorical attributes were displaced with 0s and 1s corresponding to their values. The missing values were replaced with the mean of that particular attribute. As there was a wide range of age, the age attribute was grouped in batches (Sharma, et al., 2016). The CKD dataset includes features that vary in the degree of magnitude, range, and units. In order to interpret all the features on the same scale, Feature Scaling (Data Normalization) was carried out.

The CKD dataset was parted into 70% for the purpose training and 30% for the purpose of testing data. Four different data mining procedures encompassing Decision Tree Classification, Support Vector Machine Classification, Random Forest Classification, KNN Classification

were applied to the training and testing data and the performance measurement using different metrics like precision, f1-score, recall, accuracy, specificity, and sensitivity were observed (Vijayarani, et al., 2015). Table 3 presents the different performance metrics used in this paper.

Table 3. Different overall performance analysis metrics used

Metrics	Definition	Equation
Precision	The proportion of predicted accurately positive considerations to fully predicted positive observations is referred as precision.	$\frac{TP}{TP+FP}$
Recall (Sensitivity)	Estimates the percentage of number of yes's that are effectively-recognized correctly.	$\frac{TP}{TP+FN}$
F1-Score	Precision and Recall are weighted averages which determine the F1 score.	$\frac{2*Precision*Recall}{Precision+Recall}$
Accuracy	Measures the model's ability to accurately estimate class label of latest or previously unknown information.	$\frac{TP + TN}{TP + FP + FN + TN}$
Specificity	Here, ratio of negatives (or No's) that have been correctly recognized as such is measured.	$\frac{TN}{TN + FP}$

The performance metrics of the various proposed algorithms were derived using the equations listed in Table 3. Table 4 depicts the results obtained for every algorithm.

Table 4. Performance measures of the proposed algorithms

Model	Precision		Recall		F1-Score		Specificity	Accuracy
	Not-CKD	CKD	Not-CKD	CKD	Not-CKD	CKD		
Decision Tree	0.91	1.00	1.00	0.95	0.95	0.97	1.00	0.967
SVM	0.93	1.00	1.00	0.96	0.97	0.98	1.00	0.975

Model	Precision		Recall		F1-Score		Specificity	Accuracy
	Not-CKD	CKD	Not-CKD	CKD	Not-CKD	CKD		
Random Forest	0.95	0.99	0.98	0.97	0.96	0.98	0.987	0.975
KNN	0.95	1.00	1.00	0.97	0.98	0.99	1.00	0.983

The train score is the measurement that states us in what way the model suits the training data. Similarly, the test score shows how the model reacts to the unknown data. The area under the curve (AUC) score portrays the model's overall performance at differentiating between the positive and negative classes. Figure 2 shows the comparison of the training score, test score and mean AUC score.

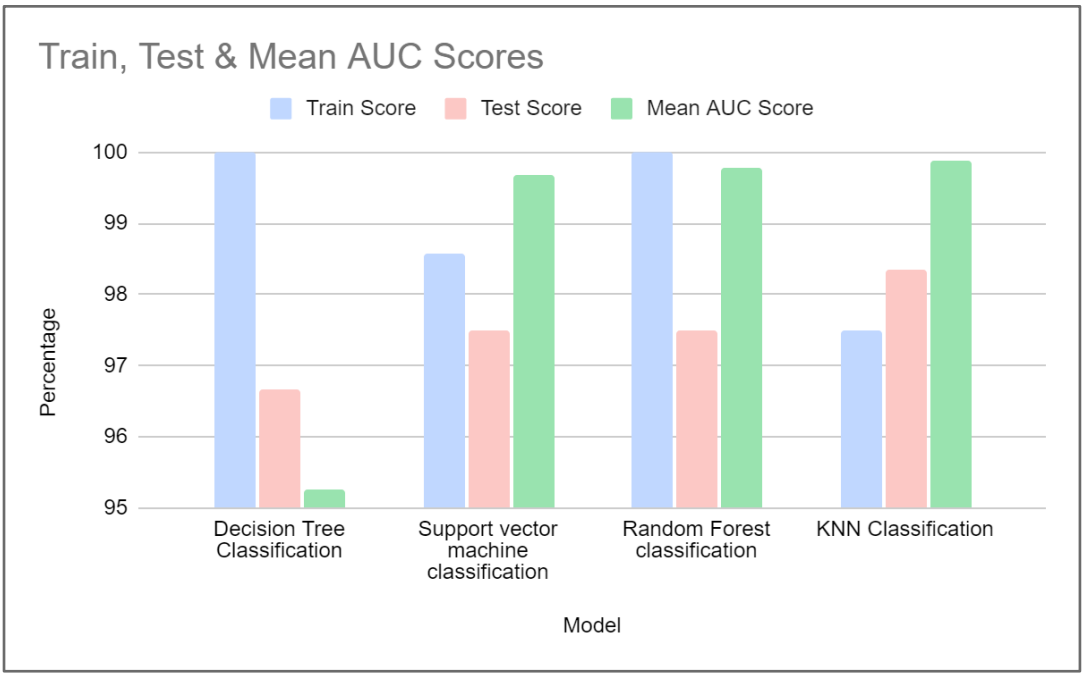

Figure 2. Depiction of Train, Test, and Mean AUC Scores of the proposed algorithms

The difference in magnitude between both the observation's prediction and its true value is termed as the mean absolute error. For the proposed algorithms, Figure 3 illustrates the mean absolute error.

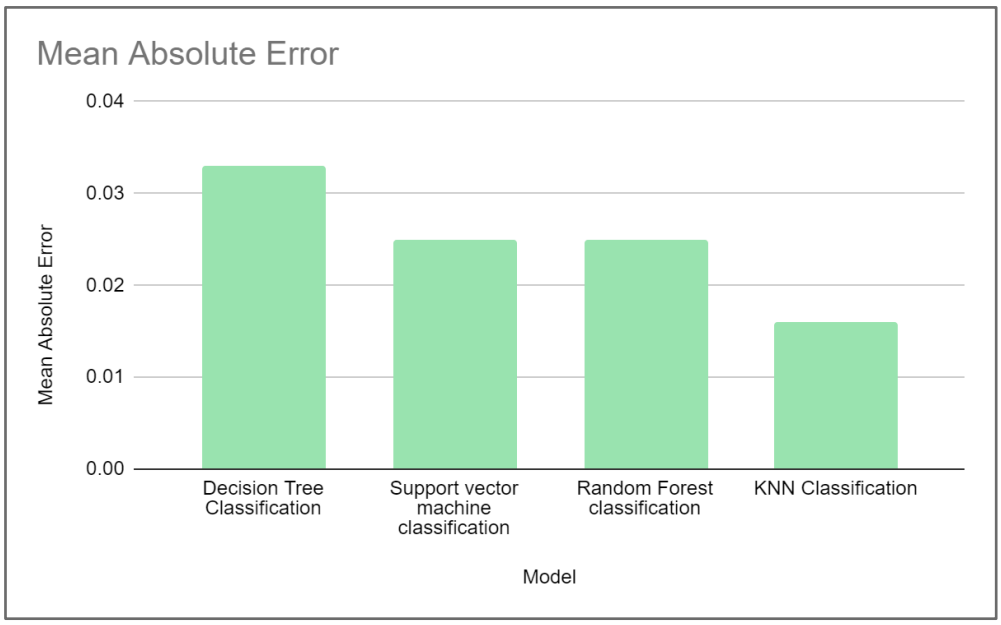

Figure 3. Illustration of the proposed algorithms' Mean Absolute Error

The Receiver operating characteristic Curve (ROC Curve) reflects the classification model's overall performance among all class thresholds [10]. The ROC curve for the algorithms employed in this study is illustrated in Figure 4.

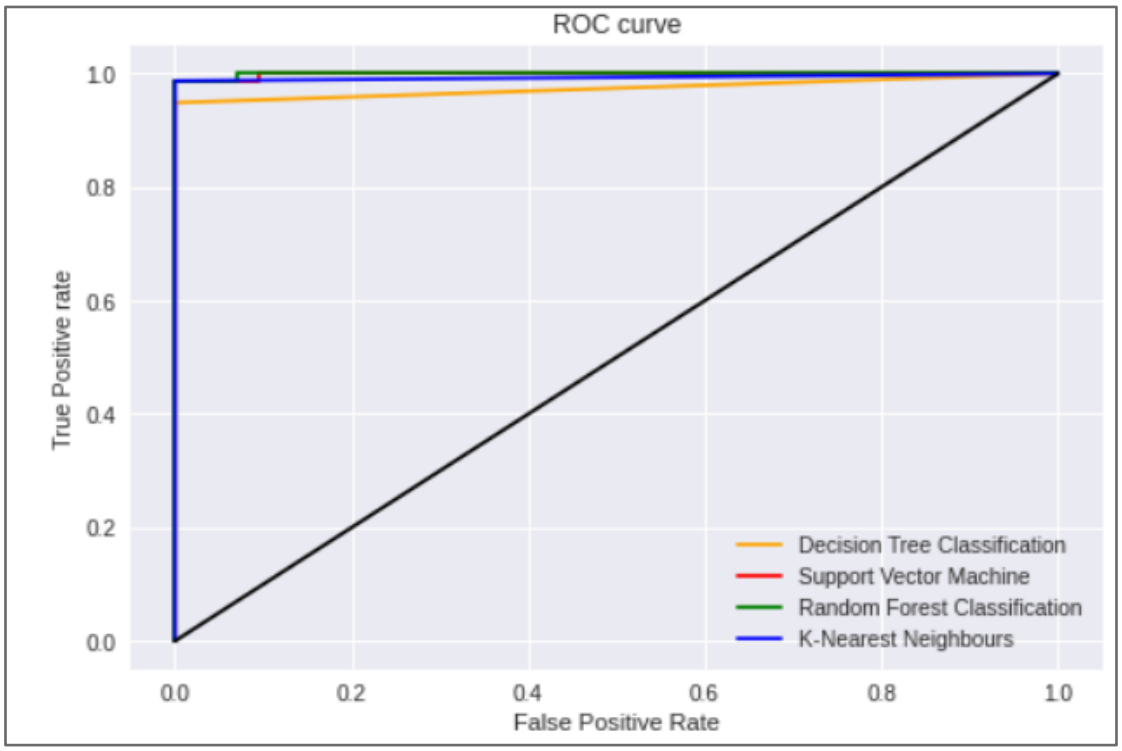

Figure 4. Plot of ROC curve for the proposed algorithms

5. CONCLUSION

The objective of this article is to analyze the variety of data mining techniques and algorithms utilized to predict Chronic Renal Disease. CKD has been predicted and diagnosed using data

mining classifiers: Decision Tree, SVM, Random Forest, and KNN. It was found that KNN results in the best accuracy. The performance of the KNN method was found to be 98.3% accurate compared to Decision Tree (96.7%), SVM (97.5%), and Random Forest (97.5%). The work can further be extended keeping into consideration the other parameters like food intake, living conditions like sanitation, availability of clean water, working environment, environmental factors like pollution, etc. for the detection of kidney disease. Further experimentation can be conducted using other classifiers like ANN or by using ensemble techniques.

REFERENCES

Almustafa, K. M. (2021). Prediction of chronic kidney disease using different classification algorithms. Informatics in Medicine Unlocked, 100631.Dobrucka, R. (2018). Synthesis of MgO nanoparticles using Artemisia abrotanum herba extract and their antioxidant and photocatalytic properties. *Iranian Journal of Science and Technology, Transactions A: Science*, *42*(2), pp. 547-555.

Aqlan, F., Markle, R., & Shamsan, A. (2017). Data mining for chronic kidney disease prediction. *In IIE Annual Conference. Proceedings* (pp. 1789-1794). Institute of Industrial and Systems Engineers (IISE).

Arasu, D., & Thirumalaiselvi, R. (2017). Review of chronic kidney disease based on data mining techniques. *International Journal of Applied Engineering Research,* 12(23), 13498-13505.

Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering,* 1(8), 1-4

Gharibdousti, M. S., Azimi, K., Hathikal, S., & Won, D. H. (2017). Prediction of chronic kidney disease using data mining techniques. *In IIE Annual Conference. Proceedings (pp. 2135-2140).* Institute of Industrial and Systems Engineers (IISE).

Ilyas, H., Ali, S., Ponum, M., Hasan, O., Mahmood, M. T., Iftikhar, M., & Malik, M. H. (2021). Chronic kidney disease diagnosis using decision tree algorithms. *BMC nephrology*, 22(1), 1-11.

Ilyas, H., Ali, S., Ponum, M., Hasan, O., Mahmood, M. T., Iftikhar, M., & Malik, M. H. (2021). Chronic kidney disease diagnosis using decision tree algorithms. *BMC nephrology*, 22(1), 1-11.

Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016, January). Chronic Kidney Disease analysis using data mining classification techniques. *In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)* (pp. 300-305). IEEE.

Milley, A. (2000). Healthcare and data mining. *Health Management Technology, 21*(8), 44-45.

Padmanaban, K. A., & Parthiban, G. (2016). Applying machine learning techniques for predicting the risk of chronic kidney disease. *Indian Journal of Science and Technology, 9*(29), 1-6.

Sharma, S., Sharma, V., & Sharma, A. (2016). Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. *arXiv preprint arXiv:*1606.09581.

Sinha, P., & Sinha, P. (2015). Comparative study of chronic kidney disease prediction using KNN and SVM. *International Journal of Engineering Research and Technology, 4*(12), 608-12.

Subasi, A., Alickovic, E., & Kevric, J. (2017). Diagnosis of chronic kidney disease by using random forest. *In CMBEBIH 2017* (pp. 589-594). Springer, Singapore.

Rubini, L. J., & Eswaran, P. (2015). UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set.

Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR), 6*(2), 1-12.