A comprehensive review on Machine Learning algorithms for prediction and analysis of Chronic Kidney Disease

Kavitha M[1], Saroja S Bhusare[2] ,Ananya R[3], Jayashree P[4], L Ashwini[5], Niharika K[6]
JSS Academy of Technical Education, Bangalore

Abstract

Due to busy life and hectic time schedule, less attention is given to personal health and it is concentrated only when symptoms appear. One of the diseases with a high mortality rate, affecting about 10% of the population is Chronic Kidney Disease (CKD). It occurs when kidney functioning gradually decreases for a relatively longer period. It rarely shows any symptoms during early stages. CKD progression is associated to an increased risk of cardiovascular disease, metabolic bone disease, hyper lipidemia and anaemia, among other significant consequences. In many parts of the world, poor access to renal replacement therapy is an issue for patients who advance to final-stage renal disease. So the early detection of the disease can control and reduce the chance of the disease being fatal to the patients. Hence various machine learning algorithms have been used to anticipate this. From the study it is observed that the feature selection method is employed, which is intended to limit the number of attributes and select only the important ones. To predict if a person has chronic illness or not at the early stages, multiple classification models have been constructed that include Support Vector Machine, Neural Networks, Random Forest and the K-Nearest Neighbour algorithm. The outcomes of these four algorithms are compared and the best algorithm is determined.

Keywords:- Machine Learning (ML), Chronic Kidney Disease (CKD),), Pre-processing, Support Vector Machine, K-Nearest Neighbour ,Neural network, Random Forest.

I INTRODUCTION

The kidney's inability to perform its regular function of blood filtering and gradual decrease in glomerular filtration rate (GFR) is called "Chronic Kidney Disease". This causes an alarming increase in the body's potassium and calcium salts. The presence of elevated levels of these salts causes a variety of other health aliments in the body. CKD can be caused by various reasons which include diabetes, polycystic kidney disease and high blood pressure . The risk of developing CKD is increased, if there is family history of kidney disease. When left untreated, an individual may develop other complications such as impaired physical function high blood pressure, malnutrition, cognitive impairment, anaemia and increased risk

of cardiovascular disease. As a result, it is important to find CKD at the earliest stage, but it is difficult to do so because the symptoms grow slowly and aren't unique to it .Studies state that CKD affects about one in every three people with diabetes. CKD can be predicted using machine learning based on few important attributes such as blood pressure, albumin, specific gravity, anaemia, and diabetes.

Machine learning can be helpful in assessing whether or not a person has CKD since certain patients have no symptoms at all. As machine learning is an evolving field in medicine, it can be used to diagnose various diseases in the health sector. The Glomerular Filtration Rate (GFR) is a test that assesses the kidneys ability to filter waste. It measures the amount of blood that passes through the glomeruli per minute in detail. Depending on the GFR, CKD is classified into several stages as shown in the Table 1.

Stage	**Quality Description**	**GFR(mL/min/1.73m^2)**
1	GFR normal or increased	≥90
2	GFR is mildly reduced	60–89
3a	GFR has slowed significantly	45-59
3b	GFR has slowed significantly	30-44
4	GFR has been severely reduced.	15–29
5	Renal Failure-Condition in which the kidneys have reached the end	<15

Table 1 Stages of chronic kidney disease [1]

GFR= Glomerular Filtration Rate;
mL/min/1.73m^2=millilitres per minute for 1.73 meter square

II RELATED WORKS

Researchers have used algorithms like Random Forest(RF),Decision Tree and Support Vector Machine(SVM) in [2] and [4]. From their analysis,it was observed that decision tree algorithms had a lower level of accuracy than SVM but when compared with Random forest, it gives much better prediction than the other two.In [3], out of 25 attributes, 5 CKD-related attributes were tested. The algorithms used were KNN and Naive Bayes for predicting the patient's CKD status. The KNN classifier correctly predicted chronic kidney disease 100% of the time, while the Naive Bayes Classifier correctly predicted 96.25% of the time.In [5],it was observed that when data mining is used in conjunction with other methods and techniques to diagnose a disease, it produces positive results. The level of chronic renal illness was predicted using the algorithmic classifier. Though the study has many parameters, attributes were limited. CKD can be strongly predicted with the aid of various classifiers in data mining, according to [6], [17] and [20].According to the results of various experiments, Neural network, RF, Naive Bayes, SVM, KNN and Radial Basis Function (RBF) are some of the classifiers that gives higher precision whereas Naive Bayes exhibits maximum accuracy

rate of 98%.Prediction of CKD has become more precise. After analysing ten algorithms in[7], it was found that three of them provided the best results: Decision Tree, Random Forest, and Gaussian Naive Bayesclassifier, all of which gave 100% accuracy rate and mimimal loss. To improve the accuracy level for other algorithms, the dataset should be processed accordingly.

Using image processing in [8], white-to-black transition locations of the renal organ are proportional to the steps of the CKD.Patients with earlier stages of CKD have higher values of indicators of transition from white to black than patients with later stages.Missing values in the chronic renal disease dataset were investigated in [9].The consistency of the model and the predicted results are harmed by missing values in the data collection. So the recalculated values substitute the missing values.According to the findings of the above survey, results can differ depending on the methods and techniques used at different stages of kidney disease diagnosis.In [10], training data set of 40,000 person was collected and information from 40 million people was used to verify the model from 2009 to 2018. Neural network approach was developed that has a 95% accuracy in predicting the likelihood of developing chronic kidney disease.It was illustrated in [11] and [14] that the classification and detection accuracy of mean, mode, and median-based pre-processing techniques with the neural network was significantly encouraging than KNN, SVM, Regression Tree and Classification Tree.It was revealed that ANN improves classification efficiency and produces accurate results, making it the best classifier algorithm among others.

Various machine learning algorithms, such as probabilistic neural networks, multilayer perceptrons, vector support machines, and radial base functions, were predicted to have high accuracy in [12].The probabilistic neural network, in contrast to the multi-layer perceptron, the SVM and the probabilistic neural network, has the highest classification accuracy percentage of 96.7%, according to the report.The proposed work in [13] looks at a variety of algorithms, including the Basic Propagation Neural Network, RBF and RF, to identify the various stages of CKD based on their severity. According to the findings, the RBF algorithm outperforms other classifiers and achieves an accuracy of 85.3% accuracy. For the purpose of disease prediction, two algorithms, KNN and Logistic Regression, were used in [15]. In the predictive model, whether a patient has CKD or not is determined. Furthermore, the patients were categorised based on a variety of factors such as bacteria, appetite, classification, anaemia and other factors. In the prediction of CKD, [16] compared the performance of six classifiers (including SVM, which had previously been reported as the best performer). The proposed method achieved superior prediction performance in terms of classification precision, according to the experimental results. The difficulty of diagnosing renal failure has been studied in [18], especially when acute renal failure is taken into account. Neural networks and deep learning were used as a classifier and for assessing their findings as a result of their study, i.e., the patient is classified as having acute or chronic renal insufficiency and is stable, then the patient is notified of the next steps.

In [19], classifiers are trained, tested, and validated using 10 cross-validations. Superior performance was achieved with the F1 gradient measurement algorithm (99.1%), sensitivity (98.8%), and specificity (99.3%). Therefore, it was concluded that CKD can only be detected with three characteristics. Furthermore, haemoglobin was discovered to have the greatest contribution to CKD identification, while albumin had the least.

In [21] and [23], authors used Naive Bayes, J48, RF, SVM and KNN classification models. The attributes like sugar levels, aluminium levels and the percentage of red blood cells can all be used to predict Chronic Kidney Disease. These parameters were used to evaluate the output of each classifier. Experiments have revealed that the Random Forest algorithm achieves a maximum accuracy rate.

Decision trees outperformed all other classifiers in three out of four efficiency parameters in [22], including predictive accuracy, specificity, and precision, with 98.6% predictive accuracy , 0.972 sensitivity , precision of 1 and specificity of 1.

III PROPOSED ALGORITHM

The approach for establishing the CKD prediction model is shown in Fig.1 and it includes the following steps:

1. Exploration of the data set.
2. Data Pre-processing

    i) Data Clean-up: The data may contain numerous irrelevant and missing pieces. To manipulate this part, the data clean-up is carried out. This involves managing missing data, noise data, etc.

    ii) Data Transformation: This step transforms the data into forms that can be used in the mining process.

    iii) Data Reduction: Data mining is a technique for processing massive amounts of data. Analytics becomes more complex in such situations when dealing with a large volume of data. We use data reduction methods to avoid this.
3. Feature selection
4. Model Fitting and Testing

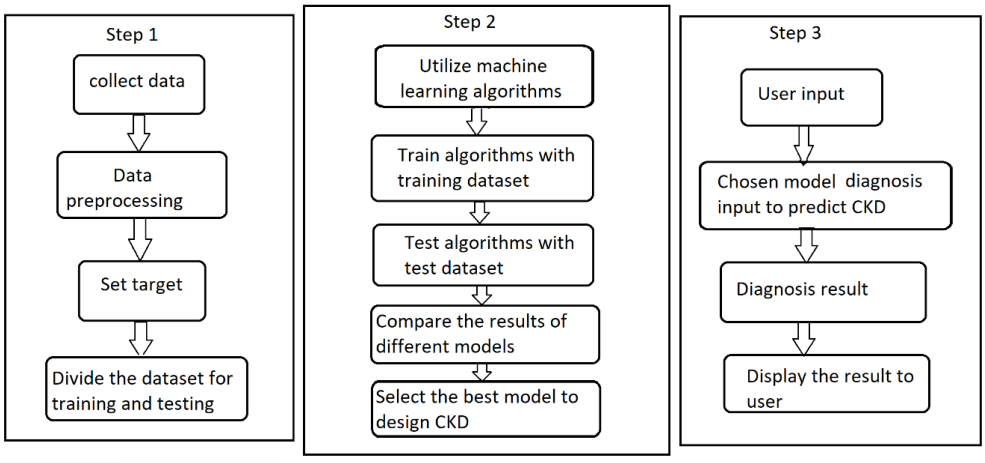

Fig.1: Workflow of the process

(i)Dataset-A set of prediction data for CKD using an automated learning algorithm is downloaded from the UCI (University of California, Irvine) repository. Dataset consists of 400 patient records as well as 25 attributes as shownin Table 2.

Sl. No.	Data type	Features	Unit
1	Numerical	Age	Years
2	Numerical	Blood Pressure(BP)	Mm/Hg
3	Nominal	Specific gravity(SG)	1.005,1.010,1.015…
4	Nominal	Albumin(Al)	0,1,2,3,4,…
5	Nominal	Sugar(Su)	0,1,2,3,4...
6	Nominal	Red blood cells(RBC)	Normal, abnormal
7	Nominal	Pus cells(PC)	Normal, abnormal
8	Nominal	Pus cell clumps(PCC)	Present, Absent
9	Nominal	Bacteria(Ba)	Present, Absent
10	Numerical	Blood glucose random(BGR)	mgs/dl
11	Numerical	Blood urea(BU)	mgs/dl
12	Numerical	Serum Creatinine(SC)	mgs/dl
13	Numerical	Sodium(Sod)	mEq/L
14	Numerical	Potassium(Pot)	mEq/L
15	Numerical	Haemoglobin(Hemo)	Gms
16	Numerical	Packed cell volume(PCV)	0,1,2…
17	Numerical	White blood cell count(WBCC)	cells/cumm
18	Numerical	Red blood cell count(RBCC)	Millions/cumm
19	Nominal	Hypertension(Htn)	Yes, No
20	Nominal	Diabetes mellitus(DM)	Yes, No
21	Nominal	Coronary artery disease(CAD)	Yes, No
22	Nominal	Appetite(Appet)	Good, Poor
23	Nominal	Pedal edema(PE)	Yes, No
24	Nominal	Anemia(Ane)	Yes, No
25	Nominal	CKD, No CKD(Class)	CKD, No CKD

Table 2: Features that influence CKD [19]

(ii)Data pre-processing is a method of converting large and noisy data into appropriate and clean data. Since real-world data includes errors, noisy data and missing values, the proposed system must clean the raw data to eliminate the inconsistent data. This is a critical component of the prediction model completion. It decreases the machine's dimensionality and helps it to produce better performance. The development of a classification prototype is one of the most time-consuming aspects of the process.

(iii) Testing and Training Dataset: There are two data sub-sets in the dataset.
(a) Training data: To train the model for CKD prediction, the larger dataset is used.
(b) Testing data: The performance is predicted using the smaller dataset. The data is used for training 70% of the time and testing 30% of the time. We can test model by making predictions against the test set after it has been dealt with using the training set.

(iv) Feature selection: Attribute selection is the automatic selection of data attributes that are most relevant to the predictive modelling issue. From the entire set of attributes, this step selects a subset of suitable attributes. This stage assists in the model's dimensionality reduction, as well as its simplification and ease of use, resulting in a short training time and

high accuracy. The 5 attributes chosen from a total of 25 attributes are: Blood Pressure, Diabetes Mellitus, Albumin, Specific gravity and Red Blood Cells

1. Blood Pressure is the force of blood circulating through the artery walls is known as blood pressure. The amount of blood the heart pumps and the amount of resistance to blood circulation within the arteries are used to calculate blood pressure. The arteries are in charge of carrying oxygenated blood throughout the body.Blood pressure is elevated if there is a lot of resistance to blood flow, which means blood isn't able to flow freely across the body.
2. Diabetes is a condition characterized by an abnormal rise in blood glucose, also known as blood sugar. Blood sugar, which is derived from the food we eat, is the primary energy source. Insulin, a pancreatic hormone, promotes the absorption of glucose in cells as energy. The body either doesn't generate enough or any insulin or it doesn't use it correctly. Glucose is still present in our system.
3. Albumin is a type of protein found in the bloodstream. Albumin should not move from a healthy kidney into the urine.Albumin is shed through urine from an impaired kidney. Lower the amount of albumin in your urine, the better is your health.The presence of too much protein in the urine results in Albuminuria.
4. Specific gravity is a measurement of the kidneys' ability to concentrate or dilute urine in comparison to plasma. Urine is a mineral, salt, and compound solution that is water soluble. The more concentrated the urine is, the higher the urine essential gravity.
5. Anemia is characterized by low red blood cell count. Anemia is diagnosed by a lower hemoglobin or hematocrit level on a standardized blood test. The key protein in red blood cells is hemoglobin. It transports and distributes oxygen in body. If it's low tissues or organs won't get enough oxygen.

v) Classification: Researchers used four machine learning algorithms to predict the early onset of CKD: neural networks, random forests and K-nearest neighbour and help vector machines. The effectiveness of each algorithm is assessed. From the process as shown in Fig.2, the model with the highest accuracy is selected.

Considering the drawbacks like time consumption of traditional methods the proposed model is a non-invasive and quick model. Among the others, these four algorithms have been identified for having the highest accuracy to predict CKD at an early stage.

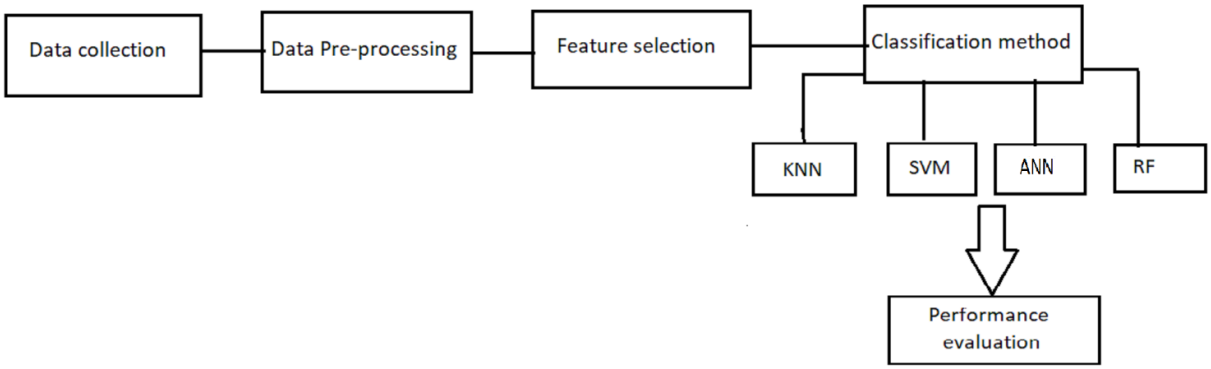

Fig.2: CKD prediction using machine learning algorithms

- K-Nearest Neighbour: When all attribute values are continuous, this form of distance-dependent algorithm is used, and it can be updated based on categorical attributes. To determine the classification of unrecognized proceedings, use the classification of the instance or instances nearest to it. The same concept is used in more cases in the training kit to describe the nearest k neighbour or known neighbour.
- Support Vector Machine: The SVM is a classification and regression linear model that can solve both linear and nonlinear problems. Individual data element will be represented as a point in n-dimensional space (where n is the number of characteristics) and the significance of each character in this algorithm will be the value of a particular coordinate. The classification is complemented by the determination of the best hyper plane to efficiently distinguish the two groups.
- Neural Network are computer systems that are based on biological neural networks seen in animal brains. Artificial neurons are a collection of connected units or nodes that make up an ANN. Each link can send a signal to other neurons, just like synapses in a human brain. An artificial neuron that receives a signal, analyses it, and can send signals to other neurons. We are able to recycle our model over and over again to get the best training, validation and testing. In NN, three layers define this classifier's architecture. They are: an input layer with the same number of neurons as that of data attributes, an output layer with exactly one output layer and zero or more hidden layers between them.
- Random Forest: The random forest algorithm generates a sequence of decision trees that can be used for grading and regression. Several decision trees are generated from random subset of training data sets. The use of a large number of decision trees allows for more accurate results. The algorithm is relatively fast to execute and accounts for missing data. The algorithm is randomised by the random forest, not the exercise dataset. The decision class is the class mode that decision trees generate.

(v) Based on their accuracy, the best model is chosen.
(vi) On this basis, when the user provides the input, the associated output is displayed.
(vii) Based on these when the user provides input the corresponding output is displayed.

IV CONCLUSION

Early and precise detection of CKD can help prevent further deterioration of the patient's health. The main purpose of the study is to determine whether or not an individual has CKD or not. To predict a patient's CKD status, KNN, SVM, Random Forest and Neural network classification algorithms are used. All algorithms shall be compared on the basis of precision. In conclusion, this study helps doctors predict the disease more accurately and in no time so the patients undergo minimal testing relative to a large number of tests required for conventional CKD prediction. Millions of records without missing values will be required to construct a 99.99% reliable CKD model using machine learning.

REFERENCES:

[1]"DaVita Kidney Care" https://www.davita.com/education/kidney-disease/stages
[2] Siddheshwar Tekale , Pranjal Shingavi , Sukanya Wandhekar , Ankit Chatorikar," Prediction of Chronic Kidney Disease Using Machine Learning Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 7, Issue 10, October 2018
[3] Gunarathne W.H.S.D,Perera K.D.M, Kahandawaarachchi K.A.D.C.P, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)",2017 IEEE 17th International Conference on Bioinformatics and Bioengineering.
[4] S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh,"Chronic Kidney Disease Prediction using Machine Learning Models", International Journal of Engineering and Advanced Technology (IJEAT) May 15, 2020
[5] Jayalakshmi V , Lipsa Nayak and K.Dharmarajan," A Survey on Chronic Kidney Disease Detection Using Novel Methods" International Journal of Pure and Applied Mathematics,April 2018
[6] Reem A. Alassaf , Khawla A. Alsulaim ,Noura Y. Alroomi , Nouf S. Alsharif , Mishael F. Aljubeir," Preemptive Diagnosis of Chronic Kidney Disease Using Machine Learning Techniques " 2018 International Conference on Innovations in Information Technology (IIT)
[7] AkmShahariar Azad Rabby,Rezwana Mamata,Monira Akter Laboni"Machine learning applied to kidney disease prediction", Conference: 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)Dec 30, 2019
[8] L.Vishnuvarthini Dr. S.Malarkhodi,"Kidney Disease Classification based on using Machine Learning using Digital Image Processing" International Journal of Innovative Scince, Engineering & Technology, Vol. 6 Issue 3, March 2019
[9] S.Dilli Arasu," Review of Chronic Kidney Disease based on Data Mining Techniques", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp
[10] Gabriel R. Vásquez-Morales , Sergio M. Martínez-Monterrubio , Pablo Moreno-Ger , and Juan A. Recio-García "Explainable Prediction of Chronic Renal Disease", date of publication October 21, 2019
[11] Sumit VibhavPrakash Singh, DhrubJyotiKalita" Detection of Chronic Kidney Disease Using Artificial Neural Network" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, November 10, 2019
[12] El- Houssainy A. Rady, Ayman S. Anwar," Prediction of kidney disease stages using data mining algorithms",Journal Informatics in Medicine Unlocked
[13] S.Ramya, Dr. N.Radha," Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms", International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 1, January 2016
[14] Vijayarani Mohan,S Dhayanand" KIDNEY DISEASE PREDICTION USINGSVM AND ANN Algorithms", International Journal of Computing and Business Research (IJCBR), Volume 6 Issue 2 March 2015

[15]S.D.Harish, K.Vinay Kumar, K.Taraka Ram, G.Pradeepini“Chronic Kidney Disease Prediction based on Blood Potassium Levels using Machine Learning”,International Journal of Innovative Technology and Exploring Engineering Volume-9 Issue-2, December 2019

[16] Manish Kumar “Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm”International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 5, Issue. 2, February 2016

[17] Parul Sinha, Dr. Poonam Sinha,“Performance evaluation of Classification Techniques on Prediction of Chronic Kidney Disease” December 2015 International Journal of Engineering and Technical Research

[18] Kilivia L. De Almeida, Lucilia Lessa, Anny Peixoto Rafael Gomes, Joaquim Celestino” Kidney Failure Detection Using Machine Learning Techniques”, 8th International Workshop on ADVANCEs in ICT Infrastructures and Services, 1 Mar 2020

[19]Marwa Almasoud, Tomas E Ward“Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors”,International Journal of Advanced Computer Science and Applications, Vol. 10, No. 8, 2019

[20]Pushpa M. Patil,”Review on prediction of chronic kidney disease using data mining technique”, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016

[21] Nikitha Saurabh, Tanzila Nargis” Chronic Disease Prediction Using Effective Feature Selection”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019

[22] Sahil Sharma,Vinod Sharma,Atul Sharma” Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis”, International Journal of Modern Computer Science (IJMCS) ISSN: 2320-7868 Volume 4, Issue 3, June, 2016

[23] Pankaj Chittora, Gaurav Kumar Ameta, Prasun Chakrabarti , Gaurav Kumawat,“Analysis of Chronic Kidney Disease (CKD) using supervised machine learning classifiers and curve fitting”, International Journal of Advanced Science and Technology Vol. 29, No. 5, (2020)