JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SURVEY ON CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING

[1]Mrs. Rashmi H, [2]Narasimha Maiya G S, [3]Gagansuri M S, [4]Maharaj S, [5]Nikil B S

[1]Assistant Professor, [2,3,4,5] Undergraduates
[1]Department of Computer Science and Engineering
[1]K S Institute of Technology, Bangalore, India

Abstract : Chronic Kidney Disease (CKD) is a condition which results in loss of kidney function gradually over a period of time. Chronic kidney disease develops when the kidneys become damaged and are unable to filter the blood properly. This damage may cause wastes to build up in the body. Because CKD patients are at a higher risk of End Stage Renal Disease (ESRD), It has become a significant public health issue. Dialysis has cost the nation's health-care system billions of dollars in recent years, and the costs are expected to rise further. Most middle and lower-income people in developing countries cannot afford expensive medical treatment. Chronic kidney disease has no signs in its early stages. The severity of kidney disease can be determined by the estimated glomerular filtration rate (GFR) and albumin levels. We will develop a system based on clinical data that will use a machine learning algorithm to predict the disease in its early stages.

IndexTerms **- Chronic Kidney Disease, Support Vector Machine, Machine Learning, End-Stage Renal Disease (ESRD).**

I. INTRODUCTION

Chronic Kidney Disease (CKD), commonly known as chronic renal failure, is associated with a gradual loss of renal function. In other words, CKD is a disease that affects the proper functioning of the kidneys, meaning that the kidneys are not functioning as expected and blood cannot be filtered properly. The kidneys are a pair of reddish-brown bean shaped organs, each kidney about 4-5 inches long. The kidney's job is to filter waste and excess water from the bloodstream and excrete it from the body through urine. All blood in our body flows through them about 40 times a day. Advanced chronic renal disease can cause the body to store harmful quantities of water, electrolytes, and waste materials. This disease is referred to as "chronic" since the kidney damage occurs progressively over time. As kidney disease advances, it can lead to kidney failure, which necessitates dialysis or kidney transplantation to maintain life.

Chronic kidney disease is identified as one of the world's most serious public health issues. CKD is the tenth biggest cause of death in the world. CKD affects 10% of the world's population. Lack of affordable treatment, killing millions of people each year. Bone disease, anemia, heart disease, high calcium, excessive potassium and fluid retention are all symptoms of kidney disease. Hypertension, High Blood Pressure, Diabetes, Family History and Old Age are considered to be the leading causes of CKD. As the numbers of CKD patients are increasing we require effective predictive measures for early detection of CKD that reduces renal failure and expensive treatment.

Machine learning techniques have recently been widely used for early sign and diagnosis of multiple diseases. Machine Learning (ML) plays an important role in diagnosing illness or disease simply by analyzing existing patient records and training models to predict new patient behavior. Machine learning(ML) is a subset of artificial intelligence in which computing machines learn automatically and thus prediction improves with training.

In a medical examination, 2 medical tests are conducted to determine chronic renal disease. That is, a urine test to look at albumin and a blood test to look at glomerular filtrate. The glomerular filtration rate (GFR) is a test that is used to find out how well our kidneys are performing. It is also the best test for measuring our level of kidney functionality and for determination of chronic kidney disease stages. There are 5 stages of damage severity based on GFR.

Stage	Description	GFR (mL/min/1.73 m^2)
1	Kidney damage with normal or ↑ GFR	≥90
2	Kidney damage with mild ↓ GFR	60–89
3	Moderate ↓ GFR	30–59
4	Severe ↓ GFR	15–29
5	Kidney failure	<15 (or dialysis)

Fig 1: Stages of Chronic Kidney Disease

Fig 1.1 shows that after the stage 2 of CKD, patient will suffer symptoms and will get to know about the reducing of kidney functionality. During initial stages, CKD has no symptoms. Therefore, early detection of CKD can reduce a patient's chances of CKD. With advances in machine learning and artificial intelligence, several classifiers and clustering algorithms are used in order to achieve this goal.

The dataset of CKD, from the UCI Machine Learning Repository is used in this project to analyse CKD using machine learning techniques. The dataset is imported in CSV format, the dataset is preprocessed and the best attributes are selected. Machine Learning Algorithm, Support Vector Machines is used to predict chronic kidney disease. At last we use the Stream-lit library to deploy machine learning model.

The primary goal of this survey is to employ a machine learning algorithm to predict chronic renal disease at an initial stage based on clinical data to avoid kidney failure and costly treatment. The survey's objective is to utilize machine learning methods to create a CKD prediction model based on clinical data and then deploy it using the stream-lit library. Another objective of our survey is to design a model for predicting CKD using the fewest possible indicators.

This paper is presented as follows: Section II is Literature Survey, Section III is description of Dataset, Section IV describes the proposed method for building predictive model, Section V includes results and discussion of the paper and Section VI contains acknowledgment.

II. LITERATURE SURVEY

[1] Imesh Udara Ekanayake and Damayanthi Herath used the following methods to forecast chronic kidney disease in their paper. The features with more than 20% missing values were removed from the analysis. The KNN Imputer technique is utilized to fill in the missing values. In the model training, 11 categorization models were explored. They are KNN regression, logistic regression, SVC with a linear kernel, decision tree classifier, SVC with an RBF kernel, XGB classifier, random forest classifier, extra trees classifier, Gaussian NB, an adaboost classifier, and a conventional neural network. As a result of their research, six algorithms outperformed among 11 algorithms in terms of training accuracy, crossvalidation accuracy and testing accuracy. Random forest classifiers, Decision tree classifiers, XGB classifiers, adaboost classifiers, extra trees classifiers and traditional neural network classifiers are among them.

[2] Bhavya Gudeti, Shashvi Mishra, Shaveta Malik, Terrance Frederick Fernandez, Amit Kumar Tyagi and Shabnam Kumari, employed three machine learning techniques, namely Logistic Regression, K-Nearest Neighbors and Support Vector Machine to classify the disease. The Support Vector Machine approach outperformed than Logistic Regression and K-Nearest Neighbors in predicting Chronic Kidney Disease. The benefit of the work, according to the author is that the prediction process takes less time.

[3] N. Vanitha & S.V. Sendhuraa, to detect Chronic Kidney Disease, the ML algorithm, Support Vector Machine algorithm was applied. To minimize the dimension of the Chronic Kidney Disease dataset, to identify the disease, two fundamental types of feature selection methods, namely wrapper and filter approaches were used. The results demonstrated that the Support Vector Machine classifier, when combined with the Best First search engine feature selection approach and a filtered subset evaluator, had a greater accuracy rate in identification of Chronic Kidney Disease, this method outperforms all others.

[4] Zixian Wang, Jae Won Chung, Xilin Jiang, Yantong Cui, Muning Wang, Anqi Zheng, using an associative and classification algorithm, they provided a machine learning technique for predicting CKD. For better results, the WEKA tool is used to pick the most important characteristics of the CKD dataset. The Apriori association technique is used to prepare the training dataset and the top 10 rules are chosen. For model training, 16 attributes are chosen. The five classifiers are chosen which are formed on association rules such as Naive-Bayes, OneR, k-nearest neighbour, ZeroR and J48. IBk (Instance Based Learning, i.e. k-nearest neighbour) with the Apriori associative algorithm obtained 99 percent accuracy, and a 10-fold cross validation procedure was used to compute the results.

[5] I.A. Pasadana, D. Hartama, M. Zarlis, A.S. Sianipar, A. Munandar, S.Baeha, A.R.M. Alam made use of data mining tools for CKD prediction. Different decision tree algorithms are used for the prediction of the CKD. They compared different decision tree algorithms by using several performance metrics to find out best decision tree for the CKD prediction. The results showed

that RandomForest gives the highest accuracy in identifying CKD. They also claimed that using a decision tree algorithm to detect CKD can help people live longer.

[6] Siddheshwar Tekale, Pranjal Shingavi, Sukanya Wandhekar, Ankit Chatorikar, used two machine learning algorithms Decision tree classifier and Support Vector Machine (SVM) to determine if a patient has CKD or not. Only 14 of the 25 features were employed in the predictive model. The missing values in the dataset are replaced with the mean values of that attribute using the WEKA function "ReplaceMissingValues". The accuracy of the decision tree algorithm was found to be 91.75 percent and the accuracy of the SVM algorithm was found to be 96.7 percent.

[7] Janani J, Sathyaraj R suggested a CKD prediction model based on a hybrid machine learning technique. KNN Imputation technique is used to deal with missing values. Information gain is utilized to help in feature selection. To achieve a clean dataset, pre-processing methods such as label encoding and Min-max normalization are used. Following pre-processing, several machine learning methods such as naive bayes, logistic regression, artificial neural networks and random forests are implemented, and their results are compared using different performance indicators. A hybrid of the Random Forest and Adaboost algorithms is proposed, and it outperforms than other individual component models in terms of accuracy.

[8] Reshma S, Salma Shaji, S R Ajina, Vishnu Priya S R, Janisha utilized the machine learning technique, Support Vector Machine, to provide a machine learning model for CKD prediction. The best attributes from the dataset are chosen using Ant Colony Optimization (ACO). Out of the 24 attributes available, the best 12 are chosen for prediction. Finally, SVM is used to train the model. Predicts CKD patients with fewer features while keeping a greater level of accuracy. They got an accuracy of around 96 percent here.

[9] Marwa Almasoud, Tomas E Ward used machine learning algorithms to determine how well a ML model can forecast chronic kidney disease using only a small collection of features. Outliers have been removed. Missing or empty values in the dataset are replaced using multiple imputations (MI) method. For continuous variables, the imputation method used is linear regression, while for categorical variables, it used logistic regression. Feature selection approaches choose features with a greater relevance to the target variable that are independent of the learning model. The dataset was subjected to 4 machine learning methods, namely Random forest (RF), Support vector machines (SVM), Logistic regression (LR) and gradient boosting (GB). Using 10-fold cross-validation, the classifiers were trained, tested and validated. The gradient boosting algorithm outperformed the competition in terms of F1-measure (99.1%), sensitivity (98.8%) and specificity (99.3%).

[10] Dr. Vijayaprabakaran.K, Pratheek Reddy.P, Puthin Kumar Reddy.T, Munnaf.K, Reddi Prasad.G applied various machine learning approaches to overcome the challenge and treat the disease at an early stage. The mean, median, mode, or constant value of their respective features is used to fill in the empty missing values in the dataset. In categorical attributes, null values are replaced with the most often occurring value in that attribute column. Label encoding is used to convert categorical attribute values into numerical values by converting each unique attribute value to an integer representation. XGradient, Random Forest, and Support Vector Machines are three of the proposed models that use the dataset. Out of the 24 attributes available, the best 12 are chosen for prediction. The suggested method's prediction accuracy in the Chronic Kidney Disease prediction is 97.5 percent using Random Forest, 96.25 percent using XGradient, and 65 percent using Support Vector Machines.

III. DATASET

The chronic kidney disease dataset is created using clinical history, physical examinations, and laboratory tests. The CKD dataset, which includes 400 patient records, was obtained from the University of California, Irvine Machine Learning Repository. In addition to the class feature, there are 24 features in total, with 11 numerical features and 13 categorical features in the dataset. The diagnostic class feature has two values: ckd and notckd. Except for the diagnostic feature, all features have missing values. Only 158 of the 400 records have no null or missing attribute values, while the remaining records all have at least one missing attribute value. The dataset consists 250 cases of "ckd" class (62.5%) and 150 cases of "notckd" class (37.5%), the dataset is unbalanced.

Sl No	Attribute	Attribute Full Name	Type	Unit/value
1	age	Age	numerical	years
2	bp	Blood Pressure	numerical	mm/Hg
3	sg	Specific Gravity	nominal	(1.005, 1.010, 1.015, 1.020, 1.025)
4	al	Albumin	nominal	(0, 1, 2, 3, 4, 5)
5	su	Sugar	nominal	(0, 1, 2, 3, 4, 5)
6	rbc	Red Blood Cells	nominal	(normal, abnormal)
7	pc	Pus Cell	nominal	(normal, abnormal)
8	pcc	Pus Cell Cumps	nominal	(present, notpresent)
9	ba	Bacteria	nominal	(present, notpresent)
10	bgr	Blood Glucose Random	numerical	mgs/dl
11	bu	Blood Urea	numerical	mgs/dl
12	sc	Serum Creatinine	numerical	mgs/dl
13	sod	Sodium	numerical	mEq/L
14	pot	Potassium	numerical	mEq/L
15	hemo	Hemoglobin	numerical	gms
16	pcv	Packed Cell Volume	numerical	
17	wc	White Blood Cell Count	numerical	cells/cumm
18	rc	Red Blood Cell Count	numerical	millions/cumm
19	htn	Hypertension	nominal	(yes, no)
20	dm	Diabetes Mellitus	nominal	(yes, no)
21	cad	Coronary Artery Disease	nominal	(yes, no)
22	appet	Appetite	nominal	(good, poor)
23	pe	Pedal Edema	nominal	(yes, no)
24	ane	Anemia	nominal	(yes, no)
25	class	Class	nominal	(ckd, notckd)

Fig 2: Description of CKD dataset

IV. METHODOLOGY

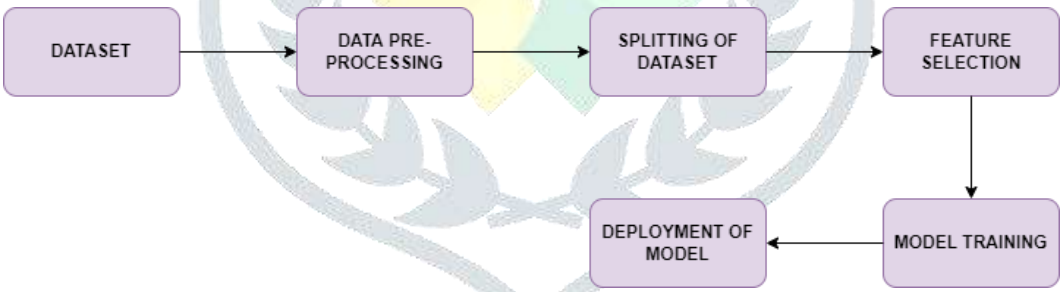

Fig 3: Methodology used for our proposed chronic kidney disease prediction using machine learning

4.1 Dataset: The chronic renal disease dataset obtained from the UCI repository. The dataset consists of 400 patient records and includes 24 features, including 11 numeric features and 13 category features, as well as classification feature such as "ckd" and "notckd".

4.2 Data Pre-Processing: Many sections of the data may be irrelevant or missing. Data pre-processing is used to manage this part. Data Pre-processing is the process of transforming raw data into data that is important and necessary. It encompasses dealing with missing data, noisy data, and so on.

- Except for the diagnostic feature, all features have missing values. It comprises 250 cases of "ckd" class (62.5%) and 150 cases of "notckd" (37.5%), the dataset is uneven or unbalanced.
- The missing numerical and nominal values will be replaced using the KNN Imputer technique.
- Conversion categorical values into numerical values using map() function.
- To handle unbalanced data, SMOTETomek class is used to perform oversampling using Synthetic Minority Oversampling Technique (SMOTE) and cleaning using Tomek links.

4.3 Splitting of Dataset: The train-test split is an approach for calculating or evaluating the performance of a machine learning algorithm. This method includes dividing a dataset into two categories.
Train Dataset: This dataset is used for training the machine learning model.
Test Dataset: This dataset is used to evaluate the trained machine learning model.

The goal of this technique is to calculate the machine learning model's performance on new data that was not used to train it before. We will be splitting the dataset into 75% for training data and 25% for testing data.

4.4 Feature selection: It is the process of lessening the amount of given input variables when designing a predictive model. It helps to minimize the computational cost of modelling and in some cases, it also helps to improve the model's performance. The number of input variables should be reduced. For selecting features, we'll use the Recursive Feature Elimination approach. Recursive Feature Elimination also commonly referred as RFE, is a prominent feature selection algorithm. RFE is one of the feature selection algorithm with a wrapper. RFE is one of the popular feature selection method because it is simple to install, run and use. It is effective in determining which features (columns) in a training dataset are more significant in predicting the outcome variable.

4.5 Model Training: The initial stage in training an ML model is to provide training data to a Machine Learning algorithm (that is, the learning algorithm). The learning algorithm looks for patterns in the provided training data that link the input data attributes to the output (the result we want to predict i.e. target attribute) and then generates an ML model that captures these patterns. The Support Vector Machine or SVM, will be used to forecast Chronic Kidney Disease. SVM is one of the most popular Supervised Learning technique that may be applied to classification problems as well as regression problems.

4.6 Deployment of model: Creating a Machine Learning model is not enough until we make it available to general use or to a specific client. Stream-lit is a popular open-source framework used for model deployment by machine learning. Stream-lit lets to create apps for machine learning project using simple python scripts. At-last step the learned machine learning model is deployed using stream-lit library.

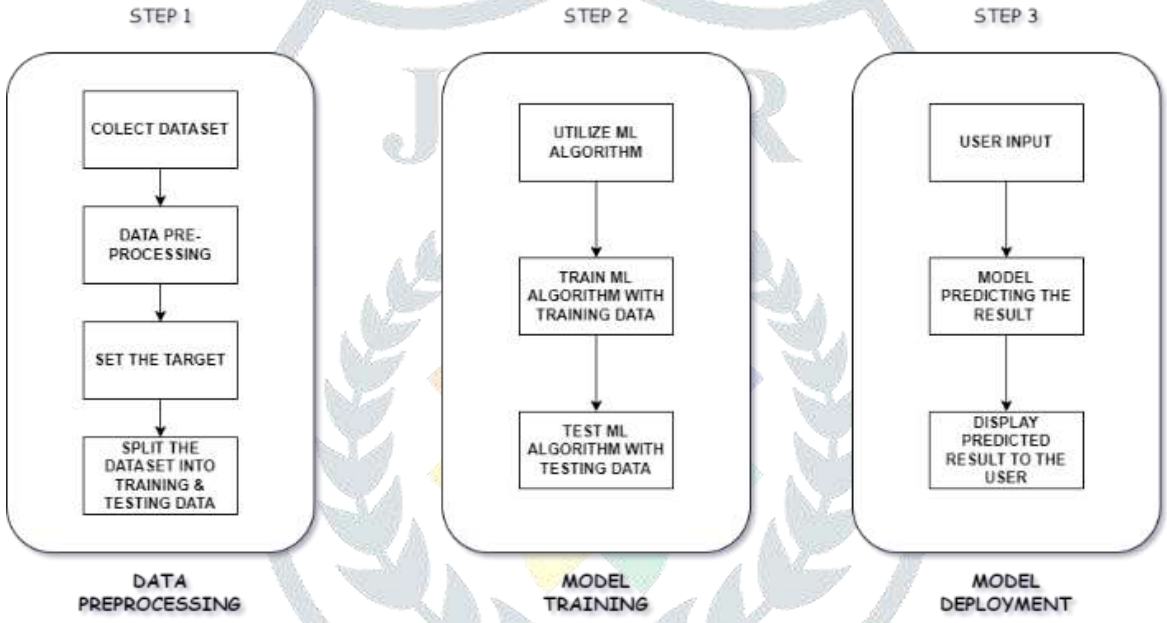

Fig 4: Steps in designing a prediction model for chronic kidney disease using ML.

V. RESULTS AND DISCUSSION

This survey helps to propose a model that helps in chronic kidney disease prediction as explained in section IV. Early detection and treatment of CKD can be implemented at low cost, thereby reducing the burden of ESRD, improving diabetic and cardiovascular disease (including hypertension) outcome and significantly lowering patient morbidity and mortality. Our intention is to provide an effective system in the simplest way which helps both doctors and patients to predict chronic kidney disease at early stages. Physicians and radiologists can use a computer-assisted diagnostic system to help them make better diagnostic conclusions. Our method enables doctors to serve a greater number of patients in less time. Proper feature selection methods aid in reducing the amount of features required by the prediction algorithm, hence reducing the number of medical tests required.

For better CKD prediction, Future research should investigate different supervised and unsupervised machine learning techniques, as well as feature selection techniques with additional performance metrics. To collect the most recent data for CKD diagnosis from various regions around the world. The sample size (400 instances) is expected to be small, which may affect the reliability of the studies. As a result, the dataset size must be increased in the future for better accuracy.

VI. ACKNOWLEDGEMENT

We would like to express our special thanks and gratitude to **Mrs. RASHMI H** for her valuable suggestion and useful advice during the planning and development of this project. We would also like to thank all the professors, staff and management of KSIT for their continuous support and encouragement.

REFERENCES

[1] Imesh Udara Ekanayake and Damayanthi Herath, "Chronic Kidney Disease Prediction Using Machine Learning Methods", IEEE Moratuwa Engineering Research Conference (MERCon), (2020), pp. 260-265.

[2] Bhavya Gudeti, Shashvi Mishra, Shaveta Malik, Terrance Frederick Fernandez, Amit Kumar Tyagi and Shabnam Kumari, "A Novel Approach to Predict Chronic Kidney Disease using Machine Learning Algorithms", IEEE Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA), (2020), pp. 1630-1635.

[3] N. Vanitha & S.V. Sendhuraa, "Chronic Kidney Disease Detection Using Machine Learning Techniques", IAR Journal of Medical Sciences, Vol.2, Iss.2, (2021), pp. 127-133.

[4] Zixian Wang, Jae Won Chung, Xilin Jiang, Yantong Cui, Muning Wang, Anqi Zheng, "Machine Learning-Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique", International Journal of Engineering & Technology, (2018), pp. 1161-1167.

[5] I.A. Pasadana, D. Hartama, M. Zarlis, A.S. Sianipar, A. Munandar, S.Baeha, A.R.M. Alam, "Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques", Journal of Physics: Conference Series 1255, (2019).

[6] 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, Iss.10, (2018), pp. 92-96.

[7] Janani J, Sathyaraj R, "Diagnosing Chronic Kidney Disease Using Hybrid Machine Learning Techniques", Turkish Journal of Computer and Mathematics Education, Vol.12, No.13, (2021), pp. 6383-6390.

[8] Reshma S, Salma Shaji, S R Ajina, Vishnu Priya S R, Janisha A, "Chronic Kidney Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT), Vol.9, Iss.7, (2020), pp. 137-140.

[9] 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 (IJACSA), Vol.10, No.8, (2019), pp. 89-96.

[10] Dr. Vijayaprabakaran.K, Pratheek Reddy.P, Puthin Kumar Reddy.T, Munnaf.K, Reddi Prasad.G, "Chronic Kidney Disease Diagnosis Using Machine Learning", International Research Journal of Engineering and Technology (IRJET), Vol.8, Iss.6, (2021), pp. 4029-4033.