**TEAM – 8**

**TEAM MEMBERS:**

KOVVURI SAI LAKSHMI NARAYANA REDDY

THULASWEERAREDDY DUDYALA

NADENDLA SURESH

POOJA REDDY BUNKUMEEDI

**Prediction of Chronic Kidney Disease**

**Using Machine Learning Algorithm**

**Abstract** :

In today's era everyone is trying to be conscious about health although due to workload and busy schedule one gives attention to the health when it shows any symptoms of some kind. But CKD is a disease which doesn't shows symptoms at all or in some cases it doesn't show any disease specific symptoms it is hard to predict, detect and prevent such a disease and this could be lead to permanently health damage, but machine learning can be hope in this problem it is best in prediction and analysis. By using data of CKD patients with 14 attributes and 400 record we are going to use various machine learning techniques like Decision Tree, Random forest, etc. To build a model with maximum accuracy of predicting whether CKD or not and if yes then its Severity.

Keywords: CKD, Decision Tree, Random Forest, Machine Learning

**I.INTRODUCTION :**

We all knows, that Kidney is essential organ in human body. Which has main functionalities like excretion and osmoregulation. In simple words we can say that all the toxic and unnecessary material from the body is collected and thrown out by kidney and excretion system. There are approximately 1 million cases of Chronic Kidney Disease (CKD) per year in India. Chronic kidney disease is also called renal failure. It is a dangerous disease of the kidney which produces gradual loss in kidney functionality. CKD is a slow and periodical loss of kidney function over a period of several years. A person will develop permanent kidney failure. If CKD is not detected and cured in early stage then

patient can show following Symptoms: Blood Pressure, anaemia, weekboans,

poor nutrition health and nerve damage, Decreased immune response because at advanced stages dangerous levels of fluids, electrolytes, and wastes can build up in your blood and body. Hence it is essential to detect CKD at its early stage but it is unpredictable as its Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all so machine learning can be helpful in this problem to predict that the patient has CKD or not. Machine learning does it by using old CKD patient data to train predicting model.

**II. LITERATURE SURVEY :**

There are many researchers who work on prediction of CKD with the help of many different classification algorithm. And those researchers get expected output of their model.

Gunarathne W.H.S.D et.al. [1] Has compared results of different models. And finally they concluded that the Multiclass Decision forest algorithm gives more accuracy than other algorithms which is around 99% for the reduced dataset of 14 attributes.

S.Ramya and Dr.N.Radha [2] worked on diagnosis time and improvement of diagnosis accuracy using different classification algorithms of machine learning. The proposed work deals with classification of different stages of CKD

according to its gravity. By analysing different algorithms like Basic Propagation Neural Network, RBF and RF. The analysis results indicates that RBF algorithm gives better results than the other classifiers and produces 85.3% accuracy.

S.Dilli Arasu and Dr. R. Thirumalaiselvi [3] has worked on missing values in a dataset of chronic Kidney Disease.

Missing values in dataset will reduce the accuracy of our model as well as prediction results. They find solution over this problem that they performed a recalculation process on CKD stages and by doing so they got up with unknown

values. They replaced missing values with recalculated values.

Asif salekin and john stankovic [7] they use novel approach to detect CKD using machine learning algorithm. They get result on dataset which having 400 records and 25 attributes which gives result of patient having CKD or not CKD.

They use k-nearest neighbours, random forest and neural network to get results. For feature reduction they use wrapper method which detect CKD with high accuracy.

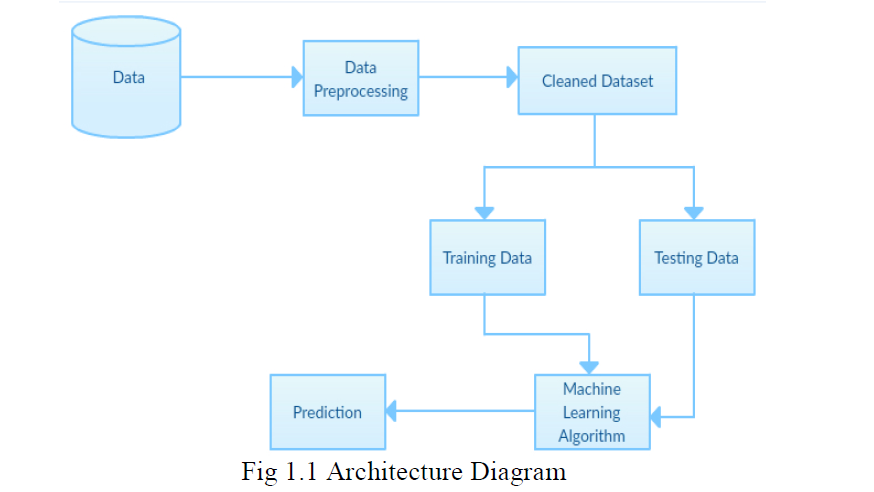
Pinar Yildirim [8] searches the effect of class imbalance when we train the data by using development of neural network algorithm for making medical decision on chronic kidney disease. In this proposed work, a comparative study

was performed using sampling algorithm. This study reveals that the performance of classification algorithms can be improved by using the sampling algorithms. It also reveals that the learning rate is a crucial parameter which significantly effect on multilayer perceptron.

Sahil Sharma, Vinod Sharma, and Atul Sharma [9], has assessed 12 different classification algorithm on dataset which having 400 records and 24 attributes. They had compared their calculated results with actual results for calculating the accuracy of prediction results. They used assessment metrics like accuracy, sensitivity, precision and specificity. They find that the decision tree technique gives accuracy up to 98.6%, sensitivity of 0.9720, and precision of 1 and specificity of 1.

Dataset and Attributes: In this paper CKD dataset [4] is downloaded from UCI repository. This dataset includes 400 patients’ records with 25 attributes. All this 25 attributes are main attributes which are related to CKD disease. Out of 25 attributes we only use 14 attributes to build our predictive model.

**III. METHODOLOGY :**

****

**i) Dataset - Dataset of prediction of chronic kidney disease using machine learning algorithm** is downloaded from KAGGLE . In that dataset there are 400 patient records are included. Also they include 25 attributes but we takeonly 14 attributes for building model. Age, Blood pressure, Albumin, Red blood cells, Pus cell, Pus cells clumps, Serum creatinine, Haemoglobin, White blood cell count, Red blood cell count, Anaemia, Classification, Appetite,

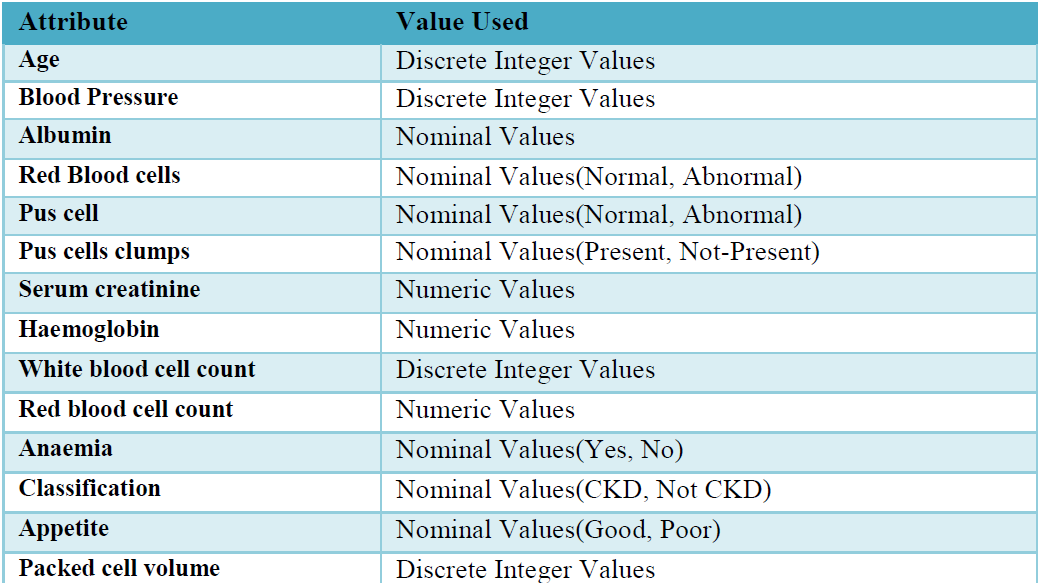
Packed cell volume all this 14 attributes are used to build model.

**ii) Data Prepossessing:**

Data Cleaning: Gather open source raw data of CKD patients available on internet. Data obtained from internet does not contains the name of the attribute so first we assigned the names to the attribute. Missing values in the dataset like NA’s or blank values are removed by using WEKA function “ReplaceMissingValues” used, which replaces NA’s with the mean values of that attribute.

Data Reduction: Out of 25 attributes present in the dataset, we have selected 24 important attributes required to build predictive model.

Following table shows the selected attributes:

****

**iii) Training and Testing Dataset:** The dataset is divided into two sub datasets both containing 24 attributes.

**Training data**: training dataset is derived from main dataset and it contains 300 out of 400 records in main dataset ofCKD.

**Testing data**: testing dataset is of 100 out of 400 records from main CKD dataset.

**iv) Classifiers:**

**Decision Tree:** Decision tree is a graphical representation of specific decision situation that used for predictive model, main component of decision tree involves root, nodes, and branching decision. Decision tree is used in those area of the medical science where numerous parameters involved in classification of data set. Since decision tree is most compressive approach among all machine learning algorithm. These clearly reflect important features in the data set. They can also generate the most affecting feature in the mass of population. Decision tree is based on entropy and Information gain clearly signifies the importance of dataset. Drawback of decision tree is that it suffers from two major problems overfitting and it is based on greedy method. overfitting happened due to decision tree split dataset aligned to axis it means it need a lot of nodes to split data, this problem is resolved by J48 explained in based on greedy method lead to less optimal tree, if dynamic approach is taken it lead to exponential number of tree which is not feasible.

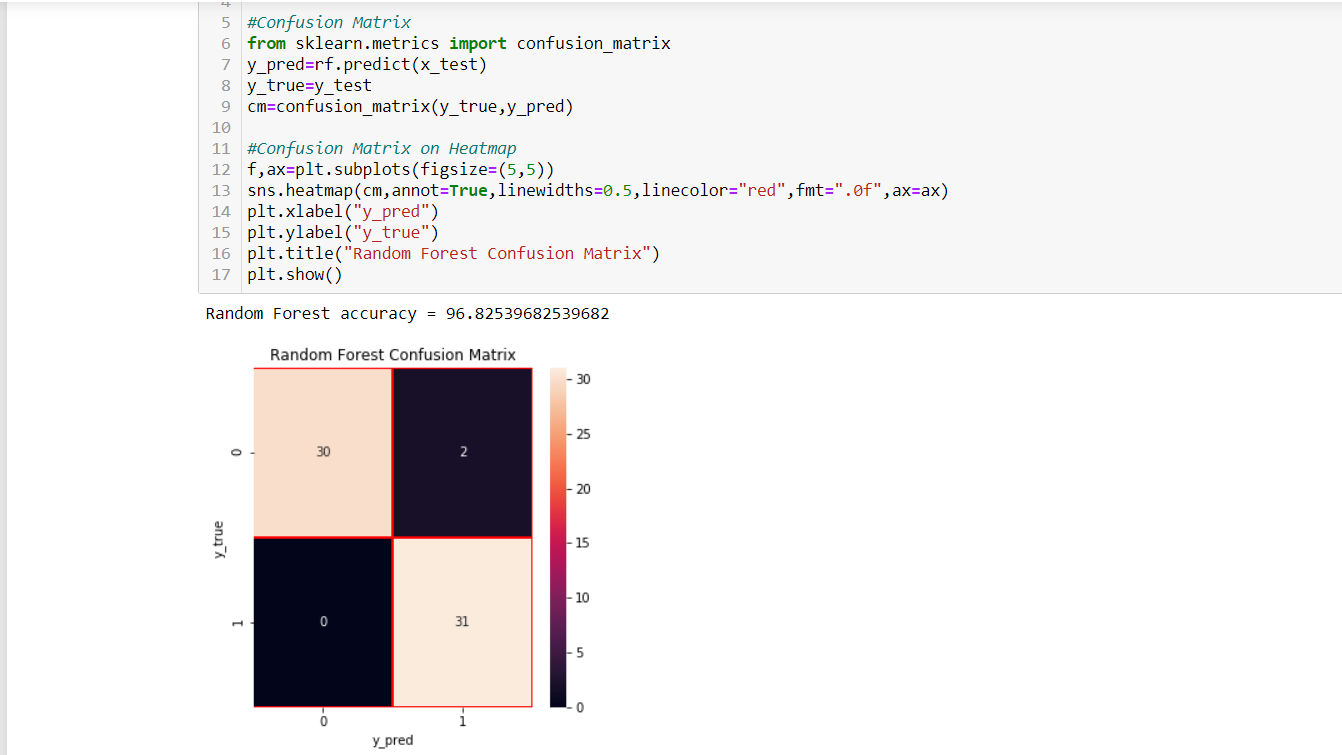
**Random Forest :**

Medical [data classification](https://www.sciencedirect.com/topics/computer-science/data-classification) is considered to be a challenging task in the field of medical informatics. Although many works have been reported in the literature, there is still scope for improvement. In this paper, a feature ranking based approach is developed and implemented for medical data classification. The features of a dataset are ranked using some suitable ranker algorithms, and subsequently the [Random Forest classifier](https://www.sciencedirect.com/topics/computer-science/random-forest-classifier) is applied only on highly ranked features to construct the predictor. Since both decision tree and random forest got same accuracy but when compared to decision tree ,random forest is giving good results .

**v) Prediction:**

Prediction done by Random Forest model trained by training CKD dataset shows following results:

**RANDOM FOREST :**



**ACCURACY SCORE** : 96.82539

**IV) FUTURE SCOPE :**

This work will be considered as basement for the healthcare system for CKD patients using machine learning Algorithms.

**CONCLUSION**

In this , we have studied different machine learning algorithms. We have analysed 24 different attributes related to CKD patients and predicted accuracy for different machine learning algorithms like Decision tree and Random forest

From the results analysis, it is observed that both the decision tree and Random forest algorithms gives the same accuracy . The advantage of this system is that, the prediction process is less time consuming. It will help the doctors to start the treatments early for the CKD patients and also it will help to diagnose more patients within a less time period. Limitations of this study are the strength of the data is not higher because of the size of the data set and the missing attribute values. To build a machine learning model targeting chronic kidney disease with overall accuracy of 99.99%, will need millions of records with zero missing values.