CHRONIC KIDNEY DISEASE DIAGNOSIS USING MACHINE LEARNING

Dr. Vijayaprabakaran.K[1], Pratheek Reddy.P[2], Puthin Kumar Reddy.T[3], Munnaf.K[4], Reddi Prasad.G[5]

[1]Assistant Professor, Department of Computer Science and Engineering, Madanapalle Institute of Technology and Science, Madanapalle
[2-5]B. Tech IV year, Department of Computer Science and Engineering, Madanapalle Institute of Technology and Science, Madanapalle

---

Abstract- *Chronic Kidney Disease(CKD) results in damage to the Kidneys. It is a global health problem and many people are losing their productive years of life. The 40% of persons with CKD are completely unaware that they have it, unlike other diseases CKD can't be cured unless it is predicted in early stages. So, in this research, blood pressure and diabetes state of patients are collected because they are important indicators of whether or not a person has CKD. The usage of various machine learning techniques such as Random Forest, XGradient boost and Support Vector Machines are proposed in this research to overcome the problem and detect the disease in early stage. In this research, CKD dataset is used to predict if a person is affected by CKD or not.*

Keywords: Machine Learning, Chronic Kidney Disease, Random Forest, XGradient, Support Vector Machines.

I. INTRODUCTION

As we all know that, the Kidney is one of the most important organs for humans and animals as well. The kidney has main functionalities like osmoregulation and Excretion. It plays a major role in purifying the blood and removes toxic materials and unwanted substances from the body. Chronic Kidney Disease(CKD) is a severe disease and can be a threat to society since this disease makes the kidney function improperly. Every year, there are approximately 10 lakh cases[1] of Chronic Kidney Disease in India. Chronic Kidney Disease can be detected by regular laboratory tests. There are some treatments to stop the development. This disease can cause permanent kidney failure. If CKD is cured in early-stage then the person can show symptoms like Blood Pressure, anaemia, poor health, weak bones and since the kidney starts to function improperly, the throw-out of waste in the person's body will be minimal. Hence it is essential to detect CKD at its early stage but some people have no symptoms. So machine learning can be helpful to predict whether the person has CKD or not. Glomerular Filtration Rate(GFR) is the best test to measure the level of kidney functionality and can determine the stage of Chronic Kidney Disease. There are five stages of damage severity based on GFR.

Table 1: Stages of Chronic Kidney Disease

Stage of Chronic Kidney Disease	Description	e-GFR level
One	Kidney function remains normal but urine findings suggest kidney disease	90 ml/min or more
Two	Slightly reduced kidney function with urine findings suggesting kidney disease	60 to 89 ml/min
Three	Moderately reduced kidney function	30 to 59 ml/min
Four	Severely reduced kidney function	15 to 29 ml/min
Five	Very severe or end-stage kidney failure	Less than 15 ml/min or on dialysis

Table 1. shows that only after the stage 2 of CKD, the patient will get to know about the reducing of kidney functionality. The early detection of CKD can reduce the chance of CKD for the patient. With the advancement in machine learning and artificial intelligence, several classifiers and clustering algorithms are being used to achieve this.

This research presents the use of machine learning algorithms for prediction of Chronic Kidney Disease. The dataset used for building the predictive models in this research are available and can be downloaded from the UCI machine learning library [2]. The data is imported in CSV format and cleaned for use. After the dataset is preprocessed and best attributes selected, machine learning algorithms including Random Forest, XGradient and Support Vector Machines, are used for prediction of Chronic Kidney Disease, and a comparison of their accuracy is done for selecting the best model for the

disease dataset. All the analysis and visualization are carried out in python 3.7.6. This paper is presented as follows: Section 3 gives the brief explanation about the machine learning algorithms used. Followed by Section 4 which describes the proposed method for building predictive models. Section 5 explains the experimental results of the models. Section 6 includes conclusion and future scope of the paper.

II.LITERATURE SURVEY

For Chronic Kidney Disease classification, lot of studies have been done and many works applied different techniques like Random Forest Classifier, KNN, Logistic Regression, Decision Tree Classifier etc.

Gunarathne W.H.S.D et.al.[3] Has compared results of different models. And finally they concluded that the Multiclass Decision forest algorithm gives more accuracy than other algorithms for the reduced dataset of 14 attributes.

S.Ramya and Dr.N.Radha[4] 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 analyzing 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 [5] 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 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 neighbors, random forest and neural network to get results. For feature reduction they use wrapper method which detect CKD with high accuracy.

J. Snegha [10] proposed a system that uses various data mining techniques like Random Forest algorithm and Back propagation neural Network. Here they compare both of the algorithm and found that Back Propagation algorithm gives the best result as it uses the supervised learning network called feedforward neural network.

Mohammed Elhoseny, 2019 described a system for CKD in which it uses Density based feature selection with ACO. The system uses wrapper methods for feature selection.

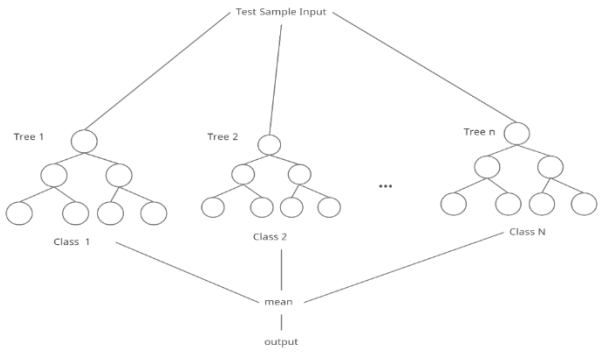
Baisakhi Chakraborty [9] proposed development of CKD prediction system using machine learning techniques such as K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine and Multi-Layer Perceptron Algorithm. These are applied and their performance are compared to the accuracy, precision, and recall results. Finally, Random forest is chosen to implement this system.

III. MACHINE LEARNING ALGORITHMS

Random Forest Classifier:

Random Forest(RF)[6] is an ensemble learning method that constructs a multitude of decision trees at training time and gives the output of mean prediction from the individual trees. Each sub-tree model does random sampling with replacement from training data and finally average results from all sub-models. Every sub-model runs in parallel without any dependency. In addition to constructing each tree using a different subset of the data, random forests differ in the way of how trees are constructed[7]. In standard decision trees, each node is branched using the optimum decision for division among all variables, so as to minimize entropy due to the splitting of the data set represented by the parent node. In a random forest, split points of each node are randomly chosen from the best split point among a subset of predictors. Random forest thus avoids overfitting, which is common with a single deep decision tree.

Figure 2. Simplified structure of Random Forest

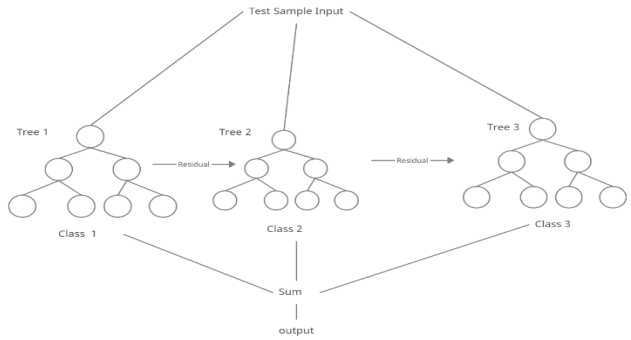


XGBoost:

XGBoost is an optimized distributed gradient boosting library of algorithms. It implements machine learning algorithms under the Gradient Boosting Decision Tree

(GBDT) framework. It is to be noted that the residual from tree-1 is fed to tree-2 so as to reduce the residual and this continues. Different from Random Forest, each tree model in XGBoost minimizes the residual from its previous tree model. The traditional GBDT uses only the first derivative of error information. XGBoost performs the second-order Taylor expansion of the cost function and uses both the first and second derivatives. In addition, the XGBoost tool supports customized cost function.

Figure 3. Simplified Structure of XGBoost



Support Vector Machines(SVM):

In machine learning, Support Vector Machine (SVM) are supervised learning models with related learning algorithms that examine data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVM works by mapping data to a high-dimensional feature space so that data points can be classified, even when the data are not otherwise linearly separable.

IV. PROPOSED METHOD

The proposed method for building the predictive models for the Chronic Kidney Disease is as follows:

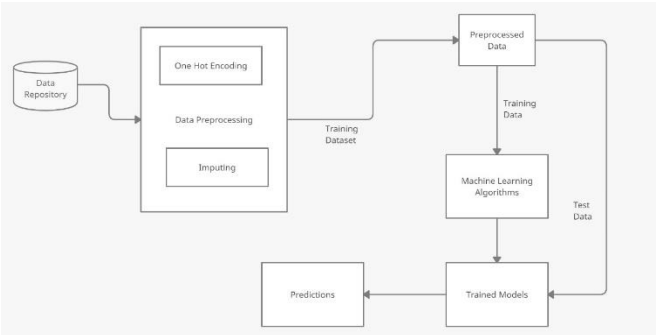


Figure 1. Steps involved in training and testing the model

4.1 Dataset:

The Dataset, CKD Dataset from the UCI repository is used for this project. It contains 400 samples with 25 attributes of two different classes. Out of 25 attributes, 11 are numeric and 13 are categorical and 1 is a class attribute. Some of the data values are missing from the dataset. The dataset contains the patient's data like age, blood pressure, red blood cells, white blood cells, hemoglobin etc.

Table.2 List of attributes present in the CKD dataset

Column	Attributes	Type
Age	Age	Numeric
Bp	Blood Pressure	Numeric
Sg	Specific Gravity	Numeric
Al	Albumin	Numeric
Su	Sugar	Numeric
Rbc	Red Blood Cells	Nominal
Pc	Pus Cell	Nominal
Pcc	Pus Cell clumps	Nominal
Ba	Bacteria	Nominal
Bgr	Blood Glucose Random	Numeric
Bu	Blood Urea	Numeric
Sc	Serum Creatinine	Numeric
Sod	Sodium	Numeric
Pot	Potassium	Numeric
Hemo	Hemoglobin	Numeric
Pcv	Packed Cell Volume	Numeric
Wc	RBC count	Numeric
Rc	WBC count	Numeric
Htn	Hypertension	Nominal
Dm	Diabetes Mellitus	Nominal
Cad	Coronary Artery Disease	Nominal
Appet	Appetite	Nominal
Pe	Pedal Edema	Nominal
Ane	Anemia	Nominal
Classification	Class	Class

4.2 . Pre-Processing:

Data Pre-Processing is the stage where the data is encoded in such a way that the machine can easily analyze it, a dataset can be observed as a group of data objects. In the dataset, there may be a chance of missing values. So, those missing values have to be treated first, they can be either estimated or eliminated from the dataset. The most common method of dealing with missing values is filling them with the mean, median, mode, or constant value of their respective features. Since the object values cannot be used for the analysis we have to convert the categorical data with object type into float64 type. Null values in the categorical attributes are changed with the most occurring value in that attribute column. Label encoding is done to translate

categorical attributes into numerical values by changing each unique attribute value to an integer representation. This automatically changes the attributes to int type. For these operations, pandas package is quiet helpful for data preprocessing.

4.3. Feature Selection:

Feature Selection, the method in which computationally selecting the features which contribute most to the prediction variable or output. This is because using more attribute columns may result in less efficiency in the machine learning model. The classifier algorithm with feature selection gives better performance and reduces the execution time of the model. It is crucial for any predictive modeling and is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

4.4. Model Fitting and Testing:

After the feature selection, the data splits into two parts in the ratio of 4:1, that is 80% of the data is for training the data and remaining 20% of the data is for testing the trained models. the data is used for the 3 proposed models such as Random Forest, XGradient and Support Vector Machines. Then the comparisons are done according with the prediction accuracy of the models.

V. RESULTS AND DISCUSSION

The metrics provided below gives us information on the quality of the outcomes that we get in this study. A confusion matrix helps us with this by describing the performance of the classifier.

Table.3 Confusion Matrix

Confusion Matrix	**CKD (Predicted)**	**Not CKD (Predicted)**
CKD(Actual)	True Positive(TP)	False Negative(FN)
Not CKD (Actual)	False Positive(FP)	True Negative(TN)

Precision: Precision or positive predictive value here is the ratio of all patients actually with CKD to all the patients predicted with CKD (true positive and false positive).

$$Precision = TP / (TP + FP)$$

Recall: It is also known as sensitivity and it is the ratio of actual number of CKD patients that are correctly identified to the total no of patients with CKD.

$$Recall = TP / (TP + FN)$$

F- Measure: It measures the accuracy of the test. It is the harmonic mean between precision and recall.

$$F\text{-}Measure = 2*((Recall * Precision) / (Recall + Precision))$$

Accuracy: It is the ratio of correctly predicted output cases to all the cases present in the data set.

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$

Support: Support is the correct number of outcomes or r esponses that are present
in each class of the predicted outcome.

Table.4 The Confusion matrix using Random Forest

	CKD (predicted)	**Not CKD (predicted)**
CKD (Actual)	28	0
Not CKD(Actual)	2	50

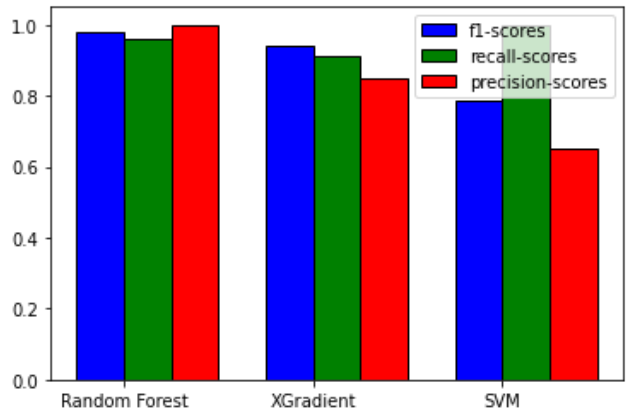
Table.5 The Confusion matrix using XGradient

	CKD (predicted)	**Not CKD (predicted)**
CKD (Actual)	28	0
Not CKD (Actual)	3	49

Table.6 The Confusion matrix using SVM

	CKD (predicted)	**Not CKD (predicted)**
CKD (Actual)	0	28
Not CKD (Actual)	0	52

Figure.4 Performance metrics scores for dataset



The **figure 4** shows the values of precision, Recall and F1 -score Performance metrics for three classifiers for our d ataset.

VI. CONCLUSION

This research deals with the prediction and early detecti on of Chronic Kidney

Disease in people by using Machine Learning algorithms. Out of 24 attributes
present 12 best attributes are taken for prediction. The p rediction accuracy of
our proposed method reaches 97.5% in Chronic Kidney Disease dataset using
Random Forest, 96.25% using XGradient and 65% using Support Vector
Machines.

REFERENCES

[1] "Centers for Disease Control and Prevention."[Online]Available: https://www.cdc.gov/kidneydisease/publications-resources/2019national-facts.html [Accessed: 1-feb-2020].
[2] "UCI Machine Learning Repository" [Online] Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease[Accessed: 24-Sep-2019].
[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.Ramya, Dr. N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering,Vol. 4, Issue 1, January 2016.
[5]. S.Dilli Arasu and Dr. R.Thirumalaiselvi, "Review of Chronic Kidney Disease based on Data Mining
[6]. L. Breiman, "Random Forest," pp. 1–33, 2001.
[7]. M. Denil, D. Matheson, and N. De Freitas "Narrowing the Gap:Random Forests In TheDenil, M., Matheson, D., & De Freitas, N.
(2014). Narrowing the Gap: Random Forests In Theory and In Practice. Proceedings of the 31st International Conference on Machine Learning,
(1998), 665–673. Retrieved from ht," Proc. 31st Int. Conf. Mach. Learn., no. 1998, pp. 665–673,2014.
[8]. Himanshu Sharma,M A Rizvi,"Prediction of Heart Disease using Machine Learning Algorithms: A Survey",International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169,Volume: 5 Issue: 8
[9] Baisakhi Chakraborty, "Development of Chronic Kidney Disease Prediction Using Machine Learning", International Conference on Intelligent Data Communication Technologies, 2019.
[10] J. Snegha, "Chronic Kidney Disease Prediction using Data Mining", International Conference on Emerging Trends, 2020