Chronic Kidney Disease Prediction Using Gradient Boosting and KNN Classifier

O.Rama Praneeth Kumar[1], T.Naga Sampath[2], M.Lakshmi Narayana[3], N.Sai Prasad[4], N Md Jubair Basha[5]
[1,2,3,4,]UG Students, [5]Associate Professor Department of CSE, Kallam Haranadhareddy Institute of Technology, Guntur, AP, India

Abstract **- Chronic kidney disease (CKD) is a global prevalent ailment that causes lives in a predominant number. Predictive analytics for healthcare using machine learning is a challenged task to help doctors decide the exact treatments for saving lives. Scientist researched collaboratively chronic kidney diseases, with the majority of their work on pure statistical models, generating numerous gaps in the development of machine-learning models. In this article we discussed the current methods and suggested improved technology based on the Gradient Boost, which combined significant characteristics of the F scores and evaluated four pre-processing scenarios. In addition, it is provided for machine training methods for anticipating chronic renal disease with Clinical information. Four techniques master Teaching are explored including Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting and Decision Tree, K-Nearest Neighbor. The components are made from UCI dataset of chronic kidney disease and the results of these models are compared to determine the best**
classification model for the prediction. From this four preprocessing cases, replacing missing values with mean values of each column and choosing important features was most logical as it allows to train with more data without dropping. However, Gradient Boosting gave the best outcomes in all four cases where it obtained 98% accuracy in case one where nulled valued are dropped, 98.75% testing accuracy for both case two and three where null values were replaced with minimum and maximum values of each column and it scores 100% accuracy in case four where null values are replaced with mean values. Thus, the system can be implemented for early stage CKD prediction in a cost efficient way which will be helpful for under developed and developing countries.

Index Terms **- Chorionic Kidney Disease, Gradient Boosting, Support Vector machine (SVM), Random Forest (RF) and Decision Tree, K-Nearest Neighbor.**

I.INTRODUCTION

A Chronic kidney disease is disease that results a gradual loss of kidney functionality which could lead in death. CKD is comprised of 5 stages [5] and there are ways that the progression of kidney failure could be slowed down or stopped. chronic kidney disease (CKD) is a significant public health problem worldwide, especially for low and medium income countries. Chronic kidney disease (CKD) means that the kidney does not work as expected and cannot correctly filter blood. Aimed in their work to detect chronic kidney disease for diabetic patients using machine learning methods. Kidneys are two bean-shaped organs, each about the size of a fist [1,2,3]. They are located just below the rib cage, one on each side of the spine. The key function of the kidneys is to remove waste products and excess fluid from the body through the urine. The production of urine involves highly complex steps of excretion and re- absorption. This process is necessary to maintain a stable balance of body chemicals. The critical regulation of the body's salt, potassium and acid content is performed by the kidneys and produce hormones that affect the function of other organs. For example, a hormone produced by the kidneys stimulates red blood cell production, regulate blood pressure and control calcium metabolism etc. Chronic kidney disease (CKD) is a major issue worldwide which is a condition characterized by a gradual loss of kidney function over time, 14% of the world population suffer from CKD. Over 2 million people worldwide currently receive treatment with dialysis or a kidney transplant to stay alive, yet this number may only represent 10% of people who need treatment to live. Chronic kidney disease causes more deaths than breast cancer or prostate cancer [2]. The stages of CKD are mainly based on the measured or estimated glomerular filtration rate (eGFR) which is based on creatinine level , gender, race and age. There are five stages of

kidney functionality . The function is normal in stage 1 and minimally reduced in stage 2 but the majority of cases are at stage 3. To predict positive CKD status and the stages of CKD machine learning can be used. [6] Machine Learning grabs a major part of artificial intelligence when it comes to doing predictions from previous data using classification and regression methods. Application of machine learning methods to predict CKD has been explored based on multiple data sets. [7] Among them, the dataset from UCI repository (referred to as UCI dataset hereafter) is identified as a benchmark dataset[8]-[10] . Similar to most of the related work, this work considers the mentioned benchmark dataset. When analyzing clinical data related to CKD, if there are instances with missing attributes then the missing values handling method should be determined based on the randomness of the way they were missed.

The rest of the paper discusses with various sections among them Section I describes about the Introduction of the work. Section II describes the related work. Section III describes the proposed work. Section IV presents the methodologies, Section V discusses about the various steps, Section VI describes about the algorithms, Section VII presents about the results and Section VIII discusses about the conclusion of the paper.

II. RELATED WORK

In his Methodology he compared most of the machine learning approaches including both supervised and unsupervised learning. WEKA Tool used for experiment and PROMISE -NASA Data set is used to train the model[8]. Introduced retrieval and classification model using (CNN) and Long Short-Term Memory (LSTM) for accurate detection. Proposed a method by using Supervised Learning algorithm mainly logistic regression, Naïve Bayes, and Decision Tree using historical data set. And used K-Fold cross validation technique. Focused on Outlier Detection and removal, followed by dimension reduction. Proposed bug detection as binary classification problem e.g. - correct and incorrect, trained the classifier which distinguish incorrect code from correct code by using deep Bugs framework. Proposed a tool or framework named as defect detector framework which works with various compiler and languages e.g. java, gcc, visual studio. Proposed an approach which uses minimum and accurate no of performing metrics at a time by using marginal R square values. Uses chose Eclipse JDT Core dataset. Proposed a one-class SFP(Software Fault Prediction) Model using One Class SVM. Focused on Vulnerability prediction of Web application using machine learning. In this paper input validation and sanitation attributes are generated. It computes static backward slice for each sink. Considering the related work based on UCI CKD data set [7], it was observed that the reasons for many to have less accuracy are the poor handling of missing values and the method of attributes selection. Approach suggested is first classifying the bugs based on their priorities based on severity and component attribute. Uses Xmean Clustering algorithm with Bayes Net Classifier. Uses Supervised learning. Datasets Used: KC1, MC1, AR1, AC6, MC2 to train the model then compares the results of naïve Bayes and Sungetal. performed a study to develop a stroke severity index. They collected 3577 patient's data with acute ischemic stroke.[9] For their predicting models, they used various data mining techniques and linear regression. Their prediction feature got the best result from the k-nearest neighbor model(95%CI). Monteiroetal. [13] performed a study to get a functional outcome prediction of ischemic stroke using machine learning. In their research, they apply this technique to a patient who was passing three months after admission. They got the AUC value above 90%. Kansadub et al. performed a study to predict stroke risk. In the study, the authors employed Naive Bayes, Decision Tree, and Neural Network to analyze data to predict stroke. In their study, they used accuracy and AUC as their pointersj48 (Decision Tree Classifier). Used Supervised learning on 10 Data Sets Provided by means of NASA especially classifiers used are bagging, guide vector machines (SVM), choice tree (DS), and random wooded area (RF) classifiers. Data is collected from an open Source Software where data will be in a form of object-oriented matrices. Model proposed is genetic based Classifier Systems.

III.PROPOSED WORK

Our goal is to implement machine learning model in order to classify, to the highest possible degree of accuracy[8]. The work proposed here uses classification techniques to predict the presence of

chronic kidney disease in humans. The classifiers used are Gradient Booster and KNN classifier. Python sklearn library was used to implement the paper .The data set for chronic kidney disease was gathered and applied on Gradient Booster and KNN classifier to predict the disease. The performance of the classifier and algorithm evaluated based on accuracy and precision.

A.BLOCK DAIGRAM:

Fig.1: The Architecture

The design of this proposed work is described as , the user has the predefined data set which contains all the information related to chronic kidney disease. Later, Python libraries are imported for the data set.

IV.METHODOLOGIES

Predictive modelling is used to analyse the data and predict the outcome[9]. Predictive modeling used to predict the unknown event which may occur in the future. In this process, we are going to create, test and validate the model. There are different methods in predictive modelling. They are learning, artificial intelligence and statistics. Once we create a model, we can use many times, to determine the probability of outcomes. So, predict model is reusable. Historical datails used to train an algorithm[6]. The predictive modelling process is an iterative processand often involves training the model, using multiple models on the same dataset.

V.STEPS

The proposed methodology consist of the 2 main steps :Data Preprocessing and Feature selection ,Model Training.

A.DATA PREPROCESSING AND FEATURE GENERATION

The data is stored in the form of a CSV file with 24 features and an output variable named- "Class" which has value, 'ckd' or 'notckd' (binary classification). This dataset also has null values which are represented by '-' which we have replaced with empty spaces so that they will be represented as null values in a pandas data frame. As we have a relatively smaller dataset of 400 records instead of dropping the null values' rows/columns-- in case of numerical values, we have replaced them with their respective column means, and in case of string values, we have replaced them with the string which has the highest frequency in their respective columns. After removing the null values to make our data numerical and fitting it on various machine

learning models, we have converted the string value containing columns into dummy/indicator columns (binary columns). By considering the the drawback (of [9]) as mentioned in the related work, the K Nearest Neighbor Imputer algorithm was used in this work to fill the missing values. Further feature engineering has been implemented through methods like correlation matrix.

B. Model Training

Splitting the data into test (30%) and training (70%) sets, we have trained the model appropriately. This is followed by a 3 fold split of the training data in sets of training set and validation set to calculate the accuracy scores. We have also calculated the confusion matrix, precision, recall, f1-score, [12] Feature importance and printed the Value of Parameters of the models. ROC curves for the models selected as the optimal subset of attributes to predict CKD.

VI.ALGORITHMS

A. SVM: Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper plane.

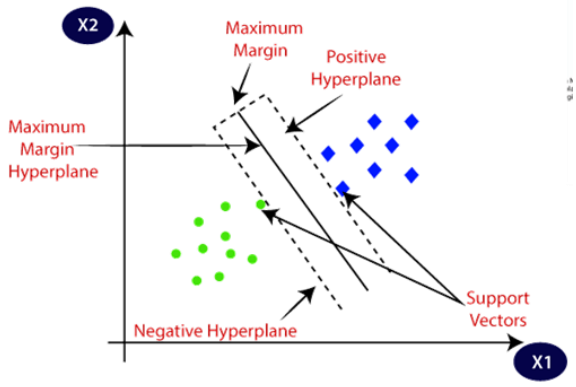

Fig 2 : Support vector machine

B. KNN:K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique's-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm's-NN algorithm can be used for Regression as well as for Classification [8] but mostly it is used for the Classification problems-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

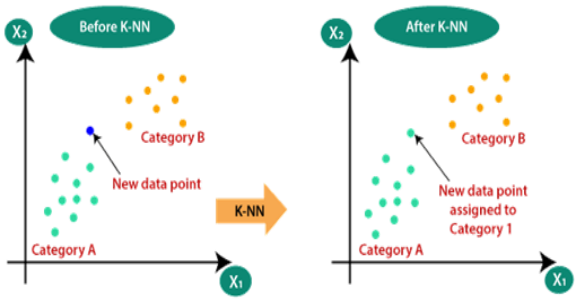

Fig 3: K-Nearest Neighbor

C. Random forest:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. [11]It is based on the concept of ensemble learning, which is a process *of* combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting. The below diagram explains the working of the Random Forest algorithm:

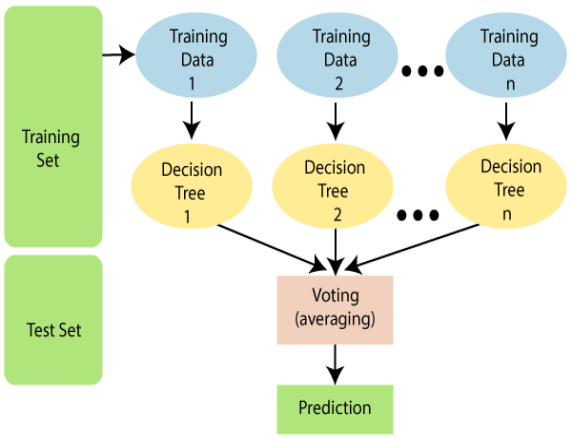

Fig 4: Random forest

D. Gradient Tree Boosting: The tree ensemble model cannot be optimized using traditional optimization methods in Euclidean space. Instead, the model is trained in an additive manner.

Shrinkage and Column Subsampling**:** Besides the regularized objective, two additional techniques are used to further prevent overfitting. The first technique is shrinkage introduced by Friedman. Shrinkage scales newly added weights by a factor η after each step of tree boosting. Similar to a learning rate in stochastic optimization, shrinkage reduces the influence of each tree and leaves space for future trees to improve the model. The second technique is the column (feature) subsampling. This technique is used in Random Forest. Column sub-sampling prevents over-fitting even more so than the traditional row sub-sampling.

The usage of column sub-samples also speeds up computations of the parallel algorithm.

E. Decision Tree Algorithm: Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given data set. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees. Below diagram explains the general structure of a decision tree:

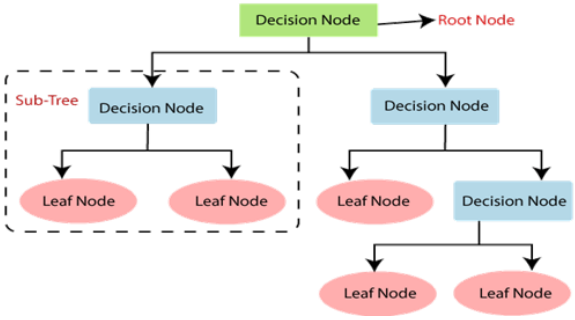

Fig 5: Decision Tree

VII.RESULTS

A. Data set:

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod
1	48	80	1.020	1	0	?	normal	notpresent	notpresent	121	36	1.2	?
2	7	50	1.020	4	0	?	normal	notpresent	notpresent	?	18	0.8	?
3	62	80	1.010	2	3	normal	normal	notpresent	notpresent	423	53	1.8	?
4	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111
5	51	80	1.010	2	0	normal	normal	notpresent	notpresent	106	26	1.4	?
6	60	90	1.015	3	0	?	?	notpresent	notpresent	74	25	1.1	142

Fig 6: Data set

B. Accuracy Tabel:

	Classifier	Accuracy	F1	Precision	Sensitivity	Specificity
0	Random forest	0.9875	0.987440	0.987731	1.000000	0.962963
1	Gradient boosting	1.0000	1.000000	1.000000	1.000000	1.000000
2	Decision tree	0.9750	0.975000	0.975000	0.981132	0.962963
3	Support vector machines	0.6625	0.528008	0.438906	1.000000	0.000000
4	Knn	0.6625	0.669568	0.753309	0.566038	0.851852

Fig 7: Accuracy table

C. Comparison Between Algorithms:

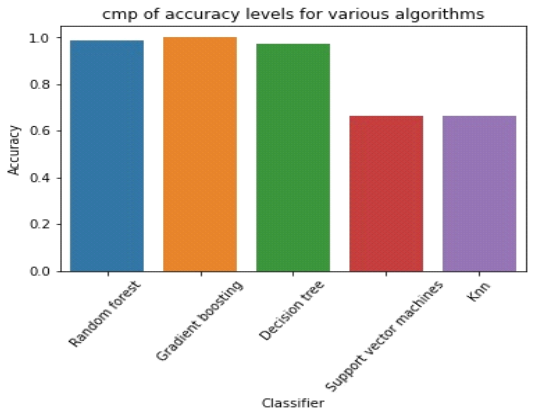

Fig 8: comparison between Algorithms

VIII. CONCLUSION

This paper is aimed to observe and analyze the results obtained by applying different machine learning algorithms in the medical field in order to predict chronic kidney failure. Here we presented a prediction algorithm to predict CKD at an early stage. The dataset shows input parameters collected from the CKD patients and the models are trained and validated for the given input parameters. Decision tree, Random Forest and Support Vector Machine and Gradient Boosting and KNN learning models are constructed to carry out the diagnosis of CKD. The performance of the models is evaluated based on the accuracy of prediction. The results of the research showed that Gradient Boosting model better predicts CKD in comparison to Decision trees and Support Vector machines. The comparision can also be done based on the time of execution, feature set selection as the improvisation of this research.

REFERENCE

[1] "Your Kidneys & How They Work | NIDDK." [Online]. Available: https://www.niddk.nih. gov/

health-information/kidneydisease/kidneys-how-they-work

[2] "Kidney Disease: The Basics," Aug. 2014. [Online].Available:https://www.kidney.org/news/newsroom/factsheets/KidneyDiseaseBasics

[3] "Global Facts: About Kidney Disease," [Online]. Available:https://www.kidney.org/kidneydisease/global-facts-about-kidneydisease/. [Accessed: 20-Feb-2020].

[4] "Estimated Glomerular Filtration Rate (eGFR)," Dec.2015.[Online].Available: https://www.kidney.org/atoz/content/gfr

[5] F. E. Murtagh, J. Addington-Hall, P. Edmonds, P. Donohoe, I. Carey, K. Jenkins, and I. J. Higginson, "Symptoms in the month before death for stage 5 chronic kidney disease patients managed without dialysis," Journal of pain and symptom management, vol. 40, no. 3, pp. 342–352, 2010.

[6] J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of translational medicine, vol. 17, no. 1, p. 119, 2019.

[7] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[8] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (ckd)," in 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2017, pp. 291–296.

[9] A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in 2018 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2018, pp. 1–9.

[10] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," Informatics in Medicine Unlocked, vol. 15, p. 100178, 2019.

[11] Eyck, Jo Van, MortezaKhavaninZadeh, Mohammad Rezapour, Abeer Y. Al-Hyari, Xudong Song, ZhanzhiQiu, Jicksy Susan Jose and N. Uma Maheswari. "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm." (2016).

[12] M. Baumgarten and T. Gehr, "chronic kidney disease: detection and evaluation", *American family physician*, vol. 84, no. 10, pp. 1138, 2011.

[13] Swathi Baby, P. &Panduranga, T. (2015). Vital, Statistical Analysis and Predicting Kidney Disease Using Machine Learning Algorithms. International Journal of Engineering Research and Technology, 4(07), 206-210.