EasyChair Preprint

№ 7854

Comparative Study on Chronic Kidney Disease Detection Using Tree-Based Models

Siddhartha Krishna, Maanvi Keesara and Mary Subaja

April 28, 2022

Comparative Study on Chronic Kidney Disease detection using Tree-Based Models

Siddhartha Krishna[1], Maanvi Keesara[2], Dr. Mary Subaja Christo[3]

[1,2]*B. Tech Student Dept. of Computer Science & Engineering, SRM Institute of Science and Technology, Kattankulathur*

[1]sk1075@srmist.edu.in

[2]mk6262@srmist.edu.in

[3]marysubc@srmist.edu.in

***Abstract*—**

Chronic Kidney disease is the slow, progressive deterioration of the functioning of the kidneys. This impairment is irreversible if it reaches the later stages and hence demands early detection and treatment to ensure prolonged functioning of the kidney.

In this project, we have developed ML model to classify whether a person has Chronic Kidney Disease (CKD) or not. The idea of this study is to calculate the performance of various decision tree-based learning algorithms and to compare their accuracies. In our work, we have used the dataset from UCI which contains real-time data. The UCI's CKD dataset has 400 entries and has missing/noise information. It has 250 patients that have CKD and 150 that have non-CKD, consisting of attributes like age, blood pressure, specific gravity, etc. A total of 5 Tree Bases ML classifiers have been used achieving accuracies as high as 99%.

***Keywords*— Chronic Kidney Disease, Machine learning algorithms, Ensemble methods,Tree-Based Algorithms, Recall, Boosting Algorithms, Precision, Accuracy.**

I. Introduction

Chronic Kidney Disease is one of the most serious conditions wherein the kidneys fail in their functioning and they are not able to filter the blood. Kidneys, they are two bean shaped liked parts in the human body whose primary function is to filter or remove waste blood from the body. If the filtering system is damaged, the protein will be able to seep in our urine and waste elements will just end up remaining in our blood. In the end stage of this disease which is called a renal disease the renal function of the kidney is damaged severely. The symptoms are not shown initially and a person will not be aware of this illness. In the first stage, the illness is often ignored. The second stage, an individual can encounter unobtrusive misfortune in the kidney, third stage is when

an individual has encountered some misfortune in their kidney work. The fourth stage is when an individual will encounter some serious misfortune. In the 5th or the end, stage is when an individual will experience total kidney failure. Chronic Kidney Disease is also called chronic renal failure, has become quite a serious problem in this world where kidneys get damaged and has become the cause for the improper function of the kidney organ. There are several factors that contribute to this but few of the main ones are: Cardiovascular Disease, Records of Hereditary failure of kidney, High Blood Pressure and Diabetes. One of the best method to reduce the death rate due to the diseases is by early treatment.

About 956,000 people lost their lives due to CKD in 2013. It ranked 18 and 27 in the years 2010 and 1990 respectively. But in countries which have not so good healthcare facilities, the patients only take treatment once they are in the serious state.

Automated systems can be built to detect the disease before it reaches the end stage. Clinical data of patients such as sex, blood pressure, age etc. can be used to achieve this. To serve as a solution for the detection of the disease, much research have been done on artificial systems. Machine Learning and Data Mining have proved to be very help full in in the field of medical sciences as there is lot of requirement of well-organized data and methodologies for the analysis, prediction and the detection of diseases. They extract the patterns from the data and these patterns can be used in for the survival of the patients. There are many classification models which have been successfully been implemented and used for this purpose. Some of them being Logistic Regression (LG), Naïve Bayes (NB), and Support Vector Machine (SVM). Boosting algorithms are a type of ensemble machine learning method which converts weak classifier to a strong model to achieve better accuracies. In our research we will be studying tree based ensemble learning models and will try to achieve better and faster results than the already existing models.

II. Literature Review

Many authors have worked extensively in regard to accuracy and effectiveness of different ML models in determining early CKD risk. For instance, [7] examined the correlation between predictors (i.e. input parameters) and the development of CKD using predictive analysis approaches, the accuracy and applicability of using one of the two tests, either blood test or urine test, for the prediction of CKD. While the output of the MLP model could be ascertained on a common metric, it was found that the MLP model's use of a computationally intensive back propagation algorithm enabled the adjust of connection weights and the identification of the ideal set of weights and bias values to predict CKD, while minimising error rate.

[6] considered how SVM algorithms can be made more efficient by means of reducing dimensions and eliminating redundant data. He proposed a combination of CFS algorithm and BPSO algorithm as feature selection to improve the accuracy of the SVM algorithm. The former ensured attributes with good correlation while the

later allowed for the best combination of attributes.

[4] applied Boosting algorithms such as J48 and Ant-Miner to raise the accuracy of CKD detection. It examined the utility of different boosting algorithms AdaBoost and LogitBoost for the diagnosis of CKD to find that both had accuracy close to 100% because they constructed a strong classifier based on several weak classifiers and thus, they improved their performances. When analyzing the decision rules inducted by J48 decision tree and Ant-Miner over CKD dataset, it was concluded that while both rule extractors gave better decision rules, Ant-Miner had a comparative edge over J48 because it iteratively tried until it achieved effective accuracy.

[5] proposed Ant Colony Optimization Based Feature Selection / Extreme Learning Machine for CKD prediction. They concluded that the ELM technique is an improvement under the Sigmoid additive

type of SLFNs. It also extended such a study to generalized SLFNs with different type of hidden nodes. This was found to be capable of providing better accuracy at an acceptable pace.

[9] established and compared nine prediction models using statistical, machine learning and neural network approaches with blood-derived outpatient clinical features and demographic features and even established an online tool for patient follow-up urinary protein severity prediction. Significant performance differences between the different models were found. Linear models LR, Elastic Net, Lasso, and Ridge had an excellent performance with an accuracy rate of 0.80. LR obtained the highest AUC value of 0.873. Elastic Net model, owing to its flexibility was found to be suitable for the early diagnosis of proteinuria progression in patients with CKDs. Also found that ALB, Scr, TG, LDL and EGFR had important impacts on the predictability of the models, while other predictors such as CRP, HDL and SNA were less significant.

[11] compare the decision tree algorithms such as DecisionStump, HoeffdingTree, J48, CTC, J48graft, LMT, NBTree, Random Forest, RandomTree, REPTree, and SimpleCart in predicting CKD based on 7 metrics seven performance metrics, namely, FACC, MAE, PRE, REC, FME, Kappa Statistics and Runtime. Random Forest was concluded to be the strongest classifier in the group, holding the highest accuracy rate of 100%. DecisionStump had the worst accuracy rate with 92%.

[12] utilised KNN imputation to fill in the missing values in the data set. Further, Logistic regression (LOG), RF, SVM, KNN, naive Bayes classifier (NB) and feed forward neural network (FNN) were used to establish CKD diagnostic models on the complete CKD data sets. Out of these the highest performing models were isolated for misjudgement analysis. It was found that by the use of KNN imputation, LOG, RF, SVM and FNN could achieve better performance than the cases when the random imputation and mean and mode imputation were used. KNN imputation would fill in the missing

values in the data set for the cases wherein the diagnostic categories are unknown, which is closer to the real-life medical situation. LOG and RF were found to be suitable component models. Such integration bettered the performance of these otherwise independent models.

[2] measured the performance of several other algorithms such as J48 Decision Tree, Support Vector Machine, Multi-Layer Perceptron, Naïve Bayes Tree, Logistic Regression, and Naïve Bayes, and Composite Hypercube on Iterated Random Projection. Such comparison was based on the metrics of Mean Absolute Error, Root Mean Squared Error, Relative Squared Error, Root Relative Squared Error, Precision, Recall, F-measure and Accuracy. They concluded that based on overall error metrics, CHIRP performance was a minimal error rate as compared to other employed techniques. CHIRP was also found to have better performance than the entire utilized techniques in the overall accuracy metrics.

[3] applied the K-Nearest Neighbour, Support Vector Machine and Ensemble technique to classify the dataset. This evaluation was based on the usage of heat map, visualization graph, Support Vector Machine using rbf kernel and K-Nearest Neighbour with hyperparameter. It was found that for Data Mining Support Vector Machine using rbf kernel was used which gave result of 87% whereas for Machine Learning, K-Nearest Neighbour with hyper parameter was used giving result above 92%. Thus, we can state that machine learning can be helpful in healthcare sector to help predict CKD.

[8] applied Random Forest and ANN model to ascertain the applicability of these 2 models in terms of identifying early CKD. The accuracy of these 2 models was considered based on 5 assessment criteria, namely, Total number of instances, Accuracy, Recall (Weighted Avg.), F1-score (Weighted Avg.), Precision (Weighted Avg.). The accuracy received for random Forest was 97.12%. Error occurrence was at 2.88%. Precision, Recall and F-Measure were 0.97. Alternatively, ANN received an accuracy percentage of 94.50%. Error Occurrence was at 5.5%. Precision, Recall and F1 score were 0.95, 0.94 and 0.95 respectively. It was concluded that as both Random Forrest and ANN models enabled the detection of CKD with reasonable accuracy, risk prediction at early stages would be feasible.

[1] compared the applicability and effectiveness of 8 different supervised classification learning algorithms under different criteria to ascertain which model most effective could be used to identify and diagnose CKD. These 8 algorithms were Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Decision Tree, Random Forest, Naïve Bayes, Multilayer Perceptron and Quadratic Discriminant Analysis. They applied on the dataset provided by the UCI ML repository Found that each algorithm aside from Quadratic Discriminant Analysis had an accuracy of 95 percent. The Random Forest algorithm performed the best with the highest value on each of the performance parameters employed.

[10] utilised estimations of perceptions for neighbouring information focuses to navigate credit missing qualities in a dataset using the KNN Imputer

by scikit-learn k-Nearest Neighbours Algorithm. Exactness, affectability, specificity, accuracy, review and F1 score were utilized to assess the models. Finally, the proposed CKD demonstrative approach was deemed plausible as far as information attribution and tests finding.

III. Methodology

This section will explain the concept of this work and will aid to understand the entire notion of this work. At first, we have collected the data and then we preprocess it. Once we preprocess it, the missing values are handled in the dataset, feature selection is done to extract the most significant features. After this whole process is done, the data is run through the 5 tree based algorithms: Decision Trees, Random Forest, Gradient Boosting Classifier, Ada Boost Classifiers and Extra Trees Classifier. Finally the algorithms are compared and the best algorithm is evaluated.

The CKD dataset taken from the UCI repository has 24 attributes. It has 400 samples to 2 different classes i.e. CKD and NON-CKD. Among the 24 attributes present, 11 of them are numeric and 13 of them are nominal. Some of the values may be absent in the set.

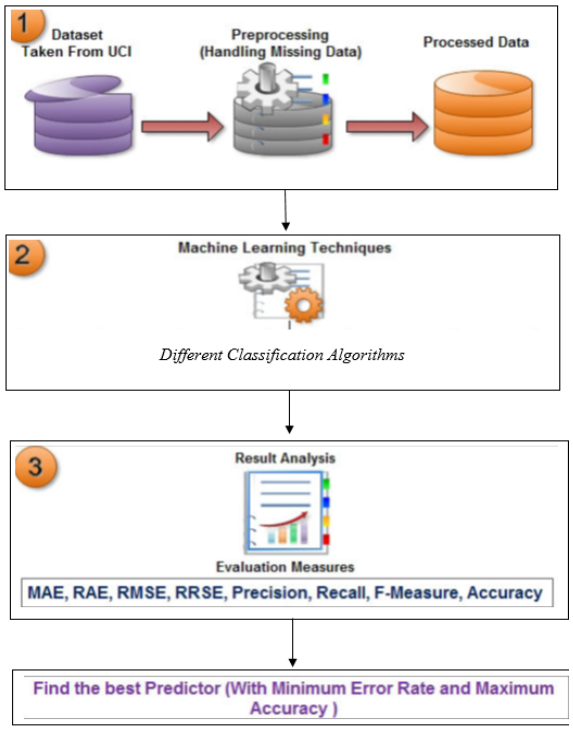

Figure 1: Methodology Workflow

SL	Feature	Data type
1	Age	Numerical
2	Blood Pressure	Numerical
3	Specific Gravity	Numerical
4	Albumin	Numerical
5	Sugar	Numerical
6	Red Blood Cell	Nominal
7	Pus Cell	Nominal
8	Pus Cell Clumps	Nominal
9	Bacteria	Nominal
10	Blood Glucose Random	Numerical
11	Blood Urea	Numerical
12	Serum Creatinine	Numerical
13	Sodium	Numerical
14	Potassium	Numerical
15	Hemoglobin	Numerical
16	Packed Cell Volume	Numerical
17	White Blood Cell Count	Numerical
18	Red Blood Cell Count	Numerical
19	Hypertension	Nominal
20	Diabetes Mellitus	Nominal
21	Coronary Artery Diseases	Nominal
22	Appetite	Nominal
23	Pedal Edema	Nominal
24	Anemia	Nominal

Table 1: List of attributes in dataset

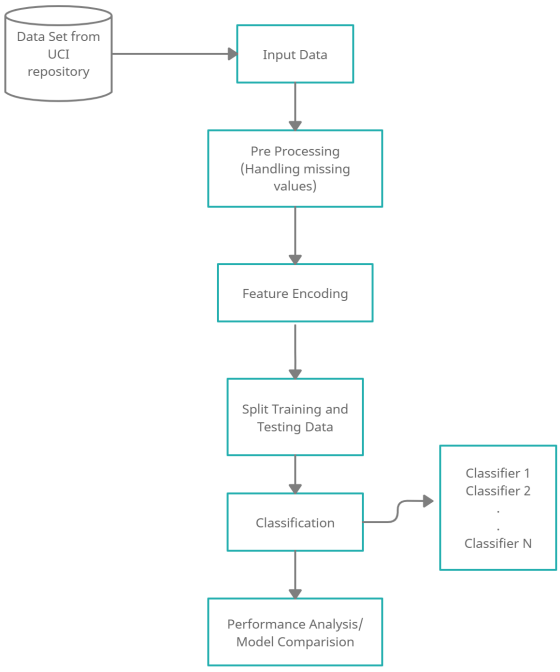

Figure 2: Proposed Architecture diagram

A. Handling Missing Data

In the 400 data samples, about more than 15% of the values are missing. With so many missing values, the performance of the classification model will be affected. Even though some ML models can handle and even ignore the missing data, a majority of them cannot. Thus there will be a waste of data and fundamental learning errors may take place. Hence, the first step is to address the absent data. While preprocessing the data, we managed to fill these absent values using the "Median" method. Often when data is missing, the gaps are filled with 0s which would skew the result slightly, using the median method gives more accurate results by using some median/mean/mode terms in places where the data is missing.

Another method we used to fill on the the missing data is random sampling method in which a random variable is picked in an completely unbiased way.

Parameters	Measurement	Missing	Percent
Glucose	Num. (mg/dL)	44	11
Urea	Num. (mg/dL)	19	4.8
Creatinine	Num. (mg/dL)	17	4.3
Sodium	Num. (mEq/L)	87	21.8
Potassium	Num. (mmol/L)	88	22
Haemoglobin	Num. (g/dL)	52	13
Packed Cell Volume	Num.	71	17.8
White Blood Cell Count	Num. (cells/mcL)	106	26.5
Red Blood Cell Count	Num. (m.c./mcL)	131	32.8
Specific Gravity	Num. (1.002-1.030)	47	11.8
Urine Glucose	Category (0-5)	49	12.3
Albumin	Category (0-5)	46	11.5
Bacteria	Binary	4	1
Red Blood Cells in Urine	Binary	152	38
Pus Cell	Binary	65	16.25
Age	Num. (years)	9	2.3
Hypertension	Binary	2	0.5
Blood Pressure	Num. (mm/Hg)	12	3
Diabetes	Binary	2	0.5
Coronary Artery Disease	Binary	2	0.5
Appetite	Binary	1	0.25
Pedal edema	Binary	1	0.25
Anemia	Binary	1	0.25

Table 2: Missing value description

B. Feature Encoding

In this procedure we turn nominal categorical data to numerical data, this procedure is necessary since Machine Learning algorithms can only process numerical data. Feature encoding converts features such as having diabetes, hypertension, etc or not into 0s and 1s where 0 represents not having the disease or attribute and 1 represents having it.

C. *Split into training and testing*

In this procedure, the dataset is taken and divided into two subsets. The first subset is used to fit the model and the second subset is taken as the testing dataset. The first is used to train the model whereas the latter is used to test it. The second subset isn't accustomed to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values.

D. *Proposed Algorithm*

The ensemble method comes under Machine Learning techniques in which several base models are combined in order to produce a final one an optimal predictive model. The main principle behind this is that a group of weak learners are grouped together to form a strong learner and to overall increase the accuracy of the model. The focus of our study is Ensemble methods specifically focused Tree Based Algorithms which are highly accurate in terms of CKD detection because they work on the principle of making weak learners to strong learners. There are many tree-based algorithms but our focal point are 5 algorithms that give high accuracies. The 5 algorithms that are being studied are Decision Trees, Random Forest, Gradient Boosting Classifier, AdaBoost Classifiers and Extra Trees Classifier. These algorithms have accuracies ranging from 96% - 99%. In our research we are using all 24 attributes used in CKD detection.

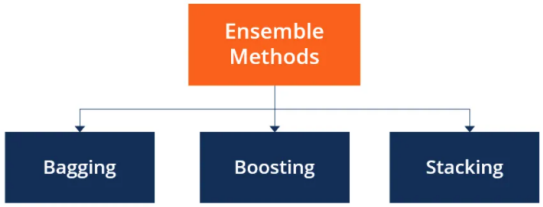

Figure 3:Types of Ensemble Methods

D.1. Decision Trees

A decision tree has a structure which includes one root node, branches and has leaf nodes. The data is divided into classes on the basis of the value of the attribute found during training the samples. The output of a decision tree is a flowchart that uses branching and then represents all possible outcomes of a decision. thus, the internal node represents the attribute, the branches represent the outcome, and the class label is represented by the leaf node. The path from the root to leaf is the classification rules. The leaf node depicts the final decision.

D.2. Random Forest

Random Forest is a ML algorithm that operates in a specific way i.e by creating many number of trees during the training period and then providing the output class of the individual trees. This algorithm can be used for regression as well as classification.We in this code use it for the classification problem of CKD. The model performs a minor tweak which utilises the de-correlated trees that builds multiple decision trees on the bootstrapped samples obtained from the training data, a process

which is called bagging. This decreases variance but inturn increases bias.

D.3. Extra Tree Classifier

Extra Tree classifier is a modified version of bagging classifiers. It still uses the basic ordinary tree techniques with an additional purpose of improving efficiency and accuracy. The differentiating point with the other tree bsed algorithms are that they split the node by randomly choosing cut points and build trees using total learning samples.

D.4. AdaBoost Algorithm

Adaptive Boosting also know as Adaboost takes extra copies of a base classifier consecutively on the same dataset. Decision stumps are used as weak learners. Decision stumps are nothing but trees which have only one split. More weight is given to difficult to classify instances whereas lesser weight is given easy to classify observations. An average of the weighted output from all the individual learners gives the final result.

D.5. Gradient Boosting Algorithm

Gradient boosting is a powerful algorithm. It helps to reduce bias error. Weak predictors are combined to form a strong predictor. The trees are connected in series to reduce the error made bt the previous trees. Due to this sequential connecting, the boosting algorithm may be slow, but is highly accurate.

IV. RESULT

Algorithm	Accuracy	Sensitivity	Precision
Extra Tree	99.167	0.99	0.99
Random forest	97.5	0.97	0.98
Descion Tree	96.667	0.97	0.97
Ada Boosting	97.5	0.97	0.98
Gradient Boosting	99.167	0.99	0.99

Table 3: accuracies and senstivities and precisions of models

We have used 3 methods to compare the algorithms which are accuracy, sensitivity and precision. Accuracy(Eq.1) is calculated using the values from the confusion matrix that include True Positive(TP), True Negative(TN), False Positive(FP) and False Negative(FN).Sensitivity(Eq.3) is calculated using the recall is calculated by dividing the True Positive(TP) with the sum of True Positve(TP) and False Negative(FN). Sensitivity helps us calculate the missed positive prediction by the algorithms these can also be determined as false negatives. Precision(Eq.2) is the fraction of positive predictions made by the algorithm and helps us determine if the algorithm has any false positives. We have chosen features such as sensitivity and precision as they are essential to the health care sectors as even one False Negative or one False Positve can be extremely dangerous to human beings life.

$Accuracy = TP + TN / TP + FP + TN + FN$
Equation 1. Accuracy

Precision = $TP / TP + FP$
Equation 2. Precision

Recall = $TP / TP + FN$
Equation 3. Recall

Using these 3 parameters we compare all the 5 algorithms. We have determined that both Exta tree and Gradient Boosting algorithms have the same high accuracy of 99.1% and sensitivity and precision of 0.99 and also have only 1 false negative and 0 false positives each. Ada Boosting algorithm has an accuracy of 97.5% sensitivity of 0.97 and precision of 0.98 and has 1 False Posotive and 2 False Negative. Random forest Algorithm has the same accuracy,sensitivity and precision as Ada boosting algorithm but has 0 false positives and 3 false negatives.

IV. Conclusion and Discussions

Detection of a disease is the most significant step for its prevention. In countries with less advanced healthcare, the risk of people getting the disease and it going unnoticed is relatively high. The research employed with the Chronic Kedney disease detection using Machine Learning can save millions of lives. The results achieved from the study

The algorithm with the least accuracy is Descion Tree with an accuracy of 96.7% sensitivity and precision of 0.97 and 2 false positives and 2 false negatives.

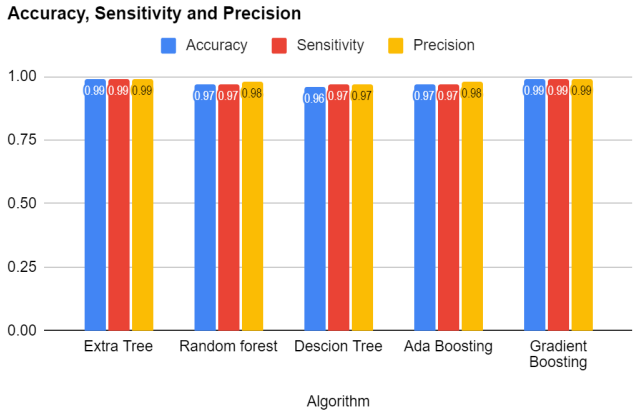

Figure 4:Comparisons of models

indicate that Extra Tree Classifier algorithm and Gradient Boosting algorithm are extremely accurate. In the future, our target is to research on analsis of other diseases that can use various Tree based algorithms for detection and give highly accurate results. We also want to focus on finding various ways to make CKD detection cheaper and easily accessible so that it can be used by people of all backgrounds. We can state that machine learning can be really helpful in the healthcare sector to help predict chronic kidney disease.

V. REFERENCES

1. Mirza Muntasir Nishat , Fahim Faisal,* , Rezuanur Rahman Dip , Sarker Md. Nasrullah, Ragib Ahsan , Fahim Shikder , Md. Asfi-Ar-Raihan Asif and Md. Ashraful Hoque “A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms”. 13 August 2021

2. BILAL KHAN 1 , RASHID NASEEM 2 , FAZAL MUHAMMAD 3 , GHULAM ABBAS 4 , (Senior Member, IEEE), AND SUNGHWAN KIM “An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy” March 18,2020.

3. Adeeba Azmi1, Amiksha Hingu2, Ruchi Dholaria3, Ms. Alvina Alphonso. “Chronic Kidney Disease Prediction using Data Mining and Machine Learning” March 2020.

4. Arif-Ul-Islam, Shamim H Ripon “Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree” February 2019.

5. S.Belina V.J Sara, Dr.K.Kalaiselvi, “Ant Colony Optimization (ACO) Based Feature Selection And Extreme Learning Machine (ELM) For Chronic Kidney Disease Detection”. 2018.

6. Doni Aprilianto, “SVM Optimization with Correlation Feature Selection Based Binary Particle Swarm Optimization for Diagnosis of Chronic Kidney Disease”, September 2020

7. Ahmed J. Aljaaf1, 2, Dhiya Al-Jumeily2, Hussein M. Haglan1, Mohamed Alloghani3, Thar Baker2, Abir J. Hussain2, and Jamila Mustafina “Ahmed J. Aljaaf1, 2, Dhiya Al-Jumeily2, Hussein M. Haglan1, Mohamed Alloghani3, Thar Baker2, Abir J. Hussain2, and Jamila Mustafina”, 2018.

8. 8.Shanila Yunus Yashfi; Md Ashikul Islam; Pritilata; Nazmus Sakib, "Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms". IEEE, October 2020

9. Jing Xiao; Ruifeng Ding; Xiulin Xu; Haochen Guan; Xinhui Feng; Tao Sun,"Comparison and development of machine learning tools in the prediction of chronic kidney disease progression", BMC 2019

10. U Abinaya, S Anitha Devi, B Haritha1 and T Raghunathan;"Noval Approach For Chronic Kidney Disease Using Machine Learning Methodology",Journal of Physics: Conference Series, Volume 1916, 2021 International Conference on Computing, Communication, Electrical and Biomedical Systems (ICCCEBS) 2021 March 2021, Coimbatore, India

11. I.A. Pasadana, D. Hartama, M. Zarlis, A.S. Sianipar, A. Munandar, S. Baeha and A.R.M. Alam, "Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques" The International Conferenczon Computer Science and Applied Mathematic October 2018

12. Qin, Jiongming; Chen, Lin; Liu, Yuhua; Liu, Chuanjun; Feng, Changhao; Chen, Bin." A Machine Learning Methodology for Diagnosing Chronic Kidney Disease." IEEE 2019