Proceedings of the International Conference on Automation, Computing and Renewable Systems (ICACRS 2022)
IEEE Xplore Part Number: CFP22CB5-ART: ISBN: 978-1-6654-6084-2

Chronic Kidney Disease Detection using AdaBoosting Ensemble Method and K-Fold Cross Validation

1st N.Mohana Suganthi
Assistant Professor, Dept.of CSE
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology
Chennai, India
mohanasuganthi@veltech.edu.in

2nd Jemin V.M
Assistant Professor, Dept.of CSE
R.M.K College of Engineering
Chennai, India
vmjemin@gmail.com

3rd P.Rama
Assistant Professor, Dept.of CSE
Bharath Institute of Higher Education & Research
Chennai, India
ramaponnuvarman@gmail.com

3rd E.Chandralekha
Assistant Professor, Dept.of CSE
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology
Chennai, India
clchandralekha08@gmail.com

2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) | 978-1-6654-6084-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACRS55517.2022.10029047

***Abstract*— CKD (Chronic Kidney Disease) is among the most serious health issues worldwide. Among the causes of total global mortality, chronic kidney disease ranked 27th in the Global Burden of Disease Study. Millions of people die every year as a result of an inability to afford treatment for chronic kidney disease. A kidney disease can be slowed or stopped if it is diagnosed and treated early. In order to predict CKD, machine learning is effective. In this paper, dataset from UCI repository is used with 24 features like age, sugar, albumin etc. The AdaBoosting ensemble algorithm, as well as K-Fold Cross Validation technique, is used to predict CKD. The missing data in this dataset is filled in using the Mean Imputation technique. Next, the classification of data is carried out using KNN, Decision Trees (DT), Random Forest(RF), Naïve Bayes(NB), Logistic Regression(LR) and its ensembles. The results were compared with KNN, Decision Tree, Naïve Bayes, Random Forest, and Logistic Regression. Experimental results shown that AdaBoost-Random Forest ensemble is 99 % accurate for early detection of kidney illness.**

Keywords*— *Chronic Kidney Disease, Decision Tree, Ensemble Method, Machine Learning, Random Forest, Naïve Bayes

I. INTRODUCTION

Blood in a person's body is filtered every 30 minutes by the kidneys. By excreting toxins and waste, the kidneys cleanse the body. Furthermore, they work to regulate your blood pressure, produce red blood cells, regulate vital blood chemicals and maintain bone health. Chronic kidney disease (CKD) is characterized by failing kidneys that cannot filter the blood properly. Thus, the body retains excess fluid and waste from blood which may lead to heart disease and strokes. In addition to anaemia and a low number of red blood cells, CKD can also increase the risk of infections, Loss of appetite and diminished eating habits are caused by little calcium levels, high potassium and high phosphorus intensities.

There are different levels of severity for CKD. The disease worsens over time, but treatment can slow it down. Untreated CKD can lead to kidney failure and early death. In the event that the kidneys stop functioning, dialysis or kidney transplants may be necessary. When a kidney fails and requires dialysis or a transplant, it is called end-stage renal disease (ESRD). There may be no symptoms or symptoms that are unexpected for patients with CKD. CKD can only be diagnosed with specific blood and urine tests. As part of these tests, creatinine in the blood is measured as well as protein in the urine. As per the Global Burden of Disease (GBD) study 2017, the contribution of over 354 sicknesses and wounds, as well as 84 risk features to disease and death in 195 countries in the year 1990 and 2017. According to this study, a new analysis estimates morbidity and mortality at global, regional, and national levels due to chronic kidney disease (CKD) and decreased kidney function. As of 2017, CKD caused 1.2 million demises globally , ranked 12th among the important causes of demise. The impairment of kidney function is estimated to contribute to 7.6% (1.4 million) of all CVD deaths. As a percentage of all deaths, 4.6% were caused by CKD. Over the last two decades, global CKD death has increased by 41.5% but age-standardized mortality decreased.

Over the last 25 years, global age-standardized mortality for cardiac disease (CVD), cancer declined. However, the decline for chronic kidney disease (CKD) was not as impressive. According to Bikbov, impaired kidney function was associated with the highest DALY rate compared to medicine use, improper hygiene, low exercise, and smoke, as well as some dietary danger factors. World economy is greatly affected by kidney disease. Approximately 2-3% of the annual income of high-income countries is spent on this disease. US spending on this disease reached 64 billion dollars in 2015. People will benefit from determining which factors contribute to these diseases. Previous research on this topic was conducted both manually and automatically. Automated research has become uneconomical because it is costly. The automatic detection system is sometimes efficient, but is not reliable. Additionally, these automatic detection approaches are not trustworthy and cannot be used by non-doctors. It is also worth noting that the literature deals almost exclusively with prediction or classification purposes in all cases of the developed approaches. Our goal in this contribution has been to integrate machine learning models with ensemble techniques to propose a model that will assist both patients and experts to take the necessary

preventive measures during the early stages of treatment. Both time and cost are inefficient when using machine learning algorithms in the literature. We developed an effective model for our society by combining machine learning and data mining concepts [1].

II. LITERATURE SURVEY

A. J. Aljaaf et al. [1] presented a simplest machine learning algorithm, taking into consideration the fewest number of features or tests. The researchers used four different machine learning algorithms on a small set of 400 records: logistic regression, SVM, random forest and AdaBoosting. By analysing the relationship between variables, the number of features could be reduced and redundancies removed. The remaining attributes were filtered using a feature selection method, with haemoglobin, albumin, and specific gravity predicting CKD best [7].

S. Vijayarani et al. [2] developed a model to detect CKD in early stage. In this work, they collected features from CKD patients and the data is validated using the features of CKD. Decision tree(DT), Random Forest(RF) and SVM algorithms are used detect CKD and the performance is calculated based on the accuracy of the algorithm. Among the three classifiers, Random Forest algorithm performs better. T Shaikhina and Torgyn, et al. [3] linked the performance of 5 classifiers in detection of CKD. Experiments shown that both RF and XGB were better at classifying data. They gave suggestion to use Ensembles of different classifiers in the future. [8]

Jaymin Patel,et al. [4] used J48, Logistic model tree and Random Forest algorithm for detecting the heart disease. They used both machine learning (ML) and data mining techniques. The performance of J48 was found to be the best algorithm because of its accuracy and model building time. Dataset is taken from UCI repository. To improve the model's performance in the future, they suggested using discretization and voting techniques.T Shaikhina, et al. [5] used Decision Tree (DT) and Random Forest (RF) classification models. They used small dataset of 80 samples as the input. It appears that the three most important factors contributing to acute rejection are the donor-specific antibody level, the IgG4 subclass antibody level, and the number of mismatches in the donor-recipient map of human leucocyte antigens.

C.T. Tran, et al. [6] implemented ensemble of classifiers by combining multiple imputation and ensemble learning. By using multiple imputations, diverse imputed datasets can be generated, which are then used to create diverse classifiers. They compared the proposed approach to dealing with incomplete data with two other popular approaches that use decision trees as classification algorithms. According to the outcomes, the proposed method is significantly more perfect at classifying data in almost all cases than other methods.

Minhaz Uddin Emon et al. [8] presented a novel technique for detecting chronic kidney disease based on a few attributes, one of the most common causes of chronic kidney disease. A total of eight machine learning classifiers were used: Logistic Regression(LG), Multilayer Perceptron(MLP), Naive Bayes(NB), Stochastic Gradient Descent(SGD), Adaptive Boosting (Adaboost), Bagging, Decision Tree(DT), Random Forest(RF) classifiers. PCA was used to extract features. According to the results,Random Forest (RF) has the best accuracy with 99% and the highest ROC curve (receiver operating characteristic curve).

Subas et al. [14] Validated different Machine Learning classifiers with the real time data taken from UCI repository and the results are compared with the previous findings in the literature survey. From the result Random Forest achieved better performance when compared to other classifiers. Huseyin Polat et al. [15] used 2 different feature selection methods namely wrapper and filter to take the important features in the dataset. In the wrapper method greedy and classifier subset evaluator were used. In filter method, greedy with correlation feature selection method is used. From the result greedy with classifier subset evaluator achieved higher accuracy of 98.5% when compared to other methods.

As the main purpose of this test is to determine whether a particular patient is suffering from Chronic Kidney Disease or not, it must be accurate and precise. The AdaBoosting ensemble algorithm, as well as K-Fold Cross Validation, is used in this work to predict CKD. For this purpose, we used CKD data set from Kaggle and examining the correlation between the growth of the CKD and predictors using a predictive approach of the analysis. Eliminating missing, duplicate and noisy data reduces the required features to detect CKD. And we have to use certain features to measure its accuracy and predictions. The dataset contains only 2 classes: CKD – affected by Chronic kidney disease ii) Not CKD - Not affected by chronic kidney disease.

III. PROPOSED WORK

The proposed work comprises two stages. Mean Imputation is used in the early stage to replace the missing values in the column. Then, the next stage involves classification using Decision Tree, Random Forest, KNN, Naïve Bayes and Logistic Regression. After that, AdaBoost Ensembles of the above-mentioned algorithms are constructed and their performance is evaluated. Fig. 1 depicts the system architecture of the above mentioned technique.

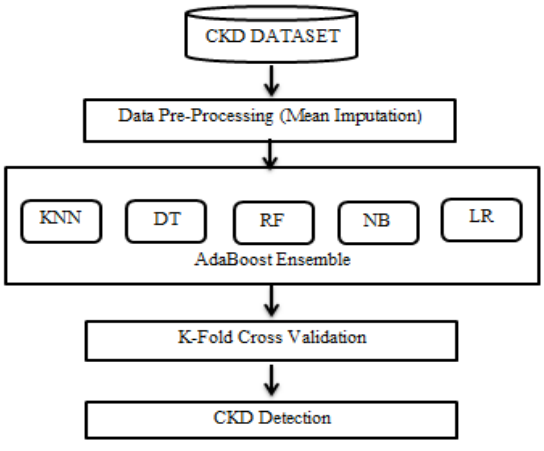

Fig. 1. System Flow Diagram

A. Dataset

Dataset for detecting CKD was downloaded from Kaggle. It contains 25 columns. In that 24 columns are the features of CKD and the last column represent the class of CKD. The features include RBC, WBC, Albumin, sugar, Bp etc for 400 patients. From this dataset, we can discover

patients at risk for CKD based on the patterns that lead to their discovery. These records are divided into two categories: training and testing. A CKD classification is the target, which is either 'CKD' or 'not CKD'. Fig. 2 Shows the attributes used in CKD dataset. Some of the features of CKD dataset are age, al(albumin), bp(blood pressure), rbc(red blood cell), sg(sugar) etc.

B. Data Pre-Processing

In our dataset, lost values in some columns is a big problem. So, it should be pre-processed before training the model. Fig. 2 shows the features of CKD dataset. Fig. 3 shows the missing values in each feature. Totally 400 patients data are there. But in this dataset some values are missing. For example, in age column only 391 values are there (9 patients data are missing). In bp column out of 400 patients 388 patients data are there (12 values missing). Before training the dataset, filling the missing values with some value is necessary. Then only the model gives good result. In order to handle the missing values in the dataset, we used mean Imputation method. The mean value of the non-missing cases replaces the missing values of a variable. The pre-processed dataset is given as the input to the classification algorithm. Fig. 4 depicts the heat map of the features. This heat map is used to visualize the strength of correlation among variables. It helps to find the features that are best for Machine Learning model building. From this heat map, we come to know that the age, bp, sg, rbc,wc features are having high correlation among them. Fig. 5 shows the important Features for detecting CKD in the dataset. Here, other column values are based on age, bp, sg, rbc values. So, age, bp, sg, rbc values are considered as important features. From Fig.. 5, we can conclude that sg, bp, wc, pot, bu, ba are the important features. The CKD detection accuracy is based upon the importance of features and correlation among the features.

Sr. No	Attribute Name	Description
1	Age	Patient age (It is in years)
2	Bp	Patient blood pressure (It is in mm/HG)
3	Sg	Patient urine specific gravity
4	Al	Patient albumin ranges from 0-5
5	Su	Patient sugar ranges from 0-5
6	Rbc	Patient red blood cells two value normal and abnormal
7	Pc	Patient pus cell two value normal and abnormal
8	Pcc	Patient pus cell clumps two values present and not present
9	Ba	Patient bacteria two values present and not present
10	Bgr	Patient blood glucose random in mg/dl
11	Bu	Patient blood urea in mg/dl
12	Sc	Patient serum creatinine
13	Sod	Patient sodium
14	Pot	Patient potassium
15	Hemo	Patient hemoglobin (protein molecule in red blood cells)
16	Pcv	Patient packed cell volume % of red blood cells in circulating blood
17	Wc	Patient white blood cell counts in per microliter
18	Rc	Patient red blood cell count in million cells per microliter
19	Htn	Patient hypertension two value Yes and No
20	Dm	Patient diabetes mellitus two value Yes and No
21	Cad	Patient coronary artery disease two value Yes and No
22	Appet	Patient appetite two value good and poor
23	Pe	Patient pedal edema two value Yes and No
24	Ane	Patient anemia two value Yes and No
25	Class	Target Variable (CKD or Not)

Fig. 2. Attributes used in CKD dataset

C. AdaBoosting Classifier

AdaBoost algorithm is used in this work to increase the accuracy of the classifier. Ideally, it should be used with weak learners. A classification problem can be classified with a high degree of accuracy using these models. Iterative ensemble methods are used in AdaBoost. By combining several poorly performing classifiers, the AdaBoost classifier creates a strong classifier with high accuracy. In this work, Decision Tree, KNN, Random Forest, Naïve Bayes (NB) classifiers and Logistic Regression are used as the base classifiers and the weights are increased based upon the results of previous prediction. Decision Tree, Random Forest, KNN, Naïve Bayes classifiers and its ensembles are used in this work.

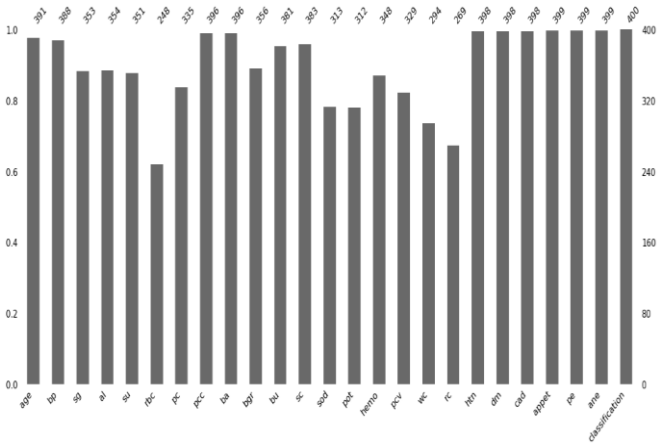

Fig. 3. Missing values in Dataset

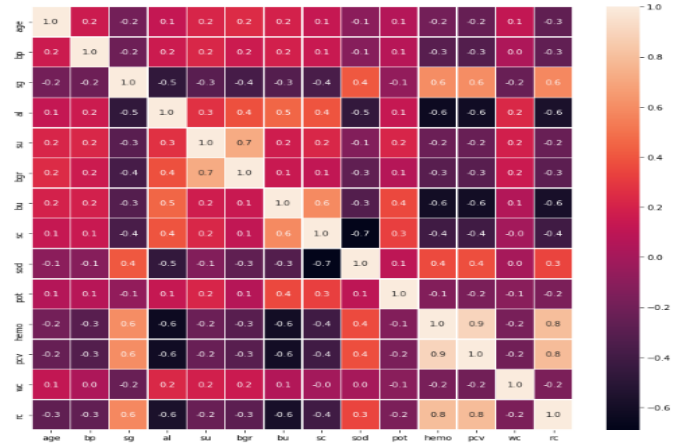

Fig. 4. Heat Map

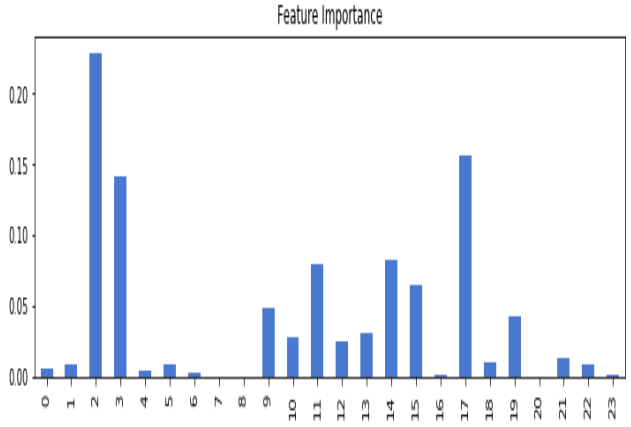

Fig. 5. Important Features

An AdaBoost classifier is created by combining several poorly performing classifiers into an iterative ensemble. Every time the data sample is trained, AdaBoost ensures

accurate predictions of unusual observations by setting the weights of the classifiers. The algorithm can be used as a base classifier as long as it accepts weights based on the training set. Two conditions must be met for AdaBoost to be effective:

1. Interacting with different weighed training examples is essential for training the classifier.
2. Each time iteration is performed, it aims to minimize training error to provide an excellent fit.

1. Initialize the weight W_i=1/N, i = 1,2,3…N
2. N → Number of data points
3. For i=1 to M (number of iterations):
   a) Fit the classifier G_i to the training data with initial weight W_i.
   b) Compute the error function by using the formula

$$\text{Error Function } EF_m = \frac{\sum_{i=1}^{N}(W_i\, I(y_i \neq G_1(x_i)))}{\sum_{i=1}^{N} W_i}$$

   c) Compute the updated weight by using the formula

$$W_{up} = log\left[\frac{1-EF_m}{EF_m}\right]$$

   d) Update the weight of the weak classifier with the new calculated weight.
   e) Fit the classifier G_i to the training data using the updated weight W_{up}

Fig. 6. Pseudo code for AdaBoost classifier

AdaBoost classifier works in the following steps:

[1] A training subset is randomly selected by AdaBoost at the beginning.
[2] Using the accuracy of the last prediction, the training set is selected iteratively to train the AdaBoost machine learning model.
[3] In the next iteration, incorrectly classified observations will receive a higher weight so they will have a higher likelihood of being classified correctly.
[4] Also, the trained classifier is assigned a weight in each iteration based on its accuracy. Classifiers with higher accuracy will be given more weight.
[5] Until no errors are detected in the entire training data or until the set of estimators has been reached, this process iterates.
[6] Classify all learning algorithms using a "vote".

Fig. 6.represents the Pseudo code for AdaBoost classifier. In this work, Decision Tree (DT), KNN, Random Forest (RF), SVM, Naïve Bayes and Logistic Regression classifiers were implemented individually in the dataset for training and testing. After that, AdaBoost using Decision Tree classifier, AdaBoost using Random Forest classifier, AdaBoost using KNN classifier, AdaBoost using SVM classifier, AdaBoost using Naïve Bayes classifier, AdaBoost using Logistic Regression classifier were implemented using python code. These codes are implemented in Google Colab using Python 3.0

IV. PERFORMANCE MEASURES AND EXPERIMENTAL RESULTS

Decision Trees (DT), Random Forest, KNN, NB(Naïve Bayes), LR(Logistic Regression) are the algorithms used in this work. The performance of each classifier is measured as below.

Precision: The level of precision can be calculated by the ratio of correctly detected positive samples (True Positives) compared to the number of positive samples classified in total. Precision is calculated by the formula in Equation (1).

$$\text{Precision} = \frac{TPs}{TPs+FPs} \quad (1)$$

Recall: Recall can be calculated as the ratio of correct Positive samples to all Positive samples. Detection of positive samples is measured by the recall of the model. A higher recall leads to a greater number of positive samples being detected. Recall is calculated by the formula in Equation (2)

$$\text{Recall} = \frac{TPs}{TPs+FNs} \quad (2)$$

Classification Accuracy: This is the number of times the model predicts correctly the output. It is calculated by dividing the number of correct predictions by the number of predictions made by a classifier. This ratio can be calculated. The formula is given below in equation (3).

$$\text{Accuracy} = \frac{TPs+TNs}{TPs+FPs+FNs+TNs} \quad (3)$$

F-measure: F1 score is to compare two models if one has a low precision and one has a high recall, and vice versa. This score was used to assess both precision and recall at once. A maximum F-score is achieved when the recall equals the precision. The formula is given below in equation (4):

$$\text{F1 Score} = \frac{2*Recall*Precision}{Recall+Precision} \quad (4)$$

TABLE 1. ACCURACY, PRECISION, RECALL, F1 SCORE FOR BASE CLASSIFIER

Classifier	Accuracy	Precision	Recall	F1 Score
Decision Tree	95	95	95	94
Random Forest	97	96	96	96
KNN	96	96	96	95
SVM	94	94	95	94
Naive Bayes	95	95	95	96
Logistic Regression	93	93	94	95

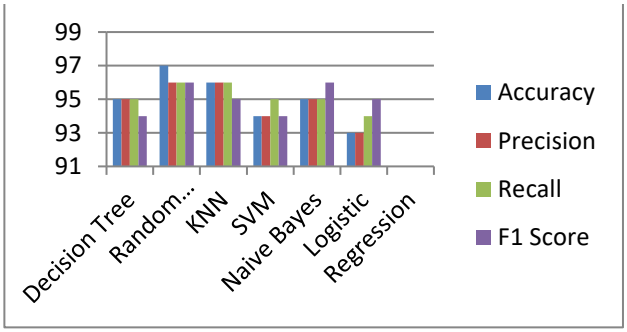

Fig. 7. Precision, Accuracy, Recall and F1 Score of Base classifier

Table I shows the accuracy, F1 score, recall and precision of KNN. Fig. 7 shows the Recall, Accuracy, Precision and F1 Score of Base classifier.

TABLE II. ADABOOST (AB) ENSEMBLE CLASSIFIER

Classifier	Accuracy	F1 Score	Recall	Precision
AB - Decision Tree	93	95	99	91
AB - Random Forest	99	99	98	98
AB - KNN	97	96	96	94
AB-SVM	95	94	95	95
AB - Naive Bayes	96	95	95	96
AB - Logistic Regression	94	93	95	94

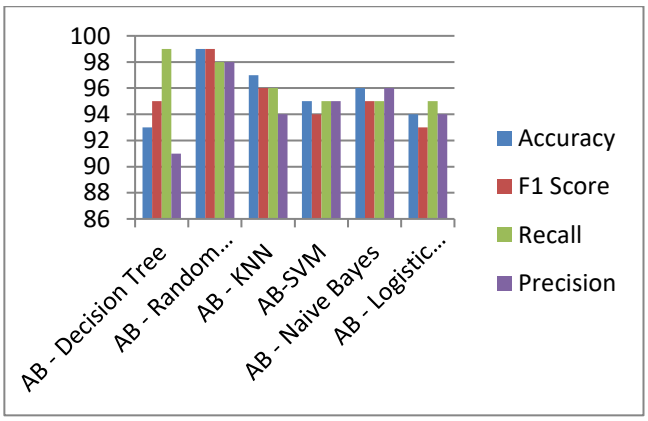

Fig. 8. Accuracy, F1 Score, Recall of AdaBoost Classifier

Table II shows the accuracy, F1 score, recall and precision of AdaBoost - DT, AdaBoost - RF, AdaBoost - SVM, AdaBoost – KNN, AdaBoost – NB and AdaBoost – LR classifier. Fig. 8 shows the Accuracy, Precision, Recall and F1 Score of AdaBoost Ensemble classifier. Models using Base classifier gave less recall rate, accuracy and precision when compared to AdaBoost model using ensembles of different classifiers. AdaBoost using Random Forest Ensemble classifier achieved 99% accuracy when compared to all classifiers.

V. CONCLUSION & FUTURE WORK

The AdaBoost Ensemble model has been used in this work to detect the CKD (Chronic Kidney Disease). In the early stage, dataset was collected from Kaggle. Then the dataset is pre-processed using mean imputation to replace the missing values. The pre-processed dataset was given as the input for the classifiers for training and testing. Ensembles of DT, RT, KNN, Naïve Bayes, SVM classifiers were used. Performance of each model was measured by accuracy, F1 score, recall and precision. From the results, AdaBoost ensembles of Random Forest classifier produced highest accuracy of 99% , AdaBoost - Decision Tree produced % 93,AdaBoost - Random Forest produced 99%, AdaBoost – KNN produced 97%,AdaBoost-SVM produced 95% , AdaBoost - Naive Bayes produced 96%, AdaBoost - Logistic Regression produced 94% accuracy. These results are compared with Base classifier accuracy. In future, heterogeneous classifiers can be used for AdaBoost classification. Also, this model cannot predict the stages of CKD. In future we can detect the stages of CKD using heterogeneous classifiers for AdaBoosting ensembles.

REFERENCES

[1] Ahmed J. Aljaaf, Dhiya Al-Jumeily, Hussein M. Haglan, Mohamed Alloghani, “Early prediction of chronic kidney disease using machine learning supported by predictive analytic”, IEEE Congress on Evolutionary Computation CEC. DOI= 10.1109/CEC.2018.8477876, 2020.

[2] Dr. S. Vijayarani , Mr.S.Dhayanan,” Kidney Disease Prediction Using SVM and ANN Algorithms”, International Journal of Computing and Business Research (IJCBR), vol. 6, no. 2, 2019.

[3] Shaikhina, Torgyn, Lowe, Dave, Daga, Sunil, Briggs, David, Higgins, Robert, Khovanova, Natasha, “Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation”, Biomedical Signal Processing and Control, S1746809417300204 doi:10.1016/j.bspc.2017.01.012, 2017.

[4] Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel,”Heart Disease Prediction Using Machine learning and Data Mining Technique” International Journal Of Computer Science Communication . Vol. 7, No. 1, pp.129 – 137, DOI = 10.090592/IJCSC.2016.018, 2019.

[5] Tran, C. T., Zhang, M., Andreae, P., Xue, B., & Bui, L. T, “Multiple Imputation and Ensemble Learning for Classification with Incomplete Data”, Intelligent and Evolutionary Systems, 401–415. DOI = 10.1007/978-3-319-49049-6_29, 2016.

[6] Saha, Anik, Saha, Abir, Mittra, Tanni,”Performance Measurements of Machine Learning Approaches for Prediction and Diagnosis of Chronic Kidney Disease (CKD)”, ACM International Conference on Computer and Communications Management, 200–204. DOI=:10.1145/3348445.3348462, 2019.

[7] T. R. Mahesh , V. Dhilip Kumar, V. Vinoth Kumar, “AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease”, Computational Intelligence and Neuroscience. Volume 2022, Article ID 9005278, 11 pages DOI=10.1155/2022/9005278, 2022.

[8] Ebrahime Mohammed Senan , Mosleh Hmoud Al-Adhaileh, “Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques”, Journal of Healthcare Engineering .Volume 2021, Article ID 1004767, DOI=10.1155/2021/1004767, 2021.

[9] Segal, Zvi, Kalifa, Dan,Radinsky, Kira, Ehrenberg,Bar, Elad, Guy,Maor,Gal,Lewis,Maor, ”Machine learning algorithm for early detection of end-stage renal disease”, BMC Nephrology.21(1),518–. DOI=10.1186/s12882-020-02093-0, 2020.

[10] S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh, “Chronic Kidney Disease Prediction using Machine Learning Models” , International Journal of Engineering and Advanced Technology. ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019, DOI= 10.35940/ijeat.A2213.109119, 2019.

[11] 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, Vol. 10, No. 8, DOI = 10.14569 / IJACSA ,2019.0100813, 2019.

[12] https://www.kaggle.com/datasets/

[13] Shahinda Mohamed Mostafa Elkholy, Amira Rezk, Ahmed Abo El Fetoh Saleh,”Early Prediction of Chronic Kidney Disease Using Deep Belief Network”, IEEE Access, Volume:9,135542 –135549,DOI= 10.1109/ ACCESS .2021 .3114306, 2021.

[14] Subasi, A., Alickovic, E., & Kevric, J, “Diagnosis of Chronic Kidney Disease by Using Random Forest”, CMBEBIH 2017, 589–594. doi:10.1007/978-981-10-4166-2_89.

[15] Polat, H., Danaei Mehr, H., & Cetin, A, “Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods”, Journal of Medical Systems, 41(4). doi:10.1007/s10916-017-0703-x,2017.