Kidney Diseases Prediction Using Hybrid Ensemble Learning

Tanush Shetty[1], Dharmik Timbadia[2]

[1]Undergraduate Student, Rajiv Gandhi Institute of Technology, Andheri, Mumbai
[2]Graduate Student, Boston University, Boston, Massachusetts, USA

---***---

Abstract - *Chronic kidney disease is a major obstacle for health care infrastructure all over the world consuming a large amount of health care budgets, mainly affecting poor countries. Machine learning plays a vital role in analyzing vast amount of medical data and to solve difficult problem of early prediction of diseases. This study is aimed to develop a chronic kidney disease prediction using machine learning techniques to predict the presence of the disease. The prediction of kidney disease is done in three phases: feature selection process to select those features which aid most in prediction. The second phase is applying the machine learning algorithms, AdaBoost, KNN & Random Forest in which the data will be trained and tested. The final phase is the user interface in which the user will enter his health related information and the machine learning models will predict whether the user will have kidney disease or not.*

Key Words: Chronic Kidney Disease, Machine Learning, Random Forest, Feature Selection.

1. INTRODUCTION

Non-Communicable Diseases (NCDs), especially chronic kidney diseases (CKD), cardiovascular, hypertension, and diabetes mellitus have now taken a place of communicable diseases that become a serious public health issue and economic cost issue worldwide, consuming a high percentage of health care budgets. CKD is a chronic disease that has significantly contributed to increased morbidity, mortality, and an admission rate of patients throughout the world.

Chronic kidney disease is a condition in which the kidney structure is uneven or function reduces for 3 months and more with a reduced glomerular filtration rate. According to the reports, CKD has now become a fast spreading and fatal disease all over the globe. According to reports, the yearly life loss of CKD increased by 90% and it is known as the 13th leading death cause in the world. Currently, 850 million people throughout the globe are likely to have kidney diseases from different factors, from which at least 2.4 million die per year and now it is the 6th fastest-growing cause of death globally. [1]

The associated extent of the increased danger of clinical occurrences, which makes it a severe public health condition globally, is affiliated with chronic kidney disease. Even though it is widely accepted that CKD has significant interactions with magnified hazards of end-stage excretory organ disease, vessel occurrences and all-cause mortality, there is still a lack of accurate information on individual patients. Excretory organ damage refers to a condition that allows the capacity of the kidney to be reduced by a considerable decrease in the vessel filtration rate (GFR). The kidneys operate as filters to remove the waste products from the blood in different tiny blood vessels. In certain cases it decomposes and kidneys lose their capacity to distinguish nutrients, which ends in nephropathy. CKD has no underlying cause, but it generally becomes irreversible and can cause severe health problems.

A concerning 195 million girls are laid low with CKD within the world and it's presently the eighth leading reason behind death among women with around 600,000 individuals' dying of this illness annually. Between 2011 and 2012. There have been 9 million adults with CKD in England as registered within the Quality Outcomes Framework (QOF).

1.1 LITERATURE SURVEY

Chronic Kidney Disease is a progressive kidney disease that kills millions of people silently all over the world regardless the age and sex. CKD disease is a condition when the kidney damaged and unable to filter wastes as much as the normal kidneys because wastes are built up in the kidneys for several time. This can also cause other complications such as cardiovascular disease (CVD), anemia, and bone disease. It is a decreased function of the kidney over several years; it often goes undetected and undiagnosed until the disease is a well-advanced stage. From different report estimations, it indicated that the CKD prevalence is around 8 – 16 % globally, which shows it is becoming one of the serious public health problems.

In most developed countries, CKD is mostly related to old age, diabetes, hypertension, obesity, cardiovascular disease, and diabetic glomerulosclerosis, and hypertensive nephroscerosis, whereas, in developing countries, the common causes of CKD are glomerular and tubulointerstitial diseases which result from infections and exposure to drugs and toxins. Moreover, in most developing countries low-level quality of life like lack of clean water, lack of appropriate diet is the main cause of CKD.

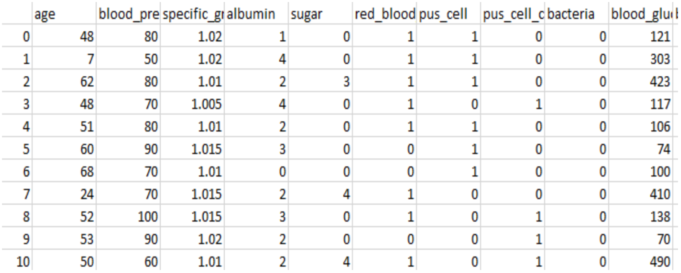
Our daily diet, what we eat and drink can damage our kidneys. Therefore, since getting a balanced diet is difficult in developing countries like Ethiopia, people suffer from kidney disease and other risk factors of this disease. Chronic kidney disease leads to kidney failure silently if do not detected early and accurately by practitioners. At the early stage of CKD, the patients may not have any symptoms. The only way to detect and identify the presence or absence of the CKD disease is urine analysis and blood test needed to know the kidney function and kidney damage. The treatment option is not affordable, on time and people have less knowledge about CKD in developing countries. CKD will grow to end-stage kidney failure slowly that is a very serious problem for which artificial filtering (dialysis) or a kidney transplant is needed.

In [3], authors conducted their research on the common risk factor in chronic kidney disease. They have stated that among the risk factor, aging was a significant predictor for renal failure in both males and females, but more prominent in males and stated that hypertension and diabetes mellitus, main causes among diseases for causing renal failure.

In [2], authors conducted their research to measure the prevalence and risk factors for chronic kidney disease (CKD) and diabetic kidney disease (DKD) in a Chinese rural population. According to their finding, they came up with age, gender, education, personal income, alcohol consumption, overweight, obesity, diabetes, hypertension, and dyslipidemia as prevalent risk factors.

2. DATASET

Dataset was derived from UCI. The data set contains 400 samples. In this CKD data set, each sample has 25 predictive variables (12 numerical variables and 13 categorical variables). Each class has two values, namely, CKD and not CKD. In the 400 samples, 250 samples belong to the category of CKD, whereas 150 samples belong to the category of not CKD. Normal and abnormal were coded as 1 and 0. [4] Present and not present were coded as 1 and 0. There is a large number of missing values in the data set, and the number of complete instances is 158.A sample dataset of 10 instances that has been used in this project has been shown:



	age	blood_pre	specific_gr	albumin	sugar	red_blood	pus_cell	pus_cell_c	bacteria	blood_glu
0	48	80	1.02	1	0	1	1	0	0	121
1	7	50	1.02	4	0	1	1	0	0	303
2	62	80	1.01	2	3	1	1	0	0	423
3	48	70	1.005	4	0	1	0	1	0	117
4	51	80	1.01	2	0	1	1	0	0	106
5	60	90	1.015	3	0	0	1	0	0	74
6	68	70	1.01	0	0	0	1	0	0	100
7	24	70	1.015	2	4	1	0	0	0	410
8	52	100	1.015	3	0	1	0	1	0	138
9	53	90	1.02	2	0	0	0	1	0	70
10	50	60	1.01	2	4	1	0	1	0	490

Fig 1 Sample Dataset

3. PROPOSED SYSTEM

The proposed system majorly focusing on the accuracy and the system betterment and overcoming the drawbacks of previous systems by developing a frontend based CKD detection solution. The proposed system also employs the feature selection method in order to select the most relevant and predictive features.

The classifiers were first established by different machine learning algorithms to diagnose the data samples. Among these models, those with better performance were selected as potential components. By analyzing their misjudgments, the component models were determined. [4] An integrated model was then established to achieve higher performance. As shown in above figure 2 first we have a preprocessing phase to convert dataset into appropriate format. Then we have feature selection phase to select features which are most useful in helping predict CKD. Then we use classification algorithms to develop a CKD prediction model. Then we have training and testing phase to evaluate performance of the model.

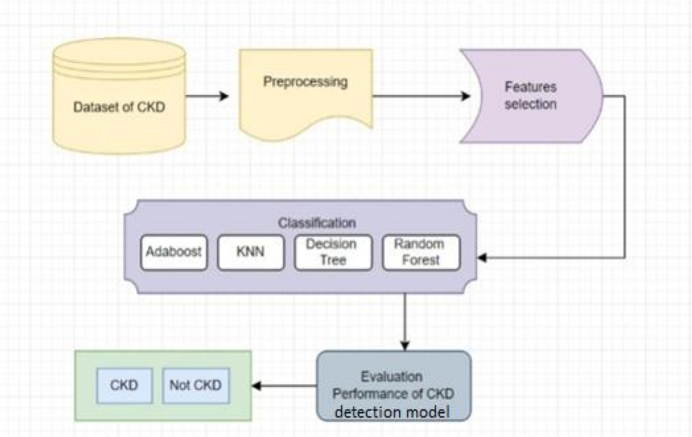


Fig 2 Proposed System

4. SYSTEM DESIGN

The system was designed used machine learning methods for chronic kidney disease detection. Various ensembles like boosting were used in the prediction of kidney disease. A frontend based CKD detection solution has been developed. Users can enter their test data such as blood pressure, creatinine levels and expect an output predicting if they have CKD or not. Supervised classification algorithms, AdaBoost, KNN and Random forest were used to develop a model to predict if someone has CKD or not. Python language was used to develop this system.

Pre-processing was done to make the data clean and appropriate for the machine learning models. In this study, the missing values have been handled using the mean replacement method and categorical values have been converted to binary.

AdaBoost combines multiple classifiers to extend the accuracy of classifiers. AdaBoost is an ensemble

methodology. AdaBoost classifier builds a powerful classifier by combining multiple poorly playing classifiers so you'll get high accuracy strong classifier. The essential idea behind AdaBoost is to line the weights of classifiers and coaching the information sample in every iteration specified it ensures the correct predictions of surprising observations. Any machine learning formula are often used as base classifier if it accepts weights on the coaching set. AdaBoost must meet 2 conditions:

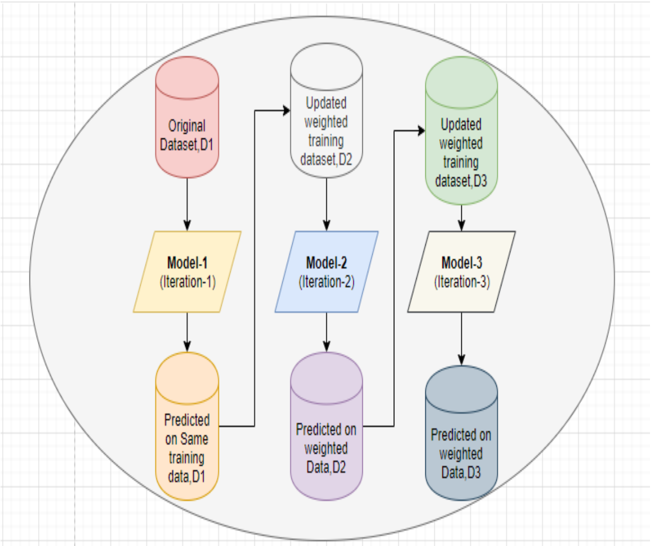

Original Dataset,D1
Updated weighted training dataset,D2
Updated weighted training dataset,D3
Model-1 (Iteration-1)
Model-2 (Iteration-2)
Model-3 (Iteration-3)
Predicted on Same training data,D1
Predicted on weighted Data,D2
Predicted on weighted Data,D3

Fig 3 AdaBoost Classifier Algorithm and Flow Chart

In the above figure 3 it is shown that:

1. The classifier should be trained iteratively on various weighted training examples.

2. In each iteration, it tries to provide an excellent fit for these examples by minimizing training errors.

The final equation of the AdaBoost algorithm can be represented as follows:

$$F(x) = sign\left(\sum_{m-1}^{M} \theta_m f_m(x)\right),$$

where f_m denotes for the m_th weak classifier and theta_m is the corresponding weight. It is exactly the weighted union of M weak classifiers. The whole procedure of the AdaBoost algorithm can be briefly explained as follows.

Given a dataset containing n points, where

$$x_i \in R^d, y_i \in \{-1,1\}.$$

Here -1 represents the negative class whereas 1 represents the positive one. Initialize the weight for every data point as:

$$w(x_i, y_i) = \frac{1}{n}, i = 1, \dots, n.$$

For iteration m=1..., M:

1. Fit weak classifiers to the data set and select the one with the lowest weighted classification error.

2. Compute the weight for the weak classifier.

3. Update the weight for each data point.

After M iteration, we can get the final prediction by summing up the weighted prediction of each classifier. [5]

- **K Nearest Neighbour (KNN)**

K Nearest Neighbour is a non-parametric algorithm, meaning it does not make any assumption on the data.

It is called a lazy learner algorithm since it does not learn from the training set at once instead it stores the dataset and during classification, it performs operations on the dataset.

Example: Suppose, we have a picture of an animal that looks similar to cat and dog, but we want to know whether it is a cat or dog. So for this identification, we can use the K Nearest Neighbour algorithm, as it works on a similarity measure. Our KNN model will find the corresponding features of the new data set to the cats and dogs pictures and based on the most corresponding features it will put it in cat or dog category.

- **Random Forest Classifier:**

Random forest is a machine learning algorithm formulated on ensemble learning. Ensemble learning is a category of learning where you join various types of algorithms or same algorithm many times to form a better prediction model. The random forest algorithm integrates many algorithm of the same category i.e. many decision trees, which leads to a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for regression as well as classification tasks.

The following are the fundamental steps involved in performing the random forest algorithm:

1. Pick N random features from the dataset.
2. Form a decision tree based on these N records.
3. Choose the number of trees you want in your algorithm and repeat step 1 and 2.

4. In case of a regression problem, for a new record, every tree in the forest predicts a value for Y (output). The final value can be computed by taking the mean of all the values predicted by all the trees in forest. Or, in case of a classification problem, every tree in the forest predicts the type to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote [6].

There are many advantages of using Random Forest Classifier as the base algorithm for machine learning models in this project which are as follows:

- The random forest algorithm is not biased, since, there are many trees and every tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd"; therefore, the overall biasedness of the algorithm is scale down.

- This algorithmic program is extremely stable. Even though brand-new information is introduced within the knowledge set the general algorithm program isn't affected. Since new data could impact one tree, however it's highly impossible for it to impact all the trees.

The random forest algorithm works better when you have both kind of data which is categorical and numerical features.
The random forest algorithm also works well when data has null values, or has not been scaled well.

5. RESULTS

Given below are the results of system. Confusion matrices of algorithms used that were generated while testing the model are included.

5.1 AdaBoost Algorithm

Below are the confusion matrix as well as classification report for AdaBoost algorithm.

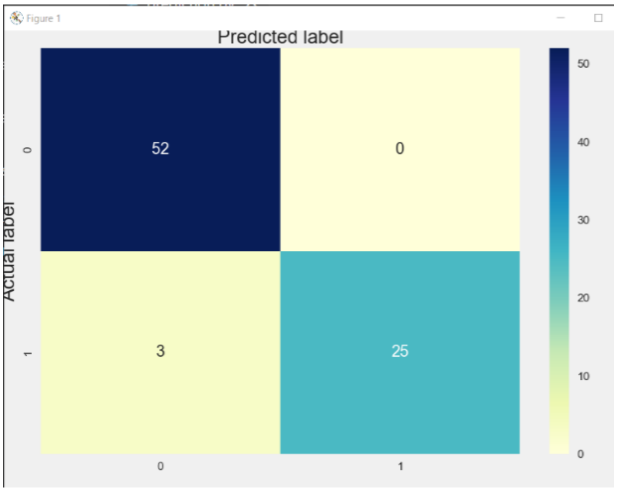


Fig 4 Confusion matrix of AdaBoost algorithm

In the above diagram (AdaBoost confusion matrix) we can see that it is predicted that 52 people do not have CKD which is correct but for 3 subjects, it has predicted wrongly, on the other hand, it has predicted correctly for 25 subjects and incorrectly for none.

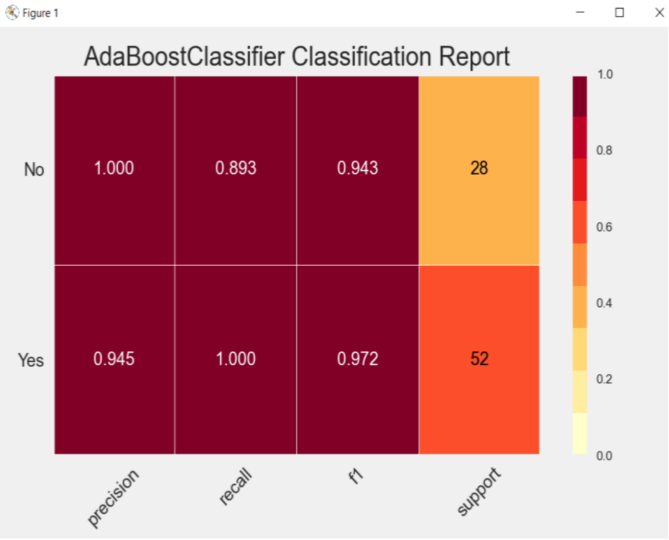


Fig 5 Classification report of AdaBoost algorithm

The above diagram tells us about the classification of the AdaBoost algorithm. As we can see that we are getting a good precision of 94.5 % for Yes and 100 % for No. Also, for recall the accuracy is good which is 100 % for Yes and 89.3 % for No.

5.2 KNN Algorithm

Below are the confusion matrix as well as classification report for KNN algorithm.

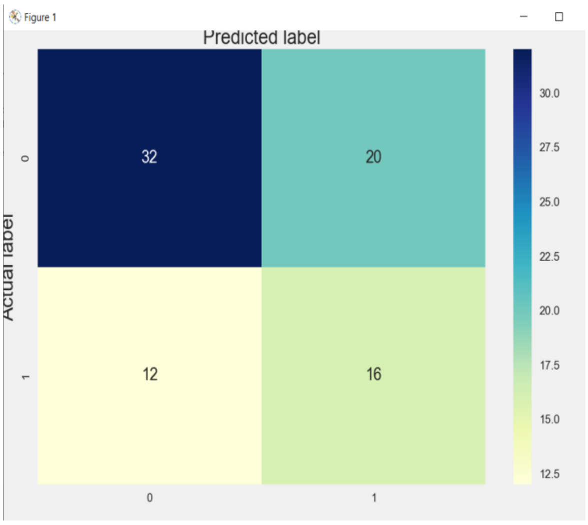
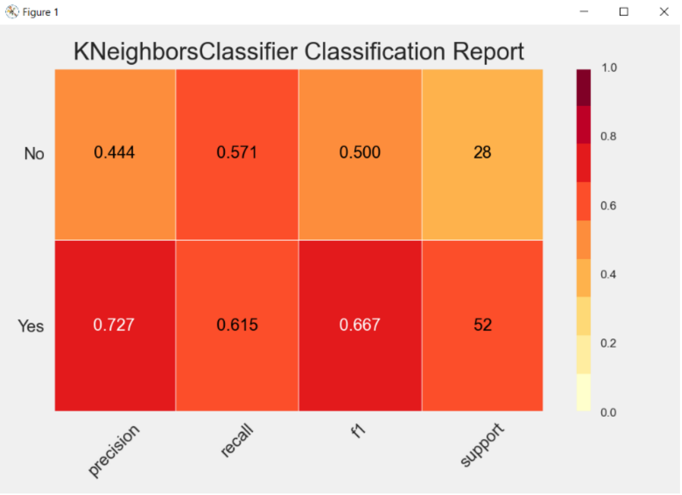


Fig 6 Confusion matrix of KNN algorithm

In the above diagram (KNN confusion matrix) we can see that it is predicted that 32 people do not have CKD which is correct but for 12 subjects, it has predicted wrongly, on the other hand, it has predicted correctly for 16 subjects and incorrectly for 20 subjects.



	precision	recall	f1	support
No	0.444	0.571	0.500	28
Yes	0.727	0.615	0.667	52

Fig 7 Classification report of KNN algorithm

The above diagram tells us about the classification of the KNN algorithm. As we can see that we are getting a good precision of 72.7% for Yes and 44.4% for No. Also, for recall the accuracy is good which is 61.5% for Yes and 57.1% for No

5.3 Random Forest Algorithm

Below are the confusion matrix as well as classification report for Random Forest algorithm

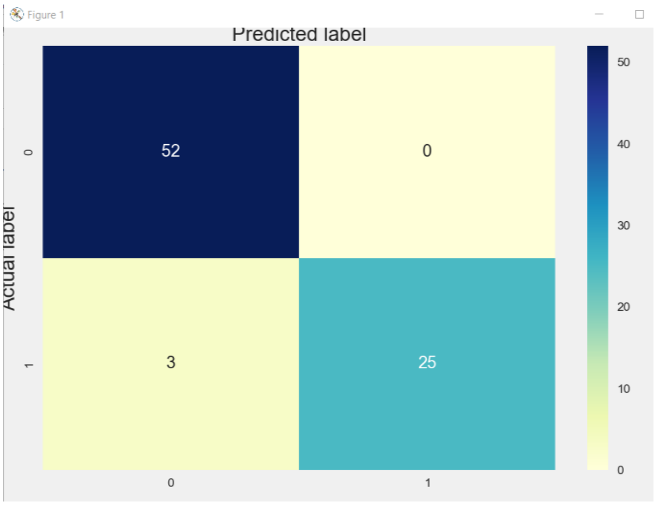
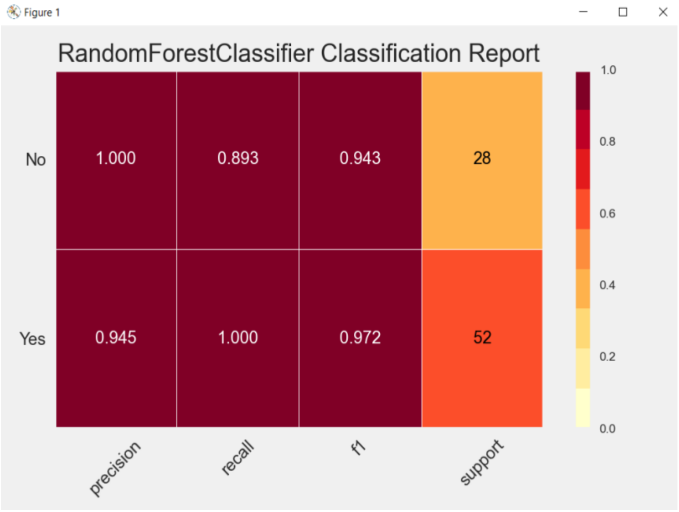


Fig 8 Confusion matrix of Random Forest algorithm

In the above diagram (Random Forest confusion matrix) we can see that it is predicted that 52 people do not have CKD which is correct but for 3 subjects, it has predicted wrongly, on the other hand, it has predicted correctly for 25 subjects and incorrectly for none.



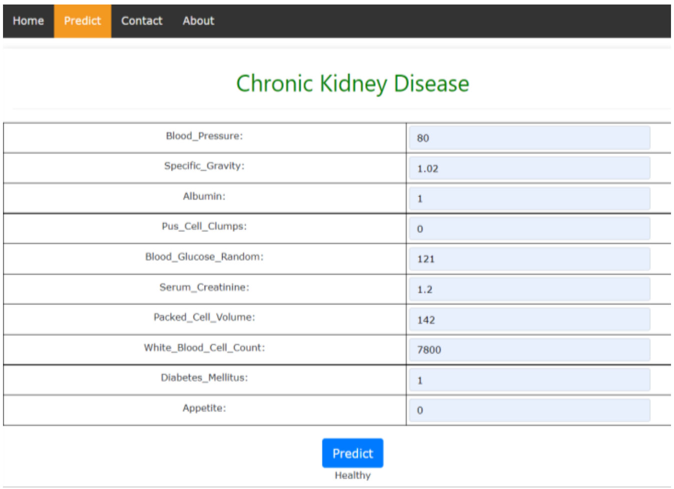
	precision	recall	f1	support
No	1.000	0.893	0.943	28
Yes	0.945	1.000	0.972	52

Fig 9 Classification report of Random Forest algorithm

The above diagram tells us about the classification of the Random Forest algorithm. As we can see that we are getting a good precision of 94.5% for Yes and 100% for No. Also, for recall the accuracy is good which is 100% for Yes and 89.3% for No

5.4 User Interface

As shown in the below image we can see the user Interface which we designed for testing the user input. We have built this UI using Flask micro framework. In the UI part, the user enters the values which are required to diagnose kidney disease by the developed trained model.

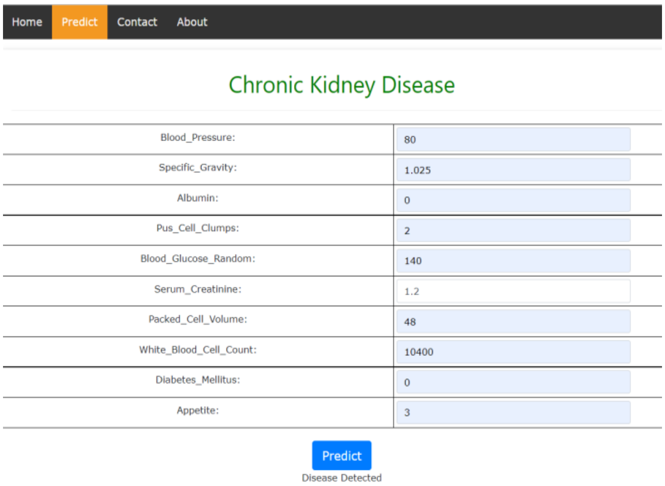


Chronic Kidney Disease

Blood_Pressure:	80
Specific_Gravity:	1.02
Albumin:	1
Pus_Cell_Clumps:	0
Blood_Glucose_Random:	121
Serum_Creatinine:	1.2
Packed_Cell_Volume:	142
White_Blood_Cell_Count:	7800
Diabetes_Mellitus:	1
Appetite:	0

Healthy

Fig 10: Healthy prediction
(Prediction page of website)



Chronic Kidney Disease

Blood_Pressure:	80
Specific_Gravity:	1.025
Albumin:	0
Pus_Cell_Clumps:	2
Blood_Glucose_Random:	140
Serum_Creatinine:	1.2
Packed_Cell_Volume:	48
White_Blood_Cell_Count:	10400
Diabetes_Mellitus:	0
Appetite:	3

Disease Detected

Fig 11 Kidney disease detected
(Prediction page of website)

6. CONCLUSION

Early prediction is very crucial for both the experts and the patients to prevent and slow down the progress of chronic kidney disease to kidney failure. In this study, the chronic kidney disease prediction, using the machine-learning technique is proposed with the deployment of the model in order to help the experts to diagnose the disease quickly.

Specific machine learning methods were evaluated by literature. We have used ensembles like boosting in the prediction of kidney disease using numeric data to achieve better accuracy results. Testing data sets setting together strategies were extremely efficient.

The proposed study also employs the feature selection method in order to select the most relevant and predictive features.

Preprocessing was done to make the data clean and appropriate for the machine learning models using relevant libraries. In this study, the missing values have been handled using the mean replacement method and categorical values have been converted to binary.

REFERENCES

[1] Prof. Sachin Wakurdekar, Munish Goswami, Anupama Sharma and Harshit Gupta, “Chronic Kidney Disease Detection Using RFA”, IJIRT, vol 8, Issue 3, pp. 714-718, Aug. 2021.

[2] Jiayu Duan, Chongjian Wang et al., “Prevalence and risk factors of chronic kidney disease and diabetic kidney disease in Chinese rural residents: a cross-sectional survey”, Nature, 2019, https://doi.org/10.1038/s41598-019-46857-7

[3] Roshni PR*, Mahitha Mathew, “Risk Factors Associated With Chronic Kidney Disease: An Overview”, Int. J. Pharm. Sci. Rev. Res., 40(2),; Article No. 47, pp. 255-257, September – October 2016

[4] Jiongming Qin, Lin Chen, et al., “A Machine Learning Methodology for Diagnosing Chronic Kidney Disease”, IEEE Access, vol 8, pp. 20991-20993, 2020

[5] T.-K. An, M.-H. Kim, A new diverse AdaBoost classifier, 2010 International Conference on Artificial

Intelligence and Computational Intelligence, IEEE, 2010, pp. 359-363.

[6] M.J.I.J.o.R.S. Pal, Random forest classifier for remote sensing classification, 26 (2005) 217-222.

[7] V. Jha , G. Garcia-Garcia , K. Iseki , Z. Li , S. Naicker, B. Plattner, R. Saran, A. Y. Wang, C.W. Yang (2013),”Chronic kidney disease: global dimension and perspectives”, The Lancet.

[8] Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016). Chronic Kidney Disease analysis using data mining classification techniques. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 300-305)

[9] L. Xun, Wu Xiaoming, Li Ningshan and Lou Tanqi, "Application of radial basis function neural network to estimate glomerular filtration rate in Chinese patients with chronic kidney disease," 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, 2010, pp. V15-332-V15-335.

[10] A. Salekin and J. Stankovic, (2016) “Detection of chronic kidney disease and selecting important predictive attributes,” IEEE International Conference on Healthcare Informatics.