International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V69I7P110

Volume 69 Issue 7, 70-75, July 2021

Original Article

Machine Learning in Healthcare for Chronic Kidney Disease Prediction

Pranjali Kasture

Assistant Professor, Department of IT, Thakur College of Engineering & Technology, Mumbai, India.

Received Date: 09 June 2021
Revised Date: 12 July 2021
Accepted Date: 22 July 2021

Abstract *- Chronic Kidney Disease (CKD) is a common yet deadly disease that can be difficult to detect at an early stage as it doesn't show too many symptoms. The proposed work is to develop and validate predictive models for the progression of CKD. The main outcome of this is to measure kidney failure, defined as the need for dialysis or pre-emptive kidney transplantation. The model will suggest the patient the way to maintain a healthy lifestyle as well as facilitate the doctor to visually represent the risk and severity of the disease and how to go about the future treatment.*

Keywords *- CKD, GFR, Machine learning, ANN.*

I. INTRODUCTION

Engineers and Researchers in the medical industry have tried to design and build machine learning algorithms as well as models to predict CKD at an early stage. The challenge is that the amount of data generated in the healthcare industry is enormous as well as quite complex, so it becomes difficult to analyze such data. But with the help of data mining technologies, we can effectively convert this data into the form of information, and then this information can be finally converted into knowledge by using machine learning algorithms.

Kidney disease severity can be classified by using an amalgamation of estimated glomerular filtration rate (GFR), age, diet, existing medical conditions, and albuminuria, but more accurate information regarding risk for progression to kidney failure is required for clinical decisions about testing, treatment, and referral.

The objective of this model is to develop and validate predictive models for the progression of CKD. The main outcome of this will be to measure kidney failure, defined as the need for dialysis or pre-emptive kidney transplantation.

Thus the model is developed using routinely obtained laboratory tests and other above-mentioned parameters to accurately predict progression to kidney failure in patients with CKD stages 1 to 5. This model also suggests the patient the way to maintain a healthy lifestyle as well as facilitate the doctor to visually represent the risk and severity of the disease and how to go about the future treatment. By applying ANN, data mining algorithms on the data set, one can find patterns in the data set, and the future possibility of some disease becoming a risk can be prevented well before.

The motive of the proposed model is to predict if the patient is suffering or can suffer from CKD in the future if he/she continues a certain lifestyle. This information can then be used for finding the stage of kidney diseases using eGFR (Estimated glomerular filtration rate), which indirectly helps the doctor to plan the treatment accordingly. eGFR - Estimated glomerular filtration rate measures the level of kidney function and determines the stage of kidney disease.

A. Chronic Kidney Disease

The main function of the kidneys is to filter the blood in the body. Kidney disease is a silent killer as a lot of functionality of the kidney can be lost without even having any symptoms or problems. CKD is a gradual decrease in renal function over a period of several months or years. Diabetes and high blood pressure are the most common causes of chronic kidney disease. CKD is a serious health concern and affects people all around the world. There can be grave consequences of not getting proper treatment for CKD and thus affects the people who can't afford the treatment. Glomerular Filtration Rate (GFR) is the best test to measure your level of kidney function and determine your stage of chronic kidney disease. It can be calculated from the results of blood creatinine, age, gender, and other factors. It's always better if an earlier disease is detected. Then there is a better chance of preventing its progression.

B. Background

Today the lifestyle of a common man is getting tougher day by day, and he is not having a sufficient amount of time to devote to fitness. Because of this, the health gets affected adversely. Therefore one may suggest that the person should do regular check-ups, but not all can afford these check-ups. If we find an alternative to this, then a lot of cost-saving can be done. Not only this but from an economic point of view, if the citizens of a country are not healthy, then it will surely impact the growth of the country in a negative manner. Thus there is a need for economical and effective ways for checking the health of the citizens who could possibly suffer from greater ailments if they are not checked on time.

C. Importance

Many patients that approach a doctor for kidney-related issues often refused to accept the advice of the doctor that the reason for their medical issue is their present lifestyle. If the patients are shown that the people suffering from CKD were also having the same kind of lifestyle that they are presently having, through a graphical manner, then it will be a lot more convincing to the patient. Also, they can be given a diet plan that will ensure the improvement of their present ailment. This will help the doctors in treating many patients in a short span of time and preventing them from getting Chronic Kidney Disease in the last stage. The result that will be generated will be very quick. Therefore one does not need to wait for the actual detailed medical reports; also, it will be very cost-effective, and ultimately a lot of patients suffering from kidney-related disease can be examined and prevented from getting CKD converted to the last stage where a kidney transplant is the only option.

D. Objectives & scope

The major objective of the project is to study the various factors/attributes responsible for CKD and then to design a model that can predict the unknown sample as possible future CKD patient or not. Also, the objective of the project is to try to represent the percentage possibility of the patient getting CKD in the future in a graphical manner for easy readability of the user. The early prediction of the high vulnerability of the patients can help save a lot of lives and also the cost and time that might be lost in the medical treatment process.

The following are the main functions of the proposed model:

- To collect data from the user(patient) from a basic test that will contain some tests for creatinine, urine albumin, etc.
- To generate a Graphical report on the possibility of the patient having CKD.
- To provide a generic diet solution that can help the patient improve current kidney disease(for which the patient might have come for a check-up).
- To help save the lives of people and thus safeguarding the valuable human resource of the nation.

The concept that has been used is applying machine learning techniques in the field of healthcare by which an immense potential can be unleashed. In the proposed model, specifically, Data Mining techniques are applied in predicting Chronic Kidney Disease (CKD). As far as the current model of CKD is concerned, the database can be further enriched with more records. That is, more data of the patients suffering from CKDcan be added (but the data should be reliable), which will improve the accuracy of the prediction system. Also, research can be done with the help of medical experts to find more factors that are responsible for causing the CKD and then incorporating those factors as attributes in the dataset which will also improve the accuracy of the prediction system furthermore.

II. LITERATURE SURVEY

There has been extensive work in the medical domain with the use of machine learning algorithms which have helped in the analysis and interpretation of medical data (in large amounts) of the patients effectively. As far as [1] is considered, it finds that MLP is the best algorithm as it is adaptive and can handle complicated predictions. The main drawback of the research work is that the attribute blood pressure is not reliable as medication of the patient is not considered, which might be affecting the blood pressure values. In [2], the dataset accuracy increases as the number of records increase, and the IBK algorithm performs the best among the five algorithms that have been applied to the dataset. The focus of this work is mainly on diabetic patients, and the accuracy of the IBK algorithm has been found to be 98.25%.In [3], the model is a GUI-based solution for the doctor to cross-verify its diagnosis as well as alert the patient if there is a suspicion of CKD. In [4], various data mining algorithms have been implemented to classify CKD(chronic kidney disease) patients and NON-CKD patients, and the paper concludes by saying that there is no one algorithm that can be used on all datasets and provide us with the same accuracy. In [5], classification algorithms like Backpropagation Neural Network, Radial Basis Function, and Random Forest have been considered for predicting CKD. Different measures have been used for the model evaluation like Kappa, Accuracy, Sensitivity, and Specificity. The deduction from the experimental result was that the Radial Basis Function has better accuracy for predicting chronic kidney disease, and it attains an accuracy of 85.3%.In [6], multilayer perceptrons are having better performance than another neural networks. The performance of the classifier is evaluated, and the results are analyzed. The multilayer perceptron classifier gave good accuracy. In [7], 14 different attributes related to CKD patients have been analyzed, and prediction of accuracy for Decision tree and Support Vector Machine algorithm has been made. The SVM algorithm has the best accuracy between them. The drawback of this study is that the strength of the data is not high because of the size of the dataset and the missing attribute values.

III. METHODOLOGY USED

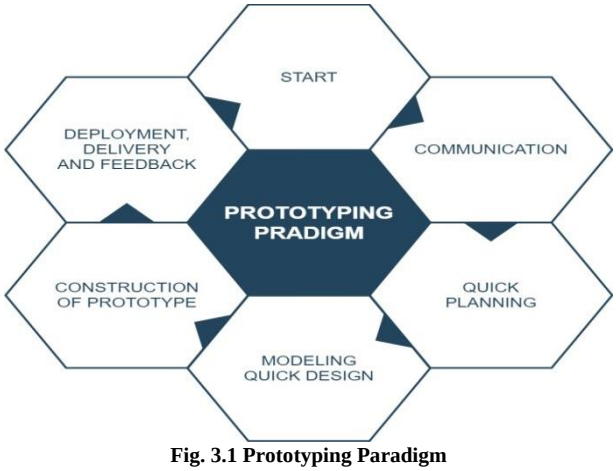

Fig. 3.1 Prototyping Paradigm

The fig. 3.1 depicts a prototyping model in which a prototype is created and then tested. If there are any changes that are needed, then those changes are worked upon until a satisfactory final product is created.

Advantages of Using Prototype Model

- Prototyping makes the requirements more clear and the system more transparent.
- The customers get to see the partial product early in the life cycle. This ensures a greater level of customer satisfaction and comfort.
- New requirements can be easily accommodated as there is scope for refinement.

The methodology that has been used for the proposed model is that first, the best algorithm will be selected for the classification of the data as CKD or NOTCKD, then. If the classification is CKD, then the CKD-EPI equation will be used to calculate the eGFR value. Using this eGFR value, we will be able to calculate the current stage of the patient.

IV. PROPOSED WORK

The use of very detailed and costly tests for checking the functioning of the kidney is not possible for patients who can't afford them. But we must be able to give them the proper health care for their kidneys. For this, we can make use of predictive models in machine learning to try to find which patient is in which stage, and from the stage, we can estimate the possibility of the patient getting chronic kidney disease in the future. This cost-effective method will improve when more and accurate data is fed in the training model, and thus people who can't afford expensive healthcare tests can also be given health care advice for their kidneys.

A. Database

The database used for the proposed model is taken from the UCI repository. The dataset is of Apollo hospital in India. The dataset has 400 instances(250 CKD, 150 notckd). The number of attributes are 24 + class = 25 (11 numeric ,14 nominal) as mentioned table 4.1. This dataset has been used for training the model. Apart from this, another real-time dataset has been used to calculate the accuracy of the model. This real-time data has been collected from different dialysis centers. This helps to calculate the accuracy of the model.

Table 4.1

Attribute Information
1. Age(numerical) age- in years 2. Blood Pressure(numerical) bp- in mm/Hg 3. Specific Gravity(nominal) sg - (1.005,1.010,1.015,1.020,1.025) 4. Albumin(nominal) al - (0,1,2,3,4,5) 5. Sugar(nominal) su - (0,1,2,3,4,5) 6. Red Blood Cells(nominal) rbc - (normal, abnormal) 7. Pus Cell (nominal) pc - (normal, abnormal) 8. Pus Cell clumps(nominal) pcc - (present, notpresent) 9. Bacteria(nominal) ba - (present, notpresent) 10. Blood Glucose Random(numerical) bgr-in mgs/dl 11. Blood Urea(numerical) bu -in mgs/dl 12. Serum Creatinine(numerical) sc- in mgs/dl 13. Sodium(numerical) sod -in mEq/L 14. Potassium(numerical) pot -in mEq/L 15. Hemoglobin (numerical) hemo -in gms 16. Packed Cell Volume(numerical) 17. White Blood Cell Count(numerical) wc- in cells/cumm 18. Red Blood Cell Count(numerical) rc- in millions/cmm 19. Hypertension(nominal) htn - (yes,no) 20. Diabetes Mellitus(nominal) dm - (yes, no)33 21. Coronary Artery Disease(nominal) cad - (yes, no) 22. Appetite(nominal) appet - (good, poor) 23. Pedal Edema (nominal) pe - (yes, no) 24. Anemia (nominal) ane - (yes, no) 25. Class (nominal) class - (ckd, notckd)

B. Block diagram

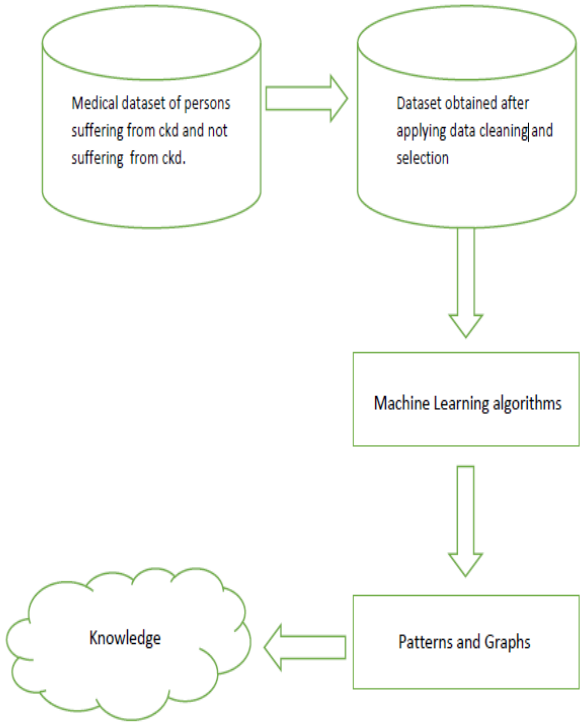

Fig. 4.2.1 Block diagram

The block diagram in fig. 4.2.1, the dataset of the medical dataset of the persons suffering from CKD or not is first taken. Then some data pre-processing is done to clean the data and handling the missing data. After the completion of this step now the data is cleaned and ready. Now we can apply machine learning algorithms on this dataset, and whatever patterns or graphs are generated will be observed, and then accuracy will be noted, and therefore knowledge will be generated as the final outcome.

C. Flow chart

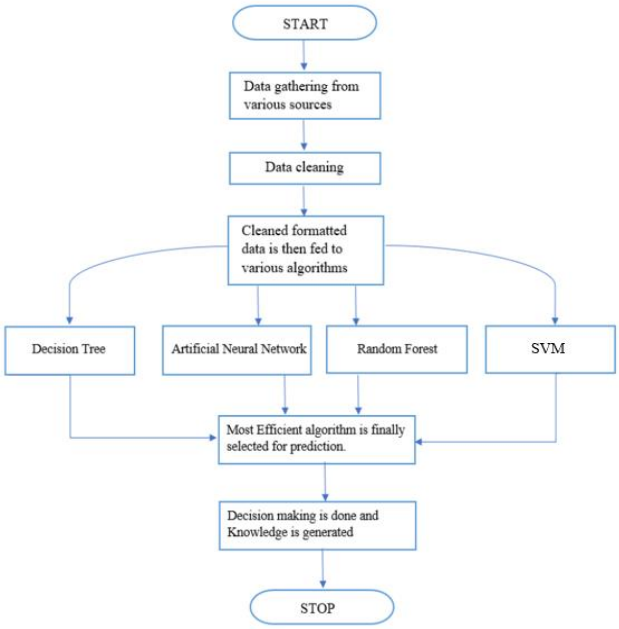

Fig. 4.3.1 Flow Chart

In the fig. 4.3.1 a flow chart of the proposed model has been depicted. First, the medical data will be collected from various sources related to the kidney. Then the dataset will be cleaned, that is missing data will be handled using various techniques. After that, the cleaned data will be fed into various algorithms, and the accuracy of the algorithms will be compared. The algorithm having the best accuracy will be finally selected for the proposed model.

V. RESULTS

ANN algorithm is applied on the given dataset, which gives fair accuracy. The attributes have not been modified. We have retained only a few columns for current analysis that is, columns_to_retain = ['sg','al','sc','hemo','pcv','wbcc','rbcc','htn','class']. Dataset has been split as 80% training and 20% testing & shuffle. We can do much more to improve the data set, and the accuracy obtained by improving the attributes can be used. A model uses Artificial Neural Network to classify the patient as CKD or NON-CKD and then tried to find out the stage of the patient suffering from CKD using the CKD-EPI equation.

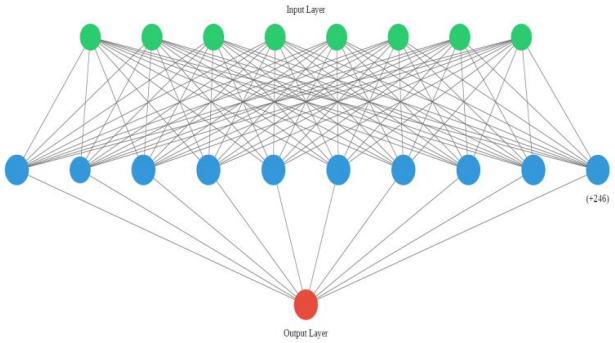

Fig. 5.1 ANN Output for CKD Model

Fig 5.1 shows an ANN visualization of the proposed model. In the Input layer, we have all the attributes that contribute most significantly to the prediction, like specific gravity, albumin, serum creatinine, hemoglobin, etc. Then we have the hidden layer, which will do all of the computation work. Finally, we have the output layer, which gives the prediction result.

The CKD-EPI equation for estimating GFR has been proposed to be more accurate than the MDRD equation. This is significant when the GFR is high. Also, it shows less bias, improved precision, and greater accuracy. This is the reason why the CKD-EPI equation has been selected in the proposed model.

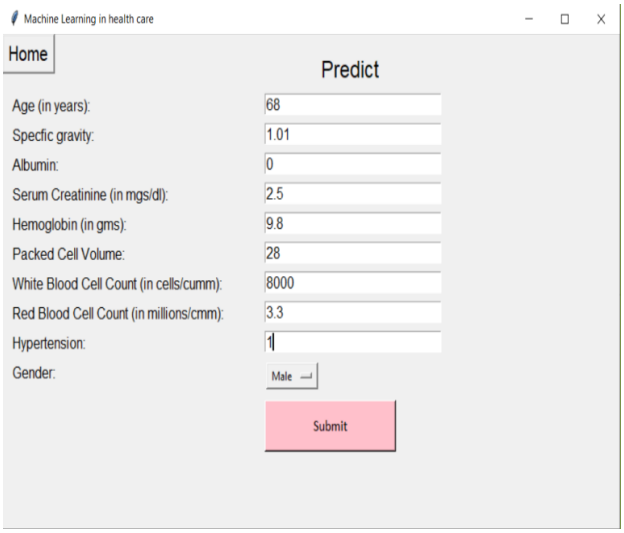

Fig. 5.2 Input GUI for the proposed model

The fig. 5.2 shows the GUI of the proposed model. The user will have to enter the values such as age, specific gravity, albumin, serum creatinine, hemoglobin, etc. Finally, after entering all the values, the user has to click on the submit button to generate the result.

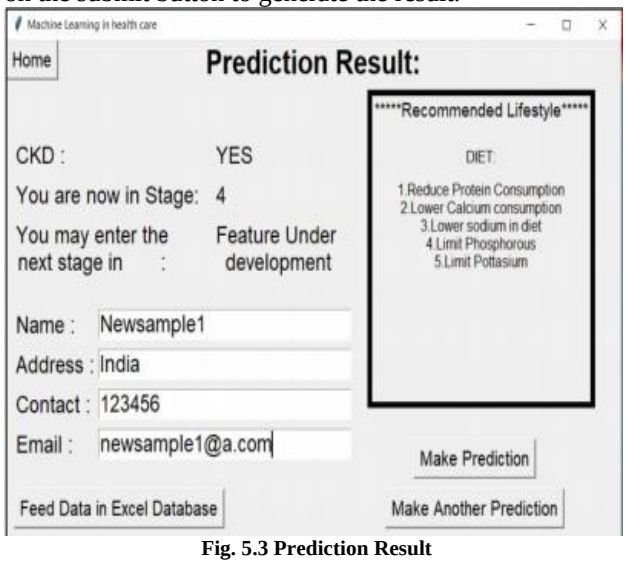

Fig. 5.3 Prediction Result

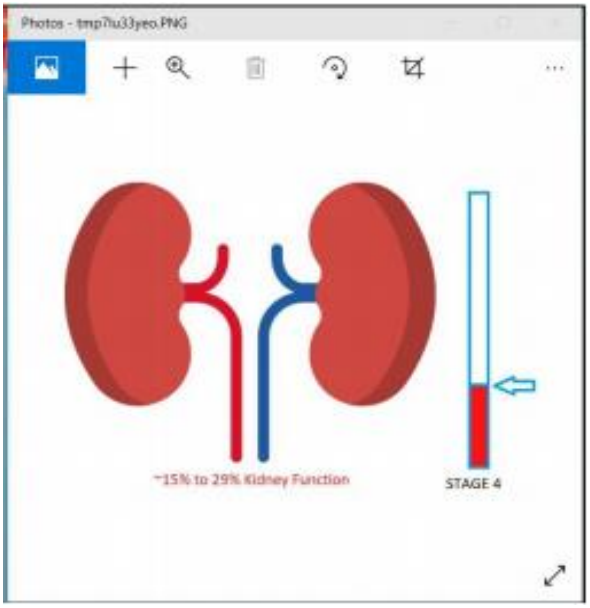

Fig. 5.4 Prediction Result for Stage

The fig. 5.3 and fig. 5.4 shows the GUI of the prediction result of the proposed model. CKD yes or no is represented along with the stage. The eGFR value is used for calculating the current stage of the patient. Also, there is a Diet recommendation plan given for the patient

We have used Decision tree, Support Vector Machine, and the Random Forest algorithm for model selection and found that the accuracies of these algorithms was not good enough for the final model selection and also what was noted that although the decision tree might be having good accuracy, it may not perform well on an entirely different dataset. Also there may be problems of overfitting in decision trees. As we wanted to select an algorithm that is not performing well only on one particular dataset, rather the selected algorithm must perform well on any other dataset. Hence the finally, Artificial Neural Network is selected as the final Model. ANN model was tested on the different datasets to check the accuracy. ANN model performs well on both testing data set and real-time different/new datasets.

Table 5.1

Algorithm	Accuracy
Decision Tree	95.83
SVM	90.24
Random Forest	90.24
ANN	97.56

Table 5.1 shows the comparison of different machine learning algorithms like Decision trees, Support vector machines, random forest, and Artificial Neural networks (ANN) with respect to accuracy. It is observed that ANN works well with the highest accuracy.

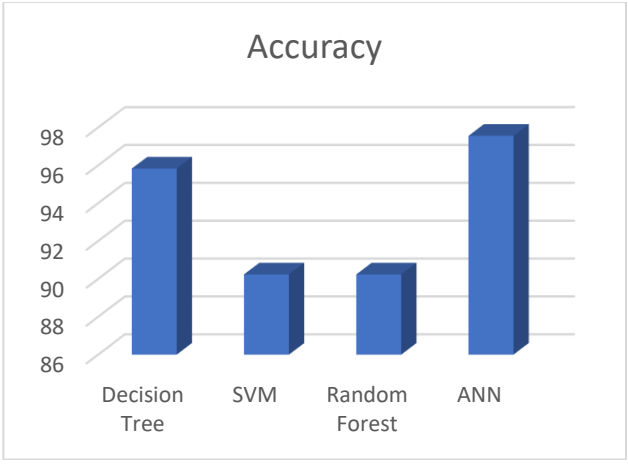

Fig. 5.4 Accuracy Bar Chart

In the fig. 5.4 an accuracy bar chart is represented depicting the accuracies of the algorithms that were examined before selecting the final algorithm. As it can be seen that ANN has the highest accuracy when compared to other algorithms.

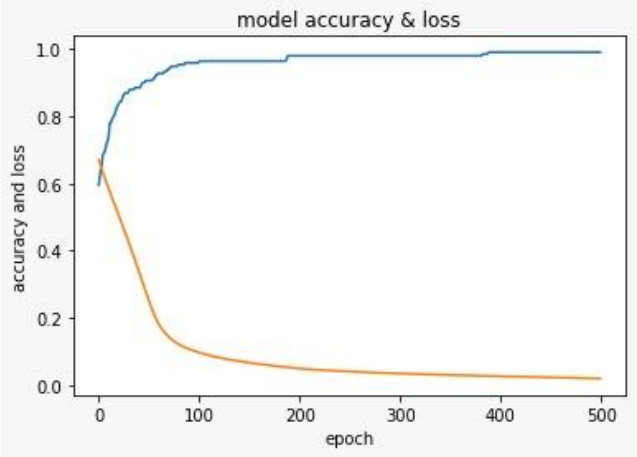

Fig. 5.5 Model Accuracy and Loss vs. epoch graph

In the fig. 5.5, we can state that, as the epochs increase, the accuracy of the model increases, and the loss decreases. Thus if we increase the number of epochs, then the model accuracy will also increase.

VI. CONCLUSION

Chronic Kidney disease is still deadly and very expensive to treat. Machine learning techniques in healthcare helps in cost savings for patients and time saving for the doctors. The doctor can use the model for treating the patients who come up with the kidney-related disease. The model will classify the patient according to results like having chronic kidney disease or not. The accuracy of the model is tested with a real-time dataset also. The system provides a low-cost solution for the problem of chronic kidney disease (CKD), and if detected early, the patient can be treated and prevented from getting into the last stage.

REFERENCES

[1] Ahmed J. Aljaaf, Dhiya Al-Jumeily, Hussein M. Haglan, Mohamed Alloghani, Thar Baker, Abir J. Hussain, and Jamila Mustafina., Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics, IEEE Congress on Evolutionary Computation (CEC), (2018).

[2] SetuBasak, Md. MahbubAlam, AniruddhaRakshit, Ahmed Al Marouf, AnupMajumder, Predicting and Staging Chronic Kidney Disease of Diabetes (Type-2) Patient using Machine Learning Algorithms, International Journal of Innovative Technology and Exploring Engineering (IJITEE), (2019).

[3] Sahil Sharma, Vinod Sharma, Atul Sharma, A Two-Stage Hybrid Ensemble Classifier Based Diagnostic Tool for Chronic Kidney Disease Diagnosis Using Optimally Selected Reduced Feature Set, International Journal of Intelligent Systems and Applications in Engineering (IJISAE), (2018).

[4] Suman Bala, Krishan Kumar, A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique, International Journal of Computer Science and Mobile Computing (IJCSMC), (2014).

[5] S.Ramya, Dr.N.Radha, Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms, International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), (2016).

[6] L.JerlinRubini, Dr.P.Eswaran, Generating comparative analysis of early-stage prediction of Chronic Kidney Disease, International Journal Of Modern Engineering Research (IJMER), (2015).

[7] SiddheshwarTekale, PranjalShingav, SukanyaWandhekar, AnkitChatorikar, Prediction of Chronic Kidney Disease Using Machine Learning Algorithm, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), (2018).

[8] https://www.geeksforgeeks.org/software-engineering-prototyping-model/

[9] https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9