e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:09/September-2020 Impact Factor- 5.354 www.irjmets.com

PREDICTION AND PREVENTIVE AWARENESS OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING ALGORITHMS

Akshay Shirahatti[*1], Vijay Yadav[*2], Nagarjuna Vadde[*3], Aman Singh[*4], Prof Pranita Mahajan[*5], Prof Rizwana Sheikh[*6]

[*1,2,3,4,5,6]Department Of Computer Engineering,SIES Graduate School Of Technology, Nerul, Maharashtra, India.

ABSTRACT

Chronic Kidney Disease(CKD) also called Chronic Kidney Failure(CKF) is a medical condition that describes the gradual loss of kidney function over a period of months to years. Early stage prediction and prevention of this disease based on severity is one of the most important problems of medical fields. The primary goal of this study is to identify the best machine learning classification algorithm for prediction of CKD and later provide a suitable prevention technique i.e diet recommendation based on severity of disease. The proposed system extracts the features which are responsible for CKD, than machine learning can automate the prediction of disease. After the prediction of disease risk assessment and classification into different stages is done according to its severity. Diet recommendation for patients will be given according to classification of disease into different stages. The dataset is based on clinical history, Electronic Medical Records(EMR), and laboratory tests. Experimental results showed over 91% success rate in classifying the patients with kidney diseases based on performance of different Machine Learning Algorithms.

Keywords: Machine Learning; Chronic Kidney Disease; Classification; Biomedical Engineering

I. INTRODUCTION

Kidney disease or renal failure is a medical condition in which the kidneys are unable to operate properly and a severe decrease in kidney function happens. The CKD is also called a chronic kidney failure where according to current medical statistics the 10% of the population worldwide is affected by CKD . There were around 58 million deaths in the year of 2005 worldwide. Where according to the World Health Organization (WHO) 35 million are attributed to chronic diseases. Currently it is estimated that one in five men, and one in four women aged 60 through 75 are going to be affected by CKD worldwide.

Undiagnosed CKD can be identified, predicting the likelihood that patients will develop chronic disease, and present patient-specific prevention interventions with Machine learning techniques[1].Data mining approaches has been recently used for attaining diagnostics effects in disease. These concepts analyze data from various sights and derive helpful information. The harmful outcomes can be avoided and prevented by early detections, according to researchers conducted. Awareness of CKD among patients is gradually increasing, but still low[1].

Management of diet depends on the current Glomerular Filtration Rate (GFR rate) and the severity of the disease. We will be classifying the disease in five stages- Stage 1, stage 2 and stage 3, Stage 4, Stage 5. Stage 1 is safe and requires a lenient diet plan to be followed. Whereas stage 2, a potential CKD patient will be given a restricted and strict diet. Keeping the balance of minerals, electrolytes, and liquids inside the body will be difficult for stage 3 to 5 patients . Therefore, they have to be under proper dietary guidance. An important diet for renal improvement and preventing further harm is essential, which also helps in keeping balance of electrolytes and water in the body[1].

Other than stages of severity, many other factors will contribute in shaping the diet. The blood potassium level, urea level, calcium level, phosphorus level and so on. In this study, to identify a suitable diet plan for a CKD patient the main focus will be on blood potassium level.

II. LITERATURE SURVEY

Our project will review the literature related to the area of study –Early Stage Detection and Prevention of Chronic Kidney Disease. This provides background to the research, it will provide the necessary backbone and support to the research. By reviewing the past publications and researches related to the study, the researcher will have an idea of how such a study has been done in the past. In this way, this research may be able to reflect, compare itself, learn from setbacks and produce a stronger and more efficient study.

During the literature review, we have identified that there are less studies performed in CKD . But there are several studies where SVM has been used abundantly. For instance in these studies researchers used SVMs as a classification model for detecting and diagnosing malignant and benign tumors based on MRI features, ultrasound features and mammographic features. This result let us try to create and test an SVM classifier over a CDK dataset.

A. Kusiak et al [6],In his study used various data preprocessing, data transformations, and data mining approaches to understand the insides of the interaction between various clinical parameters and patient survival, which are on kidney dialysis. Two different data mining algorithms were applied for the knowledge extraction in the form of decision rules. The extracted rules were used to predict survival of new unseen patients. Data mining algorithms identified the important medical parameters for carrying out the prediction process. The introduced new research concepts have been implemented and tested using data that is collected at four dialysis sites. Presented approach reduces the effort and cost of selecting patients for clinical studies. Patients can be chosen based on the prediction results and the discovered vital parameters

T. Di Noia et al [5], the researchers developed a software tool that demonstrates the capabilities of ANN for classification of patients' health status which potentially leads to End Stage of Kidney Disease (ESKD). The classifier is based on an ensemble of ten ANN networks. It has been trained by using data collected in a period of 38 years at University of Bari. The tool has been improved and made derivable both as a mobile application and as a web application. The tool is important for clinical needs based on the largest cohort worldwide.

R. Baccoli et al [4], the researchers tried to predict the Long Term Kidney Transplantation Outcome. They have performed comparative analysis between ANN and LR algorithms. The comparative analysis has been implemented based on performance metrics like accuracy, sensitivity and specificity. During the study for the kidney transplant recipients prediction of kidney rejection which was based on ten training and validating datasets. The experimental results showed that ANN can be considered a useful supportive algorithm in the prediction process of the defined problem. In summary, the ability of predicting kidney rejection (sensitivity) was 38% for LR versus 62% for ANN. The ability of predicting no-rejection (specificity) was 68% for LR compared to 85% of ANN.

M.P.N.M. Wickramasinghe, D.M. Perera, and K.A.D.C.P. Kahandawaarachchi et al [2] presented a research study, by fetching data from patient's medical records and then applying classification algorithms on these records, which would in turn give a suitable diet plan to the patients of CKD. The proposed study is to identify the different diet plans by predicting potassium zone of CKD patients according to the blood potassium level.

Akash Maurya et al [1] has proposed a system that extracts the features which are responsible for CKD, then the machine learning processes were able to automate the classification of the chronic kidney disease in different stages according to its severity. The objective of the study is to use machine learning algorithms and suggest suitable diet plans for CKD patients using classification algorithms on medical test records. Diet recommendations for patients were given according to the potassium zone which was calculated using blood potassium level to slow down the progression of CKD.

III. METHODOLOGY

a. *Data Collection:*

The CKD has been obtained from UCI machine learning repository [7]. In total it contains 400 cases, out of which 250 of the cases are patients with CDK and the rest 150 are not. The target variable indicates

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:09/September-2020 Impact Factor- 5.354 www.irjmets.com

whether a patient has a CDK or not. There are 25 attributes where 24 are clinical features and remaining is a target attribute. The features are divided into three parts clinical history, physical examination and lab tests. According to the properties of the attributes, the target attribute was classified into negative (expressed by "no disease") and positive (expressed by "presence of disease").

b. Data Preprocessing And EDA*:*

Each categorical (nominal) variable was coded to facilitate the processing in a computer. For the values of rbc and pc, normal and abnormal were coded as 1 and 0, respectively. For the values of pcc and ba, present and notpresent were coded as 1 and 0, respectively. For the values of htn, dm, cad, pe and ane, yes and no were coded as 1 and 0, respectively. For the value of appet, good and poor were coded as 1 and 0, respectively. Although the original data description defines three variables sg, al and su as categorical types, the values of these three variables are still numeric based, thus these variables were treated as numeric variables. All the categorical variables were transformed into factors.

```
#This plot shows the patient's sugar level compared to their ages
df.plot(kind='scatter', x='age',y='su');
plt.show()
```

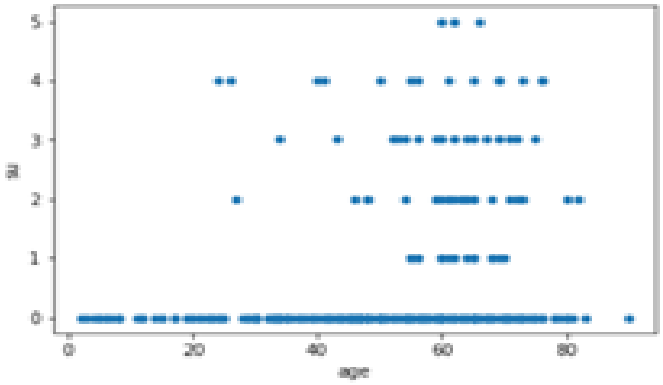


Figure 1: Scatter plot of patients sugar level compared to their ages.

From the above figure we can see that the patients under the age group of 40-80 have high sugar because of which they are more prone to CKD.

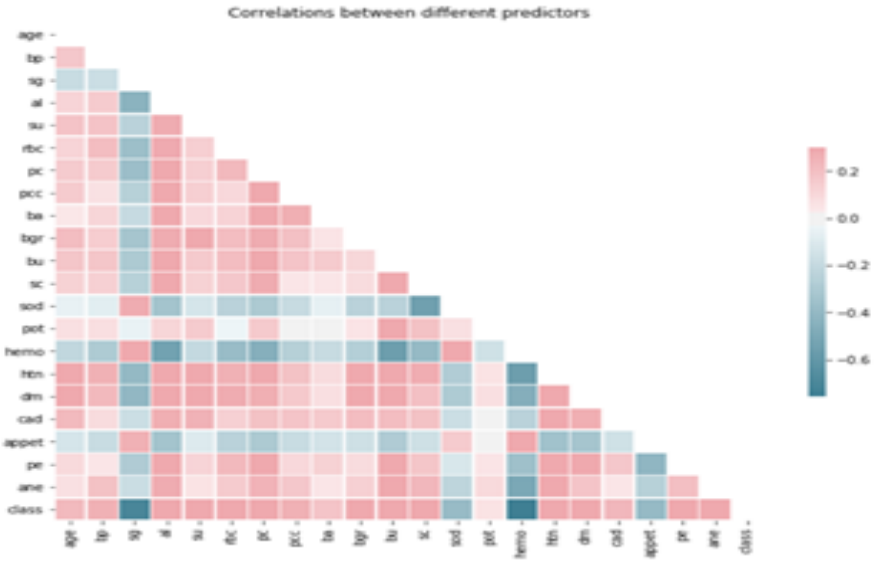


Figure 2: Correlation between different predictors

From the above figure we can see that Serum creatinine, Albumin, sugar ,hypertension are highly correlated with each other and therefore these are the most important parameters which will help in prediction of Chronic Kidney Disease.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:09/September-2020 Impact Factor- 5.354 www.irjmets.com

IV. PROPOSED MODEL

In this section, the classifiers were initially established by various machine learning algorithms to diagnose the data samples. Among these models, those with better performance were selected as potential components.

1. Support Vector Machine (SVM)

The SVM is a supervised learning algorithm that is used for data classification and regression [8]. It searches for the best hyperplane which separates between classes. The best hyperplane is considered the one which leaves the maximum margin between the two distinct classes. The margin is defined as the width of the hyperplane from the closest point of the two distinct classes. Bounds between data sets and hyperplanes are called support vectors [8,9].

The models of SVM were generated by using the RBF kernel function, and the function is described as follow:

$$K\langle x_1, x_2\rangle = e^{-\gamma\|x_1 - x_2\|^2}$$

where γ was set to [0.1, 0.5, 1, 2, 3, 4]. Parameter C represents the weight of misjudgment loss, and it was set to [0.5, 1, 2, 3]. In each calculation of the model training, the algorithm selects the best combination of parameters to establish the model by grid search. 97.5% accuracy was achieved using the Support Vector Machine.

2. Random Forest (RF):

RF was established using all variables. Two strategies were used to determine the number of decision trees generated. One is to use the default 500 trees and the other is to use the number of trees corresponding to the minimum error in the training stage. The RF was established using both strategies and evaluated on the data sets obtained by KNN imputation. The same random number seed 1234 was used to divide data and establish a model, and the accuracy obtained is 100 %. It can be seen that the default number of trees is a better choice, therefore we selected the default 500 trees to establish RF.

3. K-nearest neighbours (KNN):

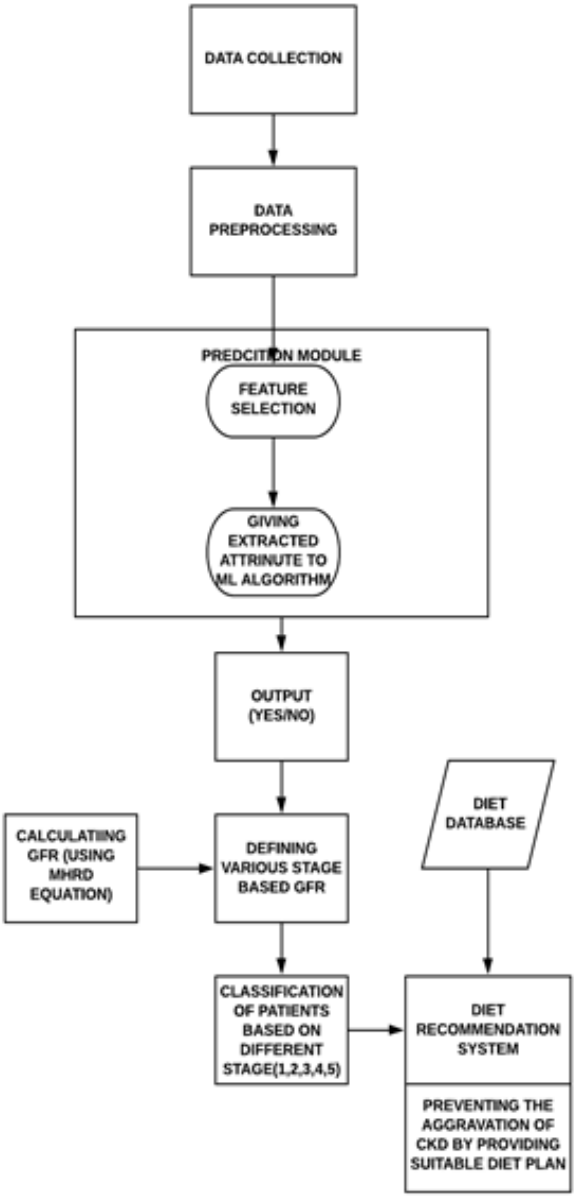
For the KNN, due to the nearest Euclidean distance with the detected sample, when the number of samples that are selected in training data set is an even number, the algorithm randomly selects a category as the output result of the detected sample in the situation wherein the number of selected samples belonging to ckd and not ckd are the same. In each calculation of model training, the algorithm selected the best parameter to establish the model by grid search. The accuracy obtained using KNN is 71.25% which is least among all other algorithms used.

4. Regression Based Model (LOG):

LOG is based on logistic regression, and it obtains the weight of each predictor and a bias. If the sum of the effects of all predictors exceeds a threshold, the category of the sample will be classified as ckd or non-ckd. The output of LOG was the probability that the sample belongs to non-ckd, and the threshold was set to 0.5.The accuracy achieved is 98.5% using Logistic Regression.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:09/September-2020 Impact Factor- 5.354 www.irjmets.com

FLOWCHART:



PREVENTION:

Adding new attributes

The Glomerular Filtrate Rate(GFR) is used to check the goodness of kidney functioning. GFR values tell how well kidneys are functioning to remove waste from the body. Hence we are adding the GFR attribute (using MDRD equation). MDRD equation for GFR calculation are as follows:

GFR = 175 × (SC)^-1.154 × (Age)^-0.203 × (0.742 if female) × (1.212 if African American)

Where,

GFR (glomerular filtration rate) = mL/min/1.73 m2

SC (serum creatinine) = mg/dL

Age = years

Classification based on different stages

CKD can be classified into 5 different stages using the MDRD equation. Detecting CKD at early stages can help in preventing the progression of kidney diseases and slow its progression. We have five stages on basis of GFR values they are as follow:

- **Stage 1** with normal or high GFR (GFR > 90 mL/min)
- **Stage 2** Mild CKD (GFR = 60-89 mL/min)
- **Stage 3A** Moderate CKD (GFR = 45-59 mL/min)
- **Stage 3B** Moderate CKD (GFR = 30-44 mL/min)
- **Stage 4** Severe CKD (GFR = 15-29 mL/min)
- **Stage 5** End Stage CKD (GFR <15 mL/min)

Diet recommendation module

As dietary management plays an important role when you have CKD - especially in the advanced stages, what you eat is an important part of your care plan, because your diet (along with exercising and taking proper medications), might help slow the damage happening to your kidneys. Mostly patients suffering from diabetes and high blood pressure conditions should have a very strict diet to prevent kidney failure. So in this module, based on the stages detected (using GFR value) and the output obtained from the classification based on different stages of patients will be given a suitable diet. Also CKD patients in higher stages(Stage3,4,5) will be given an alert for looking into treatment options.

V. ANALYSIS

It is seen that accuracy of KNN is less as compared to others which might be due to the fact that there are too many redundant variables. Random Forest achieved maximum accuracy which is best amongst all other algorithms discussed above.

Figure below shows analysis :

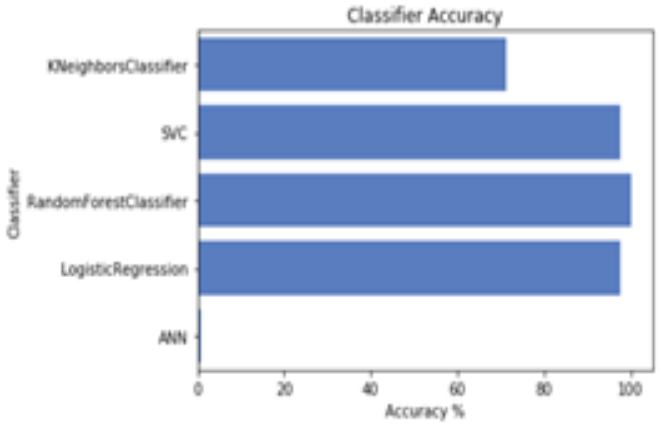


Figure 3: Classifier Accuracy

VI. RESULTS AND DISCUSSION

A comparative study between various Machine learning algorithms, with and without preprocessing has been proposed. The various classifiers were tested by performing train-test split on the dataset .The dataset was divided in a ratio of 70-30.Out of the 400 observations, 280 were used for training the classifiers and the remaining 120 were used for testing.

For each algorithm, we computed the results for the following:

1.Algorithm without any preprocessing

2.Algorithm with normalization of the data

3.Algorithm with standardization of the data

4.Algorithm with normalization and Standardization of data

Table-1: Comparison of different machine learning model

SN.	Model Name	Accuracy	F-1 Score
1	SVM	97.5%	0.98
2	Random Forest	99.25%	0.99
3	KNN	71.25%	0.75
4	Regression based Model	97.2%	0.98

VII. CONCLUSION

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. It is found out that Random Forest Algorithm performs better as compared to other algorithms performed with the accuracy of 100 %.Hence, we speculate that applying this methodology to the practical diagnosis of CKD would achieve a desirable effect. In addition, this methodology might be applicable to the clinical data of the other diseases in actual medical diagnosis. However, in the process of establishing the model, due to the limitations of the conditions, the available data samples are relatively small, including only 400 samples. Therefore, the generalization performance of the model might be limited. In addition, as there are only two categories (ckd and not ckd) of data samples in the data set, the model can not diagnose the severity of CKD. In the future, a large number of more complex and representative data will be collected to train the model to improve the generalization performance while enabling it to detect the severity of the disease. We believe that this model will be more and more perfect by the increase of size and quality of the data..

ACKNOWLEDGEMENTS

We would like to thank the UCI machine learning repository and the donors for sharing this CKD data set.

VIII. REFERENCES

[1] Akash Maurya,"Chronic kidney disease prediction and Recommendation of Suitable Diet plan by using Machine Learning" in 2019 International Conference on Nascent Technologies in Engineering (ICNTE 2019)

[2] M.P.N.M. Wickramasinghe, D.M. Perera, and K.A.D.C.P. Kahandawaarachchi,"Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," in 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, Australia, 2017.

[3] Navdeep Tangri, MD, PhD, FRCPC"Multinational Assessment of Accuracy of Equations for Predicting Risk of Kidney Failure A Meta-analysis "JAMA January 12, 2016 Volume 315, Number 2.

[4] G. Caocci, R. Baccoli, R. Littera, S. Orrù, C. Carcassi and G. La Nasa, "Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long Term Kidney Transplantation Outcome", Artificial Neural Networks Kenji Suzuki, IntechOpen, DOI: 10.5772/53104, 2013.

[5] T. Di Noia, V. C. Ostuni, F. Pesce, G. Binetti, D. Naso, F. P. Schena, and E. Di Sciascio. "An end stage kidney disease predictor based on an artificial neural networks ensemble", Expert Systems with Applications, vol. 40, pp. 4438–4445, 2013

[6] A. Kusiak, B. Dixonb, and Sh. Shaha, "Predicting survival time for kidney dialysis patients: a data mining approach", Computers in Biology and Medicine, vol. 35, pp. 311–327, 2005

[7] Soundarapandian P. (2015). UCI Machine Learning Repository

[https://archive.ics.uci.edu/ml/datasets/chronic_ kidney_disease]. Irvine, CA: University of California, School of Information and Computer Science.

[8] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, pp. 27, 2011.

[9] N. H. Sweilam, A. Tharwat, and N. A. Moniem, "Support vector machine for diagnosis cancer disease: A comparative study," Egyptian Informatics Journal, vol. 11, pp. 81-92, 2010.

[10] E. Gumus, N. Kilic, A. Sertbas, and O. N. Ucan, "Evaluation of face recognition techniques using PCA, wavelets and SVM," Expert Systems with Applications, vol. 37, pp. 6404-6408, 2010.