INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Chronic Kidney Disease Prediction UsingMachine Learning Techniques

Dasari Tejaswini
B-Tech Student

Munpally Rama Krishna
B-Tech Student

Vangaveeti Sanjay
B-Tech Student

G. DIVYA
(Assistant Professor)

Department of Information Technology
CMR Technical Campus
Kadlakoya (V), Medchal, Hyderabad-501401

Abstract: Chronic kidney disease develops slowly, with few symptoms. It is often not recognized until the disease is advanced. If it is detected early, treatment can slow down or avoid kidney function decline and diminish the negative effects on other body functions. A blood test measuring glomerular filtration rate assesses how well the kidneys clear the blood of a waste product called creatinine. A value of 60 to 90 may be an early sign of kidney disease; a value below 60 is usually considered abnormal. A test using a urine sample evaluates the presence of protein (albumin) in the urine; repeated results of 30 mg or more can indicate a problem. High blood pressure may also point to underlying chronic kidney disease. Human body organs are interconnected with each other, so if one organ does not work properly then there will be symptoms due to this improperness. When the kidney is not working properly this would cause some changes in attributes such as serum creatinine, blood pressure, blood sugar and hemoglobin. Therefore, this correlation among the attributes can be used to identify CKD. Doctors inherently use these attributes and their inter-relationships from reports such as blood reports and urine reports to identify the diseases. Chronic kidney disease is a slow and progressive loss of kidney function over a period of several years. Medical tests for other purposes sometimes contain useful information about Chronic Kidney Disease. Attributes of different medical tests are investigated to identify what attributes contain useful information about Chronic Kidney Disease. A database with several attributes of Chronic Kidney Disease is analyzed with different techniques. Common Spatial Pattern (CSP) filter and Linear Discriminant Analysis (LDA) are first used to identify the dominant attributes that could contribute in detecting Chronic Kidney Disease. These analyses suggest that Specific Gravity, Hyper Tension, Hemoglobin, Diabetes Mellitus, Albumin, Appetite, Red Blood Cell Count and Pus Cell are the most important attributes in the early detection of Chronic Kidney Disease. The main objective of this project is to determine the kidney function failure by applying the Random Forest algorithm and to classify the chronic and non-chronic kidney diseases.

I.INTRODUCTION

Chronic kidney disease (CKD) is non-communicable disease that has significantly contributed to morbidity, mortality and admission rate of patients worldwide. It is quickly expanding and becoming one of the major causes of death all over the world. A report from 1990 to 2013 indicated that the global yearly life loss caused by CKD increased by 90% and it is the 13th leading cause of death in the world. 850 million people throughout the world are likely to have kidney diseases from different factors. According to the report of world kidney day of 2019, at least 2.4 million people die every year due to kidney related disease. Currently, it is the 6th fastest-growing cause of death worldwide CKD is becoming a challenging public health problem with increasing prevalence worldwide. Its burden is even higher in low-income countries where detection, prevention and treatment remain low. Kidney disease is serious public health problem in Ethiopia affecting hundreds of thousands of people irrespective of age, sex. The lack of safe water, appropriate diet, and physical activities is believed have contributed. Additionally, communities living in rural area have limited knowledge about the CKD. According to WHO report of 2017, the number of deaths in Ethiopia due to kidney disease was 4,875. It is 0.77% of total deaths that has ranked the country 138th in the world. The age-adjusted death rate is 8.46 per 100,000 of the population and the death rate increased to 12.70 per 100,000 that has ranked the country 109 in 2018. National kidney foundation classifies stages of CKD into five based on the abnormal kidney function and reduced Glomerular Filtration Rate (GFR), which measures a level of kidney function,. The mildest stage (stage 1 and stage 2) is known with only a few symptoms and stage 5 is considered as end-stage or kidney failure. The Renal Replacement Therapy (RRT) cost for total kidney failure is very expensive. The treatment is not also available in most developing countries like Ethiopia. As a result, the management of kidney failure and its complications is very difficult in developing countries due to shortage of facilities, physicians, and the high cost to get the treatment. Hence, early detection of CKD is very essential to minimize the economic burden and maximize the effectiveness of treatments. Predictive analysis using machine learning techniques can be helpful through an early detection of CKD for efficient and timely

II.LITERATURE REVIEW

This section consists of the reviews of various technical and review articles on data mining techniques applied to predict Kidney Disease. DSVGK Kaladhar, Krishna Apparao Rayavarapu and Varahalarao Vadlapudi et al [6]. described in their research to understand machine learning techniques to predict kidney stones. They predicted good accuracy with C4.5, Classification tree and Random forest (93%) followed by Support Vector Machines (SVM) (91.98%). Logistic and NN has also shown good accuracy results with zero relative absolute error and 100% correctly classified results. ROC and Calibration curves using Naive Bayes has also been constructed for predicting accuracy of the data. Machine learning approaches provide better results in the treatment of kidney stones. J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van den Berghe, K.U.Leuven et al [7]. Explored data mining techniques for predicting acute kidney injury after elective cardiac surgery with Gaussian process & machine learning techniques (classification task & regression task). K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna et al [8]. presented performance comparison of Artificial Neural Networks, Decision Tree and Logical Regression are used for Kidney dialysis survivability. The data mining techniques were evaluated based on the accuracy measures such as classification accuracy, sensitivity and specificity. They achieved results using 10 fold cross-validations and confusion matrix for each technique. They found ANN shows better results. Hence ANN shows the concrete results with Kidney dialysis of patient records.

III. IMPLEMENTATION

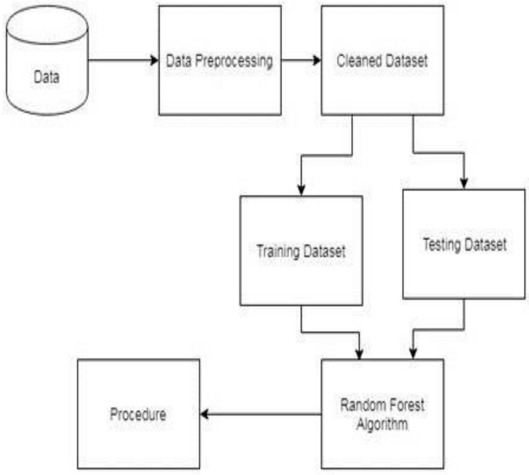


Fig-3: System Architecture

Data pre-processing is a way to convert the noisy and huge data into relevant and clean data, asthe data available is Real world data, so it contains inaccurate data, missing values and other Noisy data, for removing this inconsistent data from the Dataset, the proposed system have to clean the raw data. This is an important part to complete the prediction model. It reduces the dimensionality and helps the machine to achieve better results. This is one of the most time consuming stage in building a classification model.

Following data pre-processing steps are followed:

Looking up for proper format

As we have made our model using python, so we need a csv file (comma separated value) for ourcode. The data downloaded is in the form of RAR file, so we extract the data from the text file available and save it into a csv file so that our python code can read it. This is the first most important step, if the data is not available in requires format then we cannot design the classification model.

Finding Missing Values

When the data collected is real world data, and then it will contain missing values. This brings more change in the prediction accuracy. Sometimes these missing values can be simply deleted or ignored if they are not large in number. It is the simplest way to handle the missing data but it is not considered healthy for the model as the missing value can be an important attribute contributing to the disease. The missing values can also be replaced by zero this will not bring any change as whole, but this method cannot be much yielding. So an efficient way to handle missing values is to use mean, average of the observed attribute or value. This way we lead to more genuine data and better prediction results.

Data transformation:

In this step we transform the given real data into required format. The data downloaded consist of Nominal, Real and Decimal values. In this step we convert the Nominal data into numerical data of the form 0 and 1. The positive value is assigned the value of 1 and the negative value is assigned the value of 0. Now the resultant csv file comprises of all the integer and decimal values for different CKD related attributes. Feature Selection In this step we select subset of relevant attributes from the total give attributes. This stage helps in reducing the dimensionality and making the model simpler and easy to use, thus leading to short training time and high accuracy. To obtain highly dependent features for CKD prediction we have used Correlation and dependence method. The term correlation can be defined as mutual relationship between two. In this those attributes are chosen which highly influence the occurrence of Chronic Kidney Disease. By using the correlation it is found that 10 attributed were highly correlated to the occurrence of CKD from the total of 25 attributes.

The 10 attributes selected from a total of 25 attributes are: Specific Gravity, Hyper Tension, Hemoglobin, Diabetes Mellitus, Albumin, Appetite, Red Blood Cell Count and Pus Cell**.**

Data Set

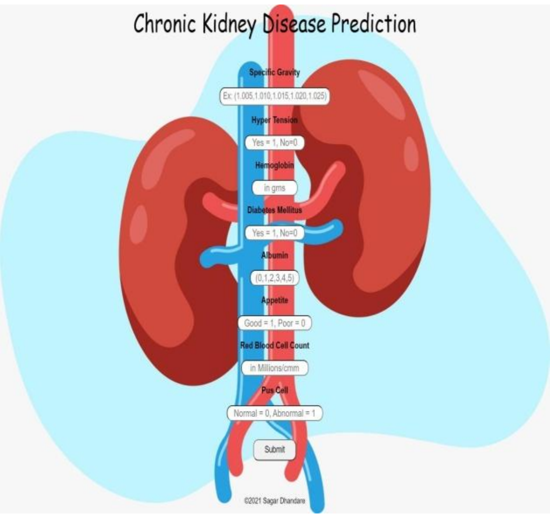
The dataset is divided into two sub datasets both containing 25 attributes.

Training data: Training dataset is derived from main dataset and it contains 300 out of 400records in main dataset of CKD.

Algorithm: random forest classification

The random forest is an ensemble approach that can also be thought of as a form of nearest neighbor predictor. Ensembles are a divide-and-conquer approach used to improve performance. The main principle behind ensemble methods is that a group of 'weak learners' cancome together to form a 'strong learner'. The random forest starts with a standard machine learning technique called a 'decision tree' which, in ensemble terms, corresponds to our weak learner. The decision tree algorithm repeatedly splits the data set according to a criterion that maximizes the separation of the data, resulting in a tree-like structure. In this algorithm an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. The random forest takes this notion to the next level by combining trees with the notion of an ensemble. Thus, in ensemble terms, the trees are weak learners and the random forest is a strong learner. The advantages of a random forest classifier are that its' runtimes are quite fast, and that it is able to deal with unbalanced and missing data. Weaknesses of this algorithm are that when used for regression it cannot predict beyond the range in the training data, and it may over-fit data sets that are particularly noisy.

IV. RESULT



Chronic Kidney Disease Prediction

Specific Gravity

Ex: (1.005,1.010,1.015,1.020,1.025)

Hyper Tension

Yes = 1, No=0

Hemoglobin

in gms

Diabetes Mellitus

Yes = 1, No=0

Albumin

(0,1,2,3,4,5)

Appetite

Good = 1, Poor = 0

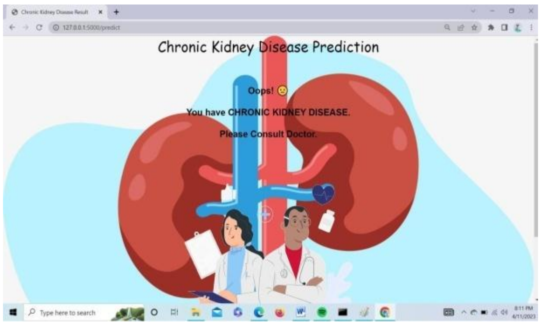
Red Blood Cell Count

in Millions/cmm

Pus Cell

Normal = 0, Abnormal = 1

Fig-8.2: input screen



Chronic Kidney Disease Prediction

Oops! 😟

You have CHRONIC KIDNEY DISEASE.

Please Consult Doctor.

Fig-: Output screen

V. CONCLUSION

This work examines the ability to detect CKD using machine learning algorithms while considering the least number of tests or features. We approach this aim by applying four machine learning classifiers: logistic regression, SVM, random forest, and gradient boosting on a small dataset of 400 records. In order to reduce the number of features and remove redundancy, the association between variables has been studied. A filter feature selection method has been applied to the remaining attributes and found that there are hemoglobin, albumin, and specific gravity have the most impact to predict the CKD. The classifiers have been trained, tested, and validated using 10-fold cross-validation. Higher performance was achieved with the gradient boosting algorithm by F1-measure (99.1 %), sensitivity (98.8%), and specificity (99.3%). This result is the highest among previous studies with less number of features and hence less cost. Therefore, we conclude that CKD can be detected with only three features. Also, we found that hemoglobin has the highest contribution in detecting CKD, whereas albumin has the lowest using RF and GB models. Since the data used in this research is small, in the future, we aim to validate our results by using big dataset or compare the results using another dataset that contains the same features. Also, in order to help in reducing the prevalence of CKD, we plan to predict if a person with CKD risk factors such as diabetes, hypertension, and family history of kidney failure will have Chronic Kidney Disease in the future or not by using appropriate dataset.

REFERENCES

1. J. Radhakrishnan et al, "Taming the chronic kidney disease epidemic: a

global view of surveillance efforts," Kidney Int., vol. 86, (2), pp. 246- 250, 2014.

2. R. Lozano et al, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," The Lancet,vol. 380, (9859), pp. 2095- 2128, 2012.
3. R. Ruiz-Arenas et al, "A Summary of Worldwide National Activities in Chronic KidneyDisease (CKD) Testing," Ejifcc, vol. 28, (4), pp. 302, 2017.
4. Q. Zhang and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: systematic review," BMC Public Health, vol. 8, (1), pp. 117, 2008.
5. T. Di Noia et al, "An end stage kidney disease predictor based on an artificial neural networks ensemble," Expert Syst. Appl., vol. 40, (11), pp. 4438-4445, 2013.
6. H. S. Chase et al, "Presence of early CKD-related metabolic complications predict progression of stage 3 CKD: a case-controlled study," BMC Nephrology, vol. 15, (1), pp. 187,2014.
7. K. A. Padmanaban and G. Parthiban, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease," Indian Journal of Science and Technology, vol. 9, (29), 2016.
8. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in Healthcare Informatics (ICHI), 2016 IEEE International Conference On,2016.
9. W. Gunarathne, K. Perera and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference On, 2017.
10. H. Polat, H. D. Mehr and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," J. Med. Syst., vol. 41, (4), pp. 55, 2017.
11. P. Yildirim, "Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction," in Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual, 2017.
12. J. Aljaaf et al, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in 2018 IEEE Congress on Evolutionary Computation (CEC),2018.