WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Prediction of Chronic Kidney Disease Using Machine Learning Perspective

Rushi Eshwar Reddy Neelam[1], D. Rushika[2], K. Rushindra[3], P. Rushitha[4], Ch. Sadwik[5], B. Sahithi[6], Prof. Sameera Sulthana[7]

[1,2,3,4,5,6]B. Tech, Artificial Intelligence and Machine Learning, Malla Reddy University, Hyderabad
[7]Professor, Department of AIML, Malla Reddy University, Hyderabad
[1]2111cs020412@mallareddyuniversity.ac.in, [2]2111cs020413@mallareddyuniversity.ac.in, [3]2111cs020414@mallareddyuniversity.ac.in,
[4]2111cs020415@mallareddyuniversity.ac.in, [5]2111cs020416@mallareddyuniversity.ac.in, [6]2111cs020417@mallareddyuniversity.ac.in

ABSTRACT –

Chronic Kidney Disease (CKD) poses a significant global health burden, demanding early detection and intervention to improve patient outcomes. Leveraging the power of machine learning, this study explores predictive models for identifying CKD risk factors and forecasting disease progression.

Utilizing a comprehensive dataset comprising demographic information and medical history from a cohort of CKD patients, various machine learning algorithms, including Random Forest Regression and Support Vector Machines were employed. Feature selection techniques were applied to discern critical predictors contributing to CKD prognosis.

This research underscores the potential of machine learning in facilitating early CKD detection and risk stratification. The developed predictive models not only offer high accuracy but also provide insights into the influential factors driving CKD development. Also , the proposed machine learning framework provides a reliable and scalable approach for early CKD detection, facilitating timely clinical interventions and personalized patient care.

Keywords—Chronic kidney disease, prediction, machine learning,Random Forest Regression

I. INTRODUCTION

The healthcare industry is producing copious amounts of data which need to be mined in order to discover hidden information for effective prediction, diagnosis and decision making. Currently, kidney disease has been a crucial problem. It is one of the leading causes of death in India.If this disease gets worse, wastes can accumulate in the blood and can cause difficulties like high blood pressure, anemia, weakening of bones, poor nutritional health and nerve damage. Also, kidney disease increases the risk of having heart and blood vessel disease.

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. The Global Burden of Disease (GBD) 2015 ranks chronic kidney disease as the eighth leading cause of death in India. All over the world, the highest count of patient with diabetes is in India with the projection figure of 57.2 million cases in 2025 and also the count of patient with hypertension is expected to double from 2000 to 2025,hence these will make India the reservoir of CKD .The burden of CKD management thus falls largely on primary care providers (PCPs). Hence an accurate, convenient, and automated CKD detection method is important for clinical practice.

Undiagnosed CKD can be identified, predicting the likelihood that patients will develop chronic disease, and present patient-specific prevention interventions with Machine learning techniques. Accurate predictive models can be created by health systems, which lower risks and eventually improve standards. The data mining techniques of classification, clustering and association helps in extracting knowledge from large amount of data. Machine learning and data mining techniques together have been the prime factors in determining and diagnosis of various critical diseases.

II. LITERATURE REVIEW

[J. Snegha][2] proposed a system that uses various data mining techniques like Random Forest algorithm and Back propagation neural Network. Here they compare both of the algorithm and found that Back Propagation algorithm gives the best result as it uses the supervised learning network called feedforward neural network.

[M.P.N.M. Wickramasinghe][3] 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.

[M. Dr. S.Vijayarani][4] discussed about a comparison made between two classification algorithms namely Support Vector Machines and Artificial Neural Networks. Based on their respective accuracies and timings, the goal to predict CKD was achieved. The one with higher accuracy and good timing was chosen.

[Guneet Kaur][5] proposed a system for predicting the CKD using Data Mining Algorithms in Hadoop. They use two data mining classifiers like KNN and SVM. Here the predictive analysis is performed based upon the manually selected data columns. SVM classifier gives the best accuracy than KNN in this system.

[Neha Sharma][6] proposed a system in which the kidney disease of a patient is analyzed and the results are to compute automatically using the data set of the patient. Here Rule based prediction method is used. This system uses neuro-fuzzy method and obtained the outcome by mathematical computation.

[Anusorn Charleonnan][7] projected that revolves around four classification algorithms which make predictive models for chronic kidney disease. The goal here was to find the best classifier amongst the four: logistic regression, Support Vector Machines, Decision trees classifier and K-nearest neighbors. Chronic kidney disease dataset was used to construct the predictive model and later comparison between their performances was done to find the best classifier amongst these to predict chronic kidney disease.

III. PROBLEM STATEMENT

Building a machine learning model for early detection and prediction of chronic kidney disease (CKD) utilizing patient demographic data, clinical attributes (e.g., blood pressure, serum creatinine levels), and medical history. The goal is to develop a reliable predictive tool that aids healthcare providers in assessing CKD risk, enabling timely interventions and personalized treatment strategies to mitigate the progression of the disease, thereby improving patient prognosis and reducing healthcare costs associated with advanced CKD complications.

IV. METHODOLOGY

The methodology employed in this Machine Learning project involves the implementation of a Random Forest Regression (RFR) for prediction of chronic kidney disease using dataset from Kaggle.

Dataset: The dataset here we use is the publically available CKD dataset. Out of 25 attributes,11 are numeric and 14 are nominal attributes. The data set contains number of missing values. Here the information of dataset uses the patients data like age, blood pressure, specific gravity, albumin, sugar, red blood cells etc. The data set contains number of missing values.

Table.1 List of attributes present in the CKD dataset

Attributes	Type
Age	Numeric
Blood Pressure	Numeric
Specific Gravity	Numeric
Albumin	Numeric
Sugar	Numeric
Red Blood Cells	Nominal
Pus Cell	Nominal
Pus Cell clumps	Nominal
Bacteria	Nominal
Blood Glucose Random	Numeric
Blood Urea	Numeric
Serum Creatinine	Numeric
Sodium	Numeric
Potassium	Numeric
Hemoglobin	Numeric
Packed Cell Volume	Numeric
Red Blood Cell count	Numeric
White Blood Cell Count	Numeric
Hypertension	Nominal
Diabetes Mellitus	Nominal
Coronary Artery Disease	Nominal
Appetite	Nominal
Pedal Edema	Nominal
Anemia	Nominal
Class	Class

CKD is caused due to diabetes and high blood pressure. Due to Diabetes our many organs get affected and it will be followed by high blood sugar. So it is important to predict the disease as early as possible. This study improvises some of the machine learning techniques to predict the disease.

Steps:

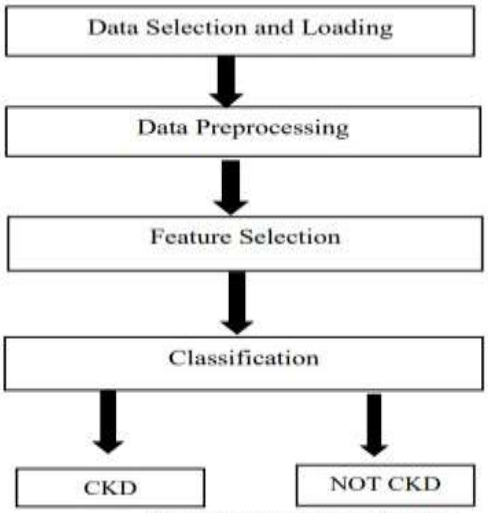


Figure.1 Flowchart of the proposed system

Data Pre Processing Techniques:

- **Cleaning Noisy Data:** Removing outlier and smoothening noisy data is an important part of preprocessing. Outliers are values that lies away from the range of the rest of the values. In clinical data, outliers may arise from the natural variance of data. The potential outliers are the data points that fall above Q3 + 1.5(IQR) and below Q1 − 1.5 (IQR), where Q1 is the first quartile, Q3 is the third quartile, and IQR = Q3 − Q1 .
- **Handling Missing Values:** Data is not always available (or missed) due to equipment malfunction, inconsistent with other recorded data and thus deleted, not entered into the database due to misunderstanding, some data may not be considered important at the time of entry.
- **Handling Categorical Data**: In this step, data has been transformed into the required format. The nominal data converted into numerical data of the form 0 and 1. For instance, 'Gender' has the nominal value that can be labeled as 0-for female and 1-for male. After preprocessing the data then the resultant CSV file comprises all the integer and float values for different CKD related features.

Model Selection:

Random Forest Regression:

- Random forest algorithm constructs multiple decision trees to act as an ensemble of classification and regression process.
- A number of decision trees are constructed using a random subsets of the training data sets.
- A large collection of decision trees provide higher accuracy of results.
- The runtime of the algorithm is comparatively fast and also accommodates missing data.
- Random forest randomizes the algorithm and not the training data set. The decision class is the mode of classes generated by decision trees.

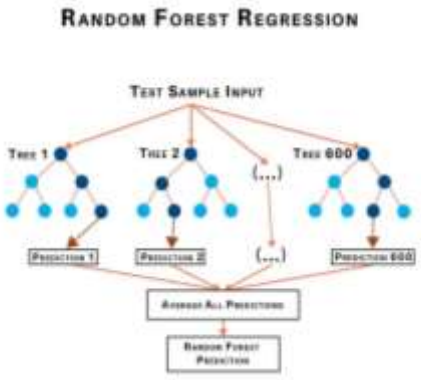


Figure.2: Random Forest Regression.

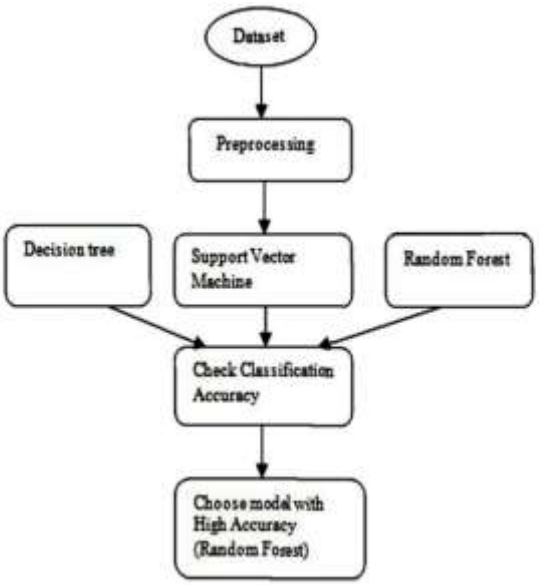


Figure.3: CKD prediction using machine learning models.

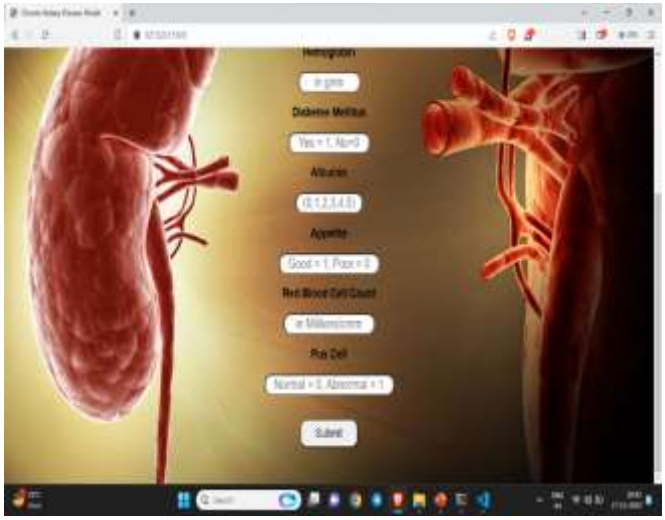
V. EXPERIMENTAL RESULTS

The experimental results showcase the robust performance of the implemented Random Forest Regression in chronic kidney disease prediction. The evaluation methodology employed rigorous metrics, including accuracy, precision, recall and f1-score metrics were calculated for each class, providing insights into the model's performance across different categories.

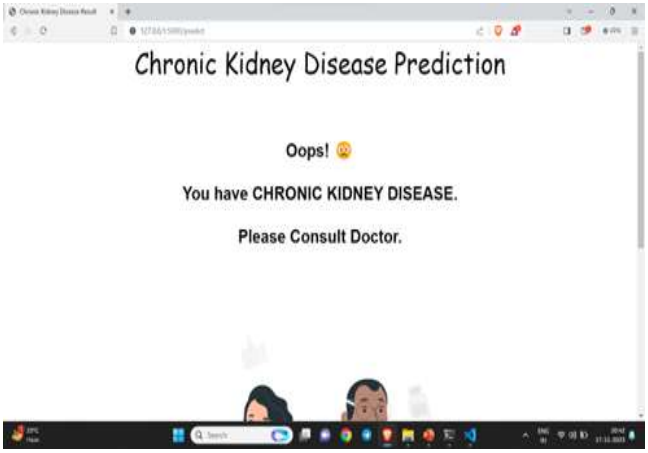
```
# Importing Performance Metrics:
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report


# RandomForestClassifier:
from sklearn.ensemble import RandomForestClassifier
RandomForest = RandomForestClassifier()
RandomForest = RandomForest.fit(X_train,y_train)

# Predictions:
y_pred = RandomForest.predict(X_test)

# Performance:
print('Accuracy:', accuracy_score(y_test,y_pred))
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
```

```
Accuracy: 0.975
[[55  3]
 [ 0 62]]
              precision    recall  f1-score   support

           0       1.00      0.95      0.97        58
           1       0.95      1.00      0.98        62

    accuracy                           0.97       120
   macro avg       0.98      0.97      0.97       120
weighted avg       0.98      0.97      0.97       120
```

1. The displayed below images shows the first/front page of the website . It shows the attributes to which proper values should be given so as to produce accurate results.



2. Once the user enters the proper values to the mentioned attributes , below is the page that will be displayed where the result of chronic kidney disease is predicted**.**



- Adding new book into the shop.
- Deleting the book which are not accessible in shop.
- Deleting the unauthenticated client the site of the book
- Adding another classification of the books

- Adding new book into the shop.
- Deleting the book which are not accessible in shop.
- Deleting the unauthenticated client the site of the book
- Adding another classification of the

VI. CONCLUSION

This paper elaborates the proposed system consisting of 4 main modules, which are data preprocessing, feature extraction, defining zones based on blood potassium level, diet recommendation module. Dataset has 25 main features, appropriate statistical method like regression extracts highly effective parameters in the decision of CKD detection. Main aim of machine learning module is to label the patient's CKD status,which is done using machine learning algorithm like Random Forest Regression. Also the main objective of this study was to predict patients with CKD using less number attributes while maintaining a higher accuracy. Here we obtain an accuracy of about 97 percentage.

VII. REFERENCES

1) "Predicting Chronic Kidney Disease from Electronic Health Records: A Machine Learning Approach" by Luis A. Martínez-Amezaga, et al. (Published in Nephron, 2020)
2) "Predicting chronic kidney disease using machine learning techniques: A systematic review of the literature" by Md Mohaimenul Islam, et al. (Published in Journal of Medical Internet Research, 2019)
3) "A machine learning approach for prediction of chronic kidney disease using cloud computing" by Irfan Ullah, et al. (Published in Journal of Ambient Intelligence and Humanized Computing, 2019)
4) "Predicting chronic kidney disease using machine learning techniques" by Rakesh Aggarwal, et al. (Published in International Journal of Computer Applications, 2014)
5) "Predicting Chronic Kidney Disease Using Random Forest Models" by Vincent Ee, et al. (Published in Nephron, 2018)
6) "Prediction of chronic kidney disease based on data mining classification algorithms" by Cen Wu, et al. (Published in Nephrology Dialysis Transplantation, 2017)
7) J. Snegha, "Chronic Kidney Disease Prediction using Data Mining", International Conference on Emerging Trends, 2020.
8) Baisakhi Chakraborty, "Development of Chronic Kidney Disease Prediction Using Machine Learning", International Conference on Intelligent Data Communication Technologies, 2019.
9) Guneet Kaur, "Predict Chronic Kidney Disease using Data Mining in Hadoop, International Conference on Inventive Computing and Informatics, 2017.
10) T. F. T. N. W. C. Anusorn Charleonnan, "Predictive Analytics for Chronic Kidney Disease," in The 2016 Management and Innovation Technology International Conference (MITiCON-2016), 2016.