Renal Disease Predicition using k-means clustering algorithm

|  |  |
| --- | --- |
| Meghamala Kunderu  *Department of Computer Science and Engineering*  *Bapatla Engineering College*  *(Autonomous)*  *(Affiliated to Acharya Nagarjuna University)*  *Bapatla, India*  Bhanu Prakash Kanumuri  *Department of Computer Science and Engineering*  *Bapatla Engineering College*  *(Autonomous)*  *(Affiliated to Acharya Nagarjuna University)*  *Bapatla, India* | Bhargavi Peetha  *Department of Computer Science and Engineering*  *Bapatla Engineering College*  *(Autonomous)*  *(Affiliated to Acharya Nagarjuna University)*  *Bapatla, India*  Navya Sri Kakarla  *Department of Computer Science and Engineering*  *Bapatla Engineering College*  *(Autonomous)*  *(Affiliated to Acharya Nagarjuna University)*  *Bapatla, India* |

***Abstract* - A novel approach for predicting renal diseases using machine learning techniques, specifically K-Means clustering, is proposed. The project aims to develop a predictive model that analyzes renal health datasets to identify distinct clusters of patients based on relevant features like demographic information, medical history, and clinical measurements. By partitioning the data into clusters, the model seeks to uncover hidden patterns and associations indicative of renal disease onset or progression. A comprehensive dataset comprising anonymized patient records will be utilized, and through rigorous preprocessing and feature engineering, the dataset's quality and relevance will be enhanced. The K-Means algorithm will then be applied to segment the dataset into clusters, each representing a subgroup of patients with similar characteristics. The outcomes of this research have significant implications for early detection and management of renal diseases. By leveraging K-Means clustering, the predictive model aims to provide healthcare practitioners with insights into patient stratification and risk assessment, facilitating personalized treatment strategies and improving overall patient outcomes. This approach has the potential to enhance early intervention efforts and ultimately reduce the burden of renal diseases on healthcare systems and individuals worldwide.**

**Top of Form**

***Keywords—* Renal disease, Unsupervised, PCA, K-means, Healthcare, CKD, Kidney disease, Clustering.**

# Introduction

Chronic Kidney Disease (CKD) is a significant health concern affecting millions worldwide, with its prevalence steadily rising over the past few decades. Early detection and timely intervention are crucial for effective management and prevention of complications associated with CKD.

“Around 10% of the Indian population suffers from CKD and every year over 1 lakh cases of renal failure are reported.” – Nephrologist Dr. Gurudev K C said. The kidneys’ main job is to get rid of waste and extra water from the blood. Renal disease means gradual failure of kidneys. This is two types- Acute Kidney Injury (AKI) and Chronic Kidney Disease (CKD).

AKI occurs when the kidneys suddenly fail due to an injury, medication or illness. Some disease and conditions can damage your kidneys and lead to AKI. Some examples include: a type of severe, life-threatening infection called “sepsis,” a type of cancer called “multiple myeloma”. AKI includes 4 stages: i) initiation, ii) oligo-anuria, iii) polyuria and iv) restitution. In this dynamic process, clinical signs of renal dysfunction emerge during stage 2.

CKD is the gradual loss of kidney function mainly caused by high blood pressure, diabetes and an inflammatory condition known as glomerulonephritis. This condition is called chronic because the damage happens slowly over a long period of time.

We can also identify in which stage of CKD the person is in according to the GFR (Glomerular Filtration Rate). CKD includes 5 stages: 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). To figure out eGFR (estimated GFR), the doctor usually uses the result of blood test that measures your creatinine level. The use of GFR as the best indicator of renal function to identify different stages of CKD with each successive stage defining a more severe decrease in GFR and the last stage defining kidney failure with a GFR <15 ml/min/1.73 m2 [4] were also advocated.

Symptoms of CKD are nausea, vomiting, loss of appetite, fatigue and weakness, sleep problems, urinating more or less, decreased mental sharpness, muscle cramps, swelling of feet and ankles, dry, itchy skin, high blood pressure that’s difficult to control, shortness of breath if fluid builds up in the lungs, chest pain if fluid builds up around the lining of the heart.

# LITERATURE SURVEY

An Approach for Classification and Detection of Chronic Kidney Disease (CKD) Using Machine Learning Algorithms was developed by O. Abuomar and O. Abuomar.[1] The urgent need for early detection of chronic kidney disease (CKD) using machine learning (ML) techniques. It successfully classifies CKD data and clusters patients based on common disease-contributing features. Decision Tree (DT) classification and Self-Organizing Maps (SOMs) clustering perform best, with DT achieving 98.8% accuracy. Minimum feature threshold values are identified, aiding healthcare practitioners in targeted testing. Neural Networks predict CKD with 85.0% accuracy, demonstrating the effectiveness of unsupervised ML techniques. This research serves as a pivotal starting point for implementing targeted testing protocols in healthcare systems for early CKD detection and management.

Chronic kidney disease (CKD) is a global healthcare challenge with significant morbidity and mortality. Early identification of high-risk CKD patients for end-stage kidney disease (ESKD) is crucial.[2] While traditional statistical models lack generalizability, machine learning (ML) offers improved accuracy. In a recent study of 748 CKD patients, machine learning algorithms, including random forest and logistic regression, were used to predict ESKD risk. The random forest algorithm showed the best performance (AUC 0.81), but the simpler Kidney Failure Risk Equation (KFRE) exhibited comparable accuracy, specificity, and precision, providing a promising approach for ESKD risk prediction.

Improvement of the handover performance and channel allocation scheme using fuzzy logic, artificial neural network and neuro-fuzzy system to reduce call drop in cellular network. Data by Islam M.A., Hasan M.R., Begum A [5] mining is the process of using specialized software to find hidden information in a large set of data. Data mining techniques are linked to each other and used in a wide range of places and situations. With data mining technologies, we can make predictions, sort the data, filter it, and put it into groups. The goal of the algorithm is to process a training set that has a collection of attributes and targets, and the objective describes how this should be done. If the dataset is very big, data mining is a good way to find patterns in it. If the dataset is very small, however, we can still reach the same goal with the help of machine learning. Data analysis and pattern recognition are 2 further capabilities of machine learning.[3] Because there is such a wide diversity of health datasets, machine learning algorithms are the most appropriate method for enhancing the accuracy of diagnosis prediction.[4][5] The prevalence of machine learning algorithms in the healthcare industry is growing as a direct result of the rapid growth of electronic healthcare datasets[6].

the UN's Sustainable Development Goals, focusing on good health and well-being, the detection and management of non-communicable diseases like chronic kidney disease (CKD) are paramount. CKD affects a significant portion of the global population and contributes to morbidity and mortality. Early detection of CKD stages is crucial to mitigate health complications and improve patient outcomes. Various machine learning techniques, including Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT),[7] have been employed to predict CKD stages. Feature selection methods such as recursive feature elimination with cross-validation (RFECV) have been utilized to enhance model performance. Results indicate that RF, particularly with RFECV, exhibits superior performance compared to SVM and DT, with high accuracy and F1-score values[8]. Multi-class classification is deemed essential for identifying disease stages and guiding appropriate treatments, underscoring the significance of early prediction in preventing CKD progression to kidney failure.

Salekin and Stankovic [9] did evaluation of classifiers such as K-NN, RF and ANN on a dataset of 400. Wrapper feature selection were implemented and five features were selected for model construction in the study. The highest classification accuracy is 98% by RF and a RMSE of 0.11. S. Tekale et al. [10] worked on “Prediction of Chronic Kidney Disease Using Machine Learning Algorithm” with a dataset consists of 400 instances and 14 features. They have used decision tree and support vector machine. The dataset has been preprocessed and the number of features has been reduced from 25 to 14. SVM is stated as a better model with an accuracy of 96.75%. Xiao et  al. [11] proposed prediction of  chronic kidney disease progression using logistic regression, Elastic Net, lasso regression, ridge regression, support vector machine, random forest, XGBoost, neural network and k-nearest neighbor and compared the models based on their performance. They have used 551 patients’ history data with proteinuria with 18 features and classified the outcome as mild, moderate, severe. They have concluded that Logistic regression performed better with AUC of 0.873, sensitivity and specificity of 0.83 and 0.82, respectively.

Mohammed and Beshah [12] conducted their research on developing a self-learning knowledge-based system for diagnosis and treatment of the first three stages of chronic kidney have been conducted using machine learning. A small number of data have been used in this research and they have developed prototype which enables the patient to query KBS to see the delivery of advice. They used decision tree in order to generate the rules. The overall performance of the prototype has been stated as 91% accurate.

1. DATASET AND PRE-PROCESSING

The dataset contains data of 400 patients. It consists of 24 clinical attributes and 1 class attribute. The datasets consist of 250 CKD cases and 150 Non-CKD cases. Missing data is a significant problem in real-world datasets, especially in the medical field. On average, every patient record and attribute have a few missing values.

**Dataset description:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable Name** | **Role** | **Type** | **Description** | **Demographic** | **Units** | **Missing**  **Values** |
| Age | Feature | Integer |  | Age | Year | Yes |
| Bp | Feature | Integer | Blood pressure |  | Mm/Hg | Yes |
| Sg | Feature | Categorical | Specific gravity |  |  | Yes |
| Al | Feature | Categorical | Albumin |  |  | Yes |
| su | Feature | Categorical | Sugar |  |  | Yes |
| Rbc | Feature | Binary | Red blood cells |  |  | Yes |
| Pc | Feature | Binary | Pus cell |  |  | Yes |
| Pcc | Feature | Binary | Pus cell clumps |  |  | Yes |
| Ba | Feature | Binary | Bacteria |  |  | Yes |
| Bgr | Feature | Integer | Blood glucose random |  | Mgs/dl | Yes |
| Bu | Feature | Integer | Blood urea |  | Mgs/dl | Yes |
| Sc | Feature | Continuous | Serum creatinine |  | Mgs/dl | Yes |
| Sod | Feature | Integer | Sodium |  | mEq/L | Yes |
| Pot | Feature | Continuous | Potassium |  | mEq/L | Yes |
| Hemo | Feature | Continuous | Hemoglobin |  | Gms | Yes |
| Pcv | Feature | Integer | Packed cell volume |  |  | Yes |
| Wbcc | Feature | Integer | White blood cell count |  | Cells  /cmm | Yes |
| Rbcc | Feature | Continuous | Red blood cell count |  | Millions /cmm | Yes |
| Htn | Feature | Binary | Hypertension |  |  | Yes |
| Dm | Feature | Binary | Diabetes mellitus |  |  | Yes |
| Cad | Feature | Binary | Coronary artery disease |  |  | Yes |
| Appet | Feature | Binary | Appetite |  |  | Yes |
| Pe | Feature | Binary | Pedal edema |  |  | Yes |
| Ane | Feature | Binary | Anemia |  |  | Yes |
| class | Target | Binary | Ckd or notckd |  |  | No |

## **Pre-processing**

Firstly, in our project after a bit of exploration, data cleaning is performed in a way to remove typos, null values, to replace missing values. Next, one-hot encoding is performed, which is a encoding technique to convert each categorical value into new categorical column and assign a binary value of 1 or 0 to those columns. One-hot encoding is perfect here as every categorical feature has only **2** values, which means we won't be increasing dimensions.

Secondly, we are performing knn-imputation, which is designed to find k-nearest neighbors for missing incomplete instance from all complete instances (without missing values) in a given dataset and then filling the missing datum with the most frequent one occurring in the neighbors with the target feature is categorical referred to as majority rule or with the mean of the neighbors if the target feature is numerical and referred to mean rule.

Next, Exploratory data analysis is done on the data, which is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected**.**

The MinMax Scaler is a form of normalization that scales the values between 0 and 1. It gets its name because the maximum and minimum values of the feature are used for normalization. The concrete formal of the MinMax Scaler is:

xscaled=x–xmin/xmax–xmin

**Standard deviation is a measure used in statistics to understand how the data points in a set are spread out from the**[**mean**](https://www.geeksforgeeks.org/mean/)**value.**It indicates**the extent of the data’s variation and shows how far individual data points deviate from the average.**

**Standard Deviation = 1/n∑in (xi – x̄)2**

1. METHODOLOGY

# Principal component analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique used to simplify complex data while retaining its essential features. By transforming the data into a new coordinate system, PCA identifies the directions (principal components) that capture the maximum variance in the data. These principal components are orthogonal to each other, meaning they are uncorrelated.

PCA is widely used in various fields such as image processing, pattern recognition, and finance. It helps in visualizing high-dimensional data, identifying patterns, reducing noise, and speeding up machine learning algorithms by reducing the number of features.

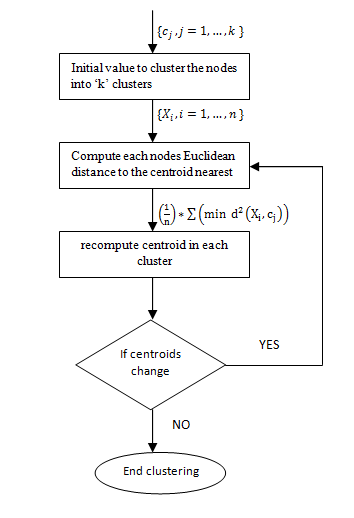
In essence, PCA offers a powerful way to summarize and compress data while preserving its inherent structure, making it a valuable tool in data analysis and machine learning.

[Unsupervised Machine Learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/)is the process of teaching a computer to use unlabeled, unclassified data and enabling the algorithm to operate on that data without supervision. Without any previous data training, the machine’s job in this case is to organize unsorted data according to parallels, patterns, and variations.

[K-Means Clustering](https://www.geeksforgeeks.org/k-means-clustering-introduction/) is an[Unsupervised Machine Learning](https://www.geeksforgeeks.org/ml-types-learning-part-2/) algorithm, which groups the unlabeled dataset into different clusters. The article aims to explore the fundamentals and working of k mean clustering along with the implementation.

K means clustering, assigns data points to one of the K clusters depending on their distance from the center of the clusters. It starts by randomly assigning the clusters centroid in the space. Then each data point assign to one of the cluster based on its distance from centroid of the cluster. After assigning each point to one of the cluster, new cluster centroids are assigned. This process runs iteratively until it finds good cluster. In the analysis we assume that number of cluster is given in advanced and we have to put points in one of the group.

In some cases, K is not clearly defined, and we have to think about the optimal number of K. K Means clustering performs best data is well separated. When data points overlapped this clustering is not suitable. K Means is faster as compare to other clustering technique. It provides strong coupling between the data points. K Means cluster do not provide clear information regarding the quality of clusters.



## **Pseudocode of k-means clustering algorithm:**

## Choose k random points as initial centroids.

## Assign each data point to the nearest centroid to form K clusters.

## Recalculate centroids as the mean of data points in each cluster.

## Continue until centroids converge (minimal change) or a maximum number of iterations is reached.

## Data points are assigned to the cluster represented by the nearest centroid.

## This iterative process partitions the data into k clusters, where each data point belongs to the cluster with the nearest centroid.

1. RESULTS AND CONCLUSION

There are only a limited number of studies, using unsupervised systems and algorithms to solve the issue of early detection of CKD. However, in detecting CKD, there were some studies based on semi-supervised and supervised

learning which were worth mentioning.

We are achieving a 99.3% accuracy rate in renal disease prediction using the k-means algorithm is a significant accomplishment, indicating the effectiveness of the model accurately identifying patterns and making predictions based on patient data. In conclusion, renal disease prediction using k-means clustering algorithm was able to identify patterns or groupings within the data related to renal disease.

Also identify any correlations discovered between features and renal disease and the potential utility of these insights for medical diagnosis or treatment planning.

1. ACKNOWLEDGMENT

The authors would like to express our sincere gratitude to M. M. MEERA DURGA for guidance, support throughout the duration of this project. Thank you to all the research paper authors whose work contributed to the success of our project.

# REFERENCES

1. O. Abuomar and P. Sogbe, "Classification and Detection of Chronic Kidney Disease (CKD) Using Machine Learning Algorithms," 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 2021, pp. 1-8, doi: 10.1109/ICECET52533.2021.9698666.
2. Bai Q, Su C, Tang W, Li Y. Machine learning to predict end stage kidney disease in chronic kidney disease. Sci Rep. 2022 May 19;12(1):8377. doi: 10.1038/s41598-022-12316-z. PMID: 35589908; PMCID: PMC9120106.
3. Akhter T., Islam M.A., Islam S. Artificial neural network based covid-19 suspected area identification. *J Eng Adv.*2020;1:188–194.
4. Islam M.A., Akter S., Hossen M.S., Keya S.A., Tisha S.A., Hossain S. *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* IEEE; 2020. Risk factor prediction of chronic kidney disease based on machine learning algorithms; pp. 952–957.
5. Islam M.A., Hasan M.R., Begum A. Improvement of the handover performance and channel allocation scheme using fuzzy logic, artificial neural network and neuro-fuzzy system to reduce call drop in cellular network. *J Eng Adv.*2020;1:130–138
6. Chittora P., Chaurasia S., Chakrabarti P., et al. Prediction of chronic kidney disease-a machine learning perspective. *IEEE Access.*2021;9:17312–17334.
7. Agrawal A, Agrawal H, Mittal S, Sharma M. Disease Prediction Using Machine Learning. SSRN Electron J. 2018;5:6937–8.
8. Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N. Predictive analytics for chronic kidney disease using machine learning techniques. Manag Innov Technol Int Conf MITiCON. 2016;80–83:2017.
9. Salekin A, Stankovic J. Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. In: Proc. - 2016 IEEE Int. Conf. Healthc. Informatics, ICHI 2016, pp. 262–270, 2016.
10. Tekale S, Shingavi P, Wandhekar S, Chatorikar A. Prediction of chronic kidney disease using machine learning algorithm. Disease. 2018;7(10):92–6.
11. Xiao J, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med. 2019;17(1):1–13.
12. Priyanka K, Science BC. Chronic kidney disease prediction based on naive Bayes technique. 2019. p. 1653–9.