IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Building A Medical System To Extract Important Attributes For CKD Prediction Using ML Algorithms

Manimala S [1] , Sanjana V [2]

Department of Computer Science and Engineering,JSS Science and Technology University, Mysuru, Karnataka, India

**Abstract** In the modern world, chronic kidney disease has become one of the most hazardous diseases. CKD is a condition in which the kidney cannot perform the proper filtering of the blood or it stooped working completely causing the left toxic to enter into the blood, leading to the patient's death. It is likely impossible to detect CKD in the early stages, and it is very difficult to save patient's lives in the last stage of CKD. A patient's life can be saved by renal transplant or the early detection the CKD. As per the medical survey there are totally 24 different attributes used to predict CKD. The attributes include gender, age, BP, sc, white blood cell, red blood cell, sugar tested, pus cells, bacteria, blood urea, sodium, potassium etc. There are so many research works on CKD prediction using machine learning algorithms. All these works used all 24 parameters for CKD prediction and model built for static datasets. In this proposed system we aim at identifying the most important attributes or parameters for CKD prediction so as to improvise the efficiency of the CKD prediction system. We'll use a variety of supervised machine learning methods before deciding the one best for the model. In our project work we build an application with model that can predict CKD disease with important attributes and provides doctors with the information of how to handle patients and treat them better. The proposed system is a real time medical system useful for hospitals and doctors and built using Microsoft tools such as Visual Studio tool and SQL Server tool.

*Keywords: CKD, Stages, Data Science, Machine Learning, Naïve Bayes, KNN Algorithm, Brute Force, GFR.*

## Introduction

The health-care 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. Chronic Kidney Disease(CKD), is delineated by the gradual loss of kidney function. Kidneys filter wastes and excess fluids from the blood, which are then excreted in the urine. If this disease gets worsened, wastes can accumulate in the blood and can cause difficulties like high blood pressure, anaemia, weakening of bones, poor nutritional health and nerve damage. Also, kidney disease increases the risk of having heart and blood vessel diseases.

The harmful outcomes can be avoided and prevented by early detection, according to researches conducted. The 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 2025and also the count of patient with hypertension is expected to double from 2000 to 2025, hence these will make India the reservoir of CKD [1]. 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 so that the undiagnosed CKD can be identified, predicting the likelihood that patients will develop chronic disease, and present patient-specific prevention interventions. Accurate predictive models can be created using Machine learning techniques to

assist health systems lowering risks and eventually improving the 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. 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 a potential CKD patient in stage 2 will be given a restricted and strict diet. Keeping the balance of minerals, electrolytes, and liquids inside body will be difficult for stage 3 to 5 patient. Therefore, they have to be under proper dietary guidance.

An important diet for a renal improvement and prevent further harm is essential, which also helps in keeping balance of electrolytes and water in the body. Other than stages of severity, many other factors such as the blood potassium level, urea level, calcium level, phosphorous level and so on will contribute in shaping the diet. In this study, to identify suitable diet plan for a CKD patient the main focus will be on blood potassium level.

## 1 Related Work

Recent research has explored various methodologies for predicting and analyzing Chronic Kidney Disease (CKD) using machine learning and data mining techniques. This section reviews relevant studies in this field. Bilal Khan et al. (2020) conducted an empirical evaluation of several machine learning techniques for CKD prediction, including NBTree, J48, Support Vector Machine (SVM), Logistic Regression, and Multi-layer Perceptron [1]. Their study highlights that the algorithms used to generate graphical outputs, which are not suitable for real-time applications. Additionally, the research was limited by the use of small datasets and predictions based solely on static data. Veenita Kunwar et al. (2016) explored CKD analysis using data mining classification techniques and implemented these techniques using tools like Rapid Miner [2]. Although the tools facilitate easy result generation, their study points out the challenge in testing these tools effectively, which can impact the robustness of the results. In another approach, Navaneeth Bhaskar and Suchetha M. (2019) developed a deep learning-based system for CKD detection, incorporating a Convolutional Neural Network (CNN) algorithm combined with a Support Vector Machine (SVM) classifier [3]. Their methodology, however, faces limitations such as relying on image data for predictions, leading to less accurate results, extended data processing times, and overall inefficiency. Anusorn Charleonnan et al. (2016) performed predictive analysis for CKD using Regression and SVM techniques [4]. They observed that these techniques produce graphical outputs, which can make distinguishing results difficult. Moreover, their methods are not suitable for real-time clinical applications, limiting their practical use. A duplicate entry by Veenita Kunwar et al. (2016) focuses on CKD analysis using static data from the UCI Machine Learning Repository [5]. This study is limited by the absence of real-time data and is based on a relatively small dataset of only 400 samples, restricting its generalizability. Lastly, Devika R et al. (2019) conducted a comparative study of classifiers for CKD prediction, evaluating Naive Bayes, K-Nearest Neighbour (KNN), and Random Forest classifiers [6]. They utilized static data from the UCI Machine Learning Repository, noting limitations such as the use of a limited dataset and the lack of real-time application, which diminishes its utility in hospital settings.

## 2 Proposed Work

Chronic kidney disease (CKD) has become a global health issue and is an area of concern. It is a condition where kidneys become damaged and cannot filter toxic wastes in the body. Our work predominantly focuses on detecting life threatening disease like Chronic Kidney Disease (CKD) using Classification algorithms. The proposed system is an automation for chronic kidney disease prediction using classification techniques. The proposed system extracts the features which are responsible for CKD, then machine learning process can automate the classification of the chronic kidney disease in different stages according to its severity. The objective is to use machine learning algorithm and suggest suitable diet plan for CKD patient using classification algorithm on medical test records. The system uses old data from "*UCI Repository*" and uses tools such as "Visual Studio" and "SQL Server" to develop application. The system is a real time application useful for doctors to identify CKD and related stages and recommending the suitable diet for the patients.

| Attribute Name | Value Range | Description |
|---|---|---|
| age | 2, ..., 90 | age |
| bp | 50, ..., 180 | blood pressure |
| sg | 1.005,1.010,1.015,1.020,1.025 | specific gravity |
| al | 0,1,2,3,4,5 | albumin |
| su | 0,1,2,3,4,5 | sugar |
| rbc | 2.1, ..., 8 | red blood cells |
| pc | normal,abnormal | pus cell |
| pcc | present,notpresent | pus cell clumps |
| ba | present,notpresent | bacteria |
| bgr | 22, ..., 490 | blood glucose random |
| bu | 1.5, ..., 391 | blood urea |
| sc | 0.4, ..., 76 | serum creatinine |
| sod | 4.5, ..., 163 | sodium |
| pot | 2.5, ..., 47 | potassium |
| hemo | 3.1, ..., 17.8 | hemoglobin |
| pcv | 9, ..., 54 | packed cell volume |
| wc | 2,200, ..., 26400 | white blood cell count |
| rc | 2.1, ..., 8 | red blood cell count |
| htn | yes, no | hypertension |
| dm | yes, no | diabetes mellitus |
| cad | yes, no | coronary artery disease |
| appet | good,poor | appetite |
| pe | yes, no | pedal edema |
| ane | yes, no | anemia |
| class | ckd,notckd | class |

Table 1: Parameters List

## 2 Methodology

### 2.1 Machine Learning

Machine learning is a process of studying a system based on data. Machine learning is a part of data science where we use machine learning algorithms to process data.

### 2.2 Supervised Learning Technique

It is a predictive model used for the tasks where it involves prediction of one value using other values in the dataset. The supervised learning will have predefined labels. It classifies an object based on the parameters to one of the predefined set of labels.

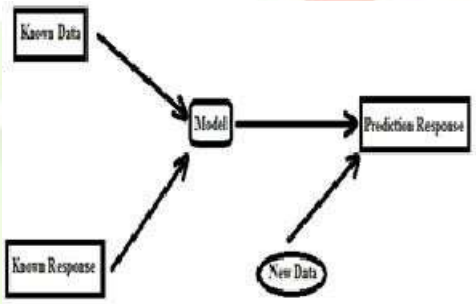


**Fig. 1: The Predictive Model**

We have many algorithms to build model in supervised learning such as KNN, Naive bayes, Decision Tree, ID3, Random Forest, SVM, Regression techniques etc. Depending on the requirement, labels, parameters

and dataset we select the appropriate algorithm for predictions. Algorithm is used to build a model that makes predictions based on evidence in the presence of uncertainty. In this project for prediction we make use to "*Bayesian Classifier or KNN algorithm*" which is an efficient and works fine for all different sets of parameters. It also generates accurate results.

### 2.3 Classification Rules

Basically classification is used to classify each item in a set of data into one of the predefined set of classes or groups. The Bayesian Algorithm or KNN is used to predict CKD. GFR used for Stage Prediction.

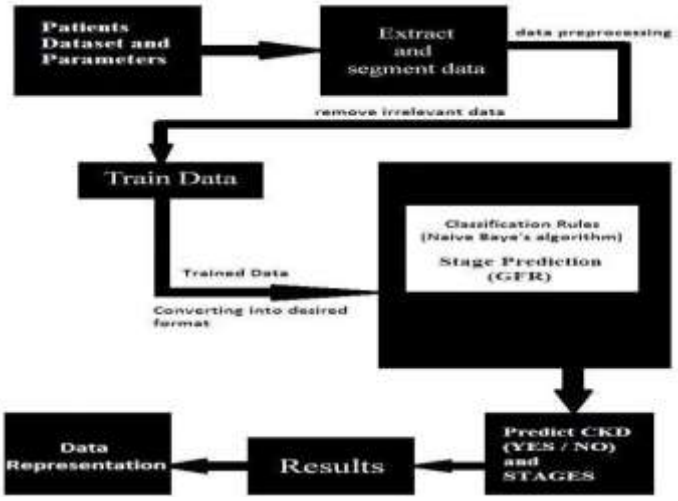


**Fig. 2: The Proposed Model**

### 2.4 KNN Algorithm

The steps involved in the working of k-nearest neighbor algorithm are as follows: Step 1: Choose the neighbour with number K.
Step 2: Determine the Euclidean distance for K number of neighbours.
Step 3: Using the computed Euclidean distance, select the K closest neighbours.
Step 4: Determine how many data points are in each category among these k neighbours.
Step 5: Put the additional data points to the category where the neighbour count is at its highest. Step 6: We've finished our model.

### 2.5 Naïve Bayes Algorithm

The steps involved in the working of k-nearest neighbor algorithm are as follows:

Step 1: Scanning the dataset (storage servers) to extract the information needed for data miningfrom servers (cloud, excel sheets, databases, etc.).
Step 2: Determine the likelihood of each attribute value in step two. [n, n_c, m, p] Here, we usethe following formula to determine the likelihood of occurrence for each property.
Step 3: Utilise the equations P (subject valuevj)/attribute value(ai)=(n_c+mp)/(n+m)

where n is the total number of training cases where v = vj nc is the number of instances where v = vj and a=aiP(aijvj) = p = a priori estimate
The corresponding sample size, m

Step 4: Apply p to the probability. Here, we multiply the values of each attribute by p for eachclass, and the final results are utilised for classification.
Step 5: Sort the attribute values into one of the predetermined classes by comparing the values.

### 2.6 Brute Force Algorithm

**Filter Method [Static/Manual Fixing] – Brute Force**

Step 1: System Training (obtaining every feature from the storage server)
Step 2: Manually fix the threshold value [Determine how many features to use for classification] Step 3: Determine the Gain [Find the gain [number of occurrences] for each feature a.]
Step 4: Determine the Model Score as 2.0 * Gain(feature)/Threshold value
Step 5: Use decreasing order to extract the features [Let the highest gain attribute be f_best]

## 3 Conclusion

This project is a medical sector application which helps the medical practitioners in predicting the CKD disease based on the CKD parameters. It is automation for CKD disease prediction and it identifies the disease, its types and complications from the clinical database in an efficient and an economically faster manner. It is successfully accomplished by applying the Naïve Bayes algorithm for classification. This classification technique comes under data mining technology. This algorithm takes CKD parameters as input and predicts the disease based on old CKD patients data.

## 4 Future Enhancements

- **SMS/Email Module**

In the proposed system, admin assigns ID and password to doctors and receptionists and is intimated manually, so we can add SMS/Email module as a future enhancement where doctors and receptionists receive an SMS or Email regarding the ID and password.

- **Query Module**

We can add the query module as a future enhancement to the application where doctor, receptionist and admin of the application can interact with each other.

## 5 References

[1] A. S. Levey, R. Atkins, and J . Coresh, “Chronic kidney disease as a global public health problem: approaches and initiatives - a position statement from Kidney Disease Improving Global Outcomes”, Kidney International, vol. 72 no.3, pp. 247-259, Aug 2007.

[2] V . Jha, G. Garcia, and K. Iseki, “Chronic kidney disease: global dimension and perspectives”, Lancet, vol. 382 no. 9888, pp. 260-272, Jul 2013.

[3] K.R Lakshmi, Y. Nagesh, and M. VeeraKrishna, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", International Journal of Advances in Engineering and Technology, vol.7, no.1, pp. 242-254, March 2014.

[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] J. Levman, T. Leung, P. Causer, D. Plewes, and A. L. Martel, "Classification of dynamic contrast-enhanced magnetic resonance breast lesions by support vector machines," Medical Imaging, IEEE Transactions on, vol. 27, pp. 688-696, 2008.

[8] 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.

[9] 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.

[10] V. Vapnik, "The nature of statistical learning theory" Springer Science and Business Media, 2013.

[11] 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.

[12] K. Tufan, "Noninvasive diagnosis of atherosclerosis by using empirical mode decom-position, singular spectral analysis, and support vector machines," Biomedical Research, vol. 24, pp. 303-313, 2013.

[13] 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.

[14] Sh. Shamiluulu, M.M. Boukar, Z. Yussupova. “Medical Tool for Assisting Patients in Kazakhstan Polyclinics”. Proceedings: 11th IEEE International Conference on Application of Information and Communication Technologies (ICECCO 2017). Abuja, Nigeria pp: 80-84