*International Journal of Engineering & Technology*, 7 (3.10) (2018) 27-30

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*

***Research paper***

# Data Mining Techniques to Predict Chronic Kidney Disease and its Stages

**Ms. Nisarga P[1], Ms. Kanchana V[2]**

*[1,2]Department of Computer Science*
*[1,2]Amrita School of Arts and Sciences, Mysuru*
*[1,2]Amrita Vishwa Vidyapeetham*
*India*
*[*]Corresponding author: E-mail: nisarga95rao@gmail.com[1]*

**Abstract**

Chronic Kidney Disease incorporates the state where the kidneys fail to function and reduce the potential to keep a person suffering from the disease healthy. When the condition of the kidneys gets worse, the wastes in the blood are formed in high level. Data mining has been a present pattern for accomplishing analytic outcomes. Colossal measure of un-mined data is gathered by the human services industry so as to find concealed data for powerful analysis and basic leadership. Data mining is the way towards extricating concealed data from gigantic datasets. The goal of our paper is to anticipate CKD utilizing the classification strategy Naïve Bayes. The phases of CKD are anticipated in the light of Glomerular Filtration Rate (GFR).

***Keywords***: *Chronic Kidney Disease, Naïve Bayes, Glomerular Filtration Rate.*

## 1. Introduction

These days, human service enterprises are giving a few advantages like fraud recognition in health care coverage, accessibility of therapeutic offices to patients at economical costs, ID of more astute treatment techniques, development of powerful social insurance arrangements, successful doctor's facility asset administration, better client connection, enhanced patient care and clinic contamination control. Illness recognition is likewise one of the noteworthy regions of research in medicinal. The approaches of data mining have turned out to be fundamental for social insurance industry in selling on choices in view of the investigation of the monstrous clinical information. It is the way toward extricating concealed data from enormous dataset. Methods like clustering, association, classification, and regression have been utilized by therapeutic field to identify and anticipate illness movement and to settle on choice with respect to the treatment given to the patients. The directed approach that dole out articles in gathering to target classes is the classification technique, the methodology that arranges the things or data into social affairs, the people which have no less than one trademark in like way. ANN, Naïve Bayes, SVM, Decision tree and so forth are the various techniques. Clustering unites the entities of same kind into a group. K-medoids, k-means, agglomerative, DBscan etc are some of the clustering techniques.

Data mining propounds certain equipments and mechanisms for applying the processed data, gives learning to medicinal services experts to settling on suitable choices and improving the execution of patient administration errands. Patients with comparative medical problems can be assembled together and compelling treatment designs could be proposed in view of patient's previous reports, physical examination, finding and past treatment designs. CKD has turned into a worldwide medical problem and is a region of concern. Here, the kidneys end up harm end and cannot channel poisonous squanders in the body. Our work overwhelmingly centers around distinguishing dangerous ailments like CKD by classification mechanism Naïve Bayes. The stage prediction is based on the Glomerular Filtration Rate. The concept can be implemented for a clinic or hospital for analyzing the CKD patients data. The concept can be implemented as an online health community system where patients can gather CKD and stages information. The concept can be implemented for a research department for analyzing the relationship between CKD and its different stages.

## 2. Related Work

Sunil D et al[1] brought up a work that spotlights on distinguishing hazardous sickness like CKD utilizing classification calculations. The framework is a computerization for incessant kidney infection expectation utilizing the classification strategy and counterfeit neural system methodology.

Mostafa GhannadRezaie et al [2] developed a system, allows to interact with the data mining procedure. Provides a place to manually generate generatedrules. The algorithm regulates the rule set about manipulation.

Jenn-Lung Su et al [3] introduced a system to access the execution of three diverse mainstream calculation with various therapeutic information and to discover impediments on applying for learning disclosure in medicinal database.

Paolo Bonato et al [4] work is to exhibit preparatory proof that information mining and counterfeit consciousness frameworks

may enable one to perceive the nearness and seriousness of engine variances in patients with Parkinson's ailment.

Wang et al [5] thought about the use of fluffy group investigation for restorative picture information mining. Decision tree computations are chosen for the therapeutic picture classifier.

Xing et al [6] brought up the work for predicting the CHD(coronary heart disease) was a challenging work for the improvement in medical field.

Sellappan Palaniappan et al [7] brought up a methodology to predict heart disease. The techniques used were neural network, decision trees, naïve bayes.

Heon Gyu Lee et al [8] brought up a new system to improve the different characteristics of HRV(heart rate variability). so ,determine cardiovascular disease.

K Srinivas et al [9] built a surveillance system and performed a survey to test cardiovascular disease rates comparing with other regions.

Sa'diyah Noor et al [10] proposed a work to recognize liver disease using 10 attributes comparing with different algorithms. NBTree gave the highest accuracy. Naïve Bayes was performed in quickest computation time.

Debabrata Pal et al [11] A computer aided screening system was brought up which assists the perception. Uses the limited resources to enhance the patient administration.

Sonam Nikhar et al [12] investigated developing and new zones of information mining methods utilized as a part of social insurance administration. This paper investigates distinctive information mining systems which are utilized as a part of medicinal services field for the estimation of heart ailments utilizing the information mining strategies.

Jenn Long Liu et al [13] brought up the metamorphic techniques of data mining algorithms to cluster the dataset present in the malady and depict the certainty.

Geeta yadav et al [14] aimed at detecting the people who were afflicted by PD at its high accuracy.

Ilayaraja M et al [15] built up a strategy to distinguish recurrence of ailments specifically topographical zone at given era with the guide of affiliation control based Apriori information mining method.

Shivagowri S et al [16] performed the inquiries about with numerous cross breed methods for diagnosing coronary illness. This paper manages a general review of utilizing the information mining in coronary illness outlook.

K.Vasantha Kokilam et al [17] brought up a system which contemplates to provide an overview on the evolution of existing techniques in organic databases that are existing.

Syed Umar Amin et al [18]aimed at building up a smart information mining framework in view of hereditary calculation upgraded neural systems for the forecast coronary illness in light of hazard variable classification.

Girija D K et al [19] proposed a famework utilizing MATLAB R2012a. in this paper Fibroid Disease Prediction Ssytem(FDPS) is developed. It predicts the probability of patient getting a fibroid illness.

Juliet Rani Rajan et al [20] a system was brought up that associated the development of information mining tools that would assist to classify patients into certain categories that might detect positive lung cancer.

## 3. Methodology

In order to predict the presence of Chronic Kidney Disease, a classification data mining technique is used that is, Naïve Bayes classifier which is highly scalable. Even if we are working on a dataset with millions of records with less constraint, the classifier can yield best results. The data is collected form UCI data repository, hospitals and acquired into database.

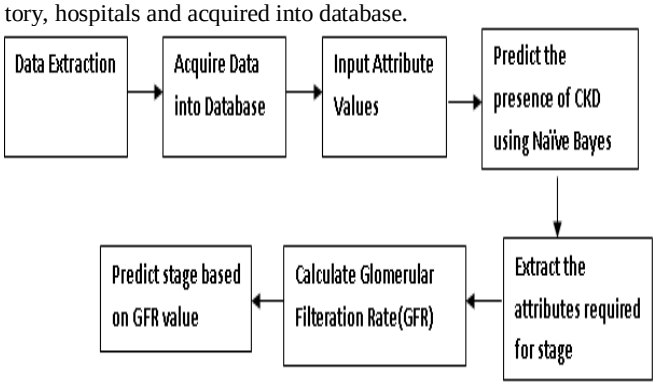


**Fig 1**: Architecture Diagram of the system

The proposed methodology adopts Naïve Bayes classification technique to predict chronic kidney disease. It contains 24 constraints(haemoglobin/WBC count/RBC count/packed cell volume/coronary artery disease/hypertension/pedal edema/ diabetes mellitus/ blood urea/ pus cells/ albumin/ pus cell clumps/ serum creatinine/ bacteria/ potassium/ blood glucose random/ sodium/blood pressure/ red blood cells/specific gravity/ age/appetite/anaemia/sugar) based on which the disease prediction is performed. Classification technique Naïve Bayes is performed to predict the presence of CKD with the above 24 constraints as input. If CKD is present, further the stage of CKD has to be predicted. There are five stages of CKD.

To figure out the stages, GFR (Glomerular Filtration Rate) is used, which is the best measure to detect the function of kidney. The rate of blood flow through the kidneys is the GFR. It can be varied depending upon age, gender, size, weight, and ethnicity. Creatinine is the main aspect for the estimation of GFR. To determine GFR, three constraints are extracted i.e., age, gender and serum creatinine. Creatinine is a waste product that forms during the malfunction of muscle tissue. Serum creatinine measures the amount of creatinine in the blood. It provides an estimation that how well the kidney is filtering. The typical range of creatinine:

**Table** 1: normal ranges of creatinine in male and female

| | |
|---|---|
| Male | 0.6 – 1.2 |
| Female | 0.5 – 1.1 |

The creatinine is measured in terms of mg/dl. Higher the blood creatinine level, lower the assessed GFR.
GRF is calculated using the mathematical formula:

$$GFR = X * (\frac{SC}{Y})^{Z} * 0.993^{age} \quad (1)$$

Where SC- Serum Creatinine
X,Y,Z are the variables used as given in the Table 2.

**Table 2**: values of the variables in the GFR formula

| Male | | Female | |
|---|---|---|---|
| SC<=0.9 | X=141 | SC>=0.7 | X=144 |
| | Y=0.9 | | Y=0.7 |
| | Z= -0.411 | | Z= -0.329 |

| | | | |
|---|---|---|---|
| SC>0.9 | X=141 | SC>0.7 | X=144 |
| | Y=0.9 | | Y=0.7 |
| | Z= -1.209 | | Z= -1.209 |

When the GFR is calculated, based on certain range, the stage is predicted.

**Table 3**: Specific ranges of CKD stages with kidney functioning.

| Stage | GFR Range | Kidney functioning |
|---|---|---|
| S1 | >90 | Normal |
| S2 | 60 to 89 | Moderate |
| S3 | 30 to 59 | Average |
| S4 | 15 to 29 | Critical |
| S5 | <15 | Kidney failure |

Finally, we identify the presence of Chronic Kidney Disease, and if the disease is present, the stage is predicted. The experimental results presented in this article show the efficiency of methods in the upcoming sections.

## 4. Experimental Results

To predict the presence of Chronic Kidney Disease, 24 attributes are required. Figure 2 illustrates the Graphical User Interface which accepts the input for the process of CKD prediction.

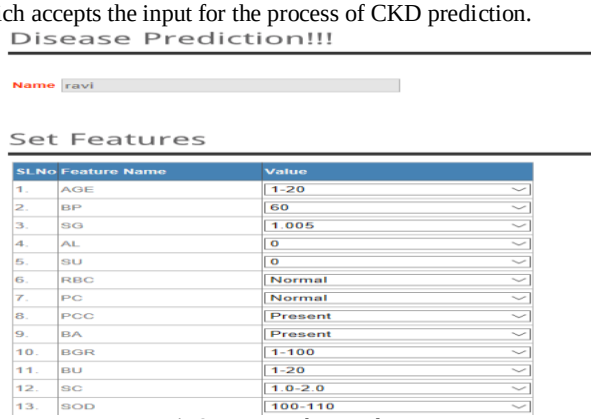


**Fig 2**: Input attributes values

If CKD is present, CKD prediction output is generated. Then, the stage is predicted. Figure 3 illustrates the GUI of the output of presence of CKD. Here, CKD is present, hence the stage prediction is performed.

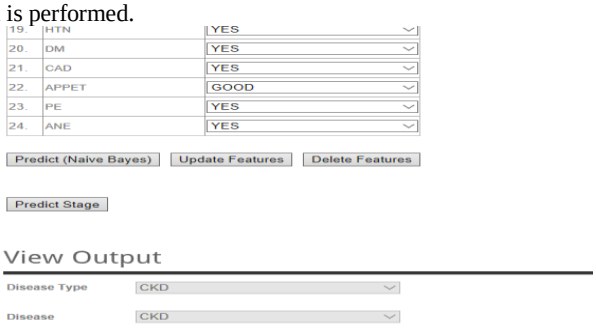


**Fig 3:** Output of prediction of CKD

For stage prediction, three attributes are extracted. They are age, gender and serum creatinine. Using the mathematical formula, GFR is calculated and the stage is predicted. Figure 4 illustrates the output of stage prediction.

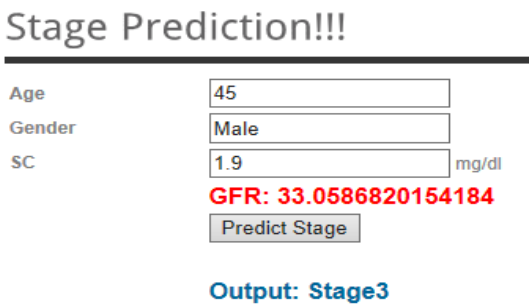


**Fig 4:** Output of stage prediction.

The end result that we get at the end of the experiment using data mining classification technique, we can predict the presence of CKD and its stages. With added datasets the efficiency of the tool becomes even more accurate

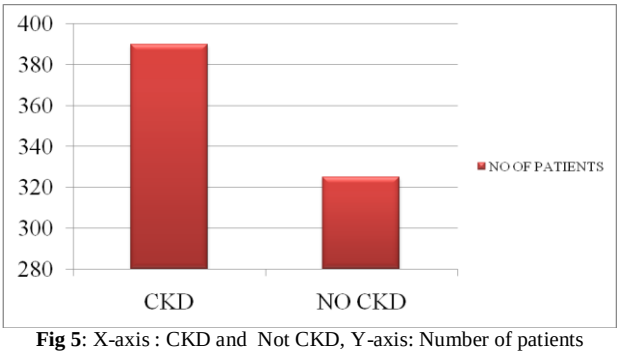


**Fig 5**: X-axis : CKD and Not CKD, Y-axis: Number of patients

In figure 5, the graph shows the number of patients who have CKD and number of patients who does not have CKD. In figure 6, the graph shows Number of patients(x-axis), Number of male and female patients suffering from different stages of CKD(y-axis).

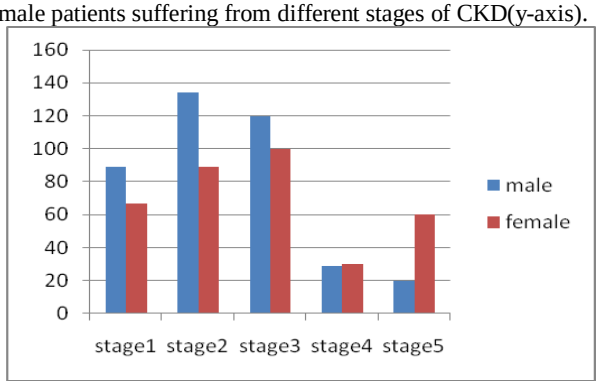


**Fig 6:** Number of patients (x-axis), Number of male and female patients suffering from various CKD stages(y-axis)

## 5. Conclusion

Data mining is especially valuable in therapeutical area. Expansive measure of complicated information is composed from social insurance industry of patients, maladies, healing centers, medicinal supplies, claims, treatment cost and so forth which needs preparing, examining by learning extracting. It concocts an arrangement of equipments and strategies which when connected to the processed data, gives learning to human services experts to settling on suitable choices and improving the execution of patient administration errands. The patients with comparative medical problems can be assembled together and successful treatments could be proposed in view of patient's past records, physical examination, reports, and past treatment designs. There is no automation for chronic kidney disease prediction along with its stages. Existing system is a manual approach, requires medical equipment, more expensive, lack of user satisfaction, less efficient, less accurate. From the examination, it is realized that Naïve Bayes arrangement calculation gives 100% precision when contrasted with ANN. Hence Naïve Bayes is used to predict the presence of CKD

and based on Glomerular Filtration Rate, CKD stages are predicted.

## Acknowledgement

First and foremost, we feel deeply indebted to Her Holiness Most Revered Mata Amritanandamayi Devi(Amma) for her inspiration and guidance both in unseen and unconcealed ways.
Whole heartedly, we thank our college, Amrita School of Arts and Sciences, Mysuru campus, Karnataka, India, for providing the necessary environment, infrastructure, encouragement and for extending the support possible at each stage of project.
We express our sincere gratitude and indebtedness to our parents who have bestowed there great guidance at appropriate times by providing encouragement in planning and carrying out the project.

## References

[1] Sunil D and Prof. B. P. Sowmya, Chronic Kidney Disease Analysis using Data Mining, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, ISSN : 2456-3307, Volume-2, Issue-4.

[2] Mostafa Ghannad-Rezaie and Hamid Soltanian-Zadeh,(2008). Interactive knowledge discovery for temporal lobe epilepsy. INTECH Open Access Publisher.

[3] Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao, (2001). The approach of data mining methods for medical database. In Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE(Vol. 4, pp. 3824-3826). IEEE.

[4] Paolo Bonato, , Delsey M. Sherrill, David G. Standaert, Sara S. Salles, and Metin Akay, (2004, September). Data mining techniques to detect motor fluctuations in Parkinson's disease. In Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE (Vol. 2, pp. 4766 4769). IEEE.

[5] Wang, S., Zhou, M., & Geng, G. (2005). Application of fuzzy cluster analysis for medical image data mining. Mechatronics and Automation, 2, 631-636.

[6] Xing, Wang J, Zhao, Z, & Gao Y, (2007, November), Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In Convergence Information Technology, 2007. International Conference on (pp. 868-872). IEEE.

[7] Sellappan Palaniappan, Rafiah Awang. (2008, March). Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on (pp. 108-115). IEEE.

[8] Heon Gyu Lee, Ki Yong Noh and Keun Ho Ryu. (2008, May). A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness. In BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on (Vol. 1, pp.200-206). IEEE.

[9] K.Srinivas, Dr.G.Raghavendra Rao, Dr. A.Govardhan. (2010, August). Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In Computer Science and Education (ICCSE), 2010 5th International Conference on (pp. 1344- 1349). IEEE.

[10] Sa'diyah Noor Novita Alfisahrin and Teddy Mantoro (2013, December). Data Mining Techniques for Optimization of Liver Disease Classification. In Advanced Computer Science Applications and Technologies (ACSAT), 2013 International Conference on (pp. 379-384). IEEE.

[11] Debabrata Pal ,Chandan Chakraborty and K.M Mandana (2011, November). Data mining approach for coronary artery disease screening. In Image Information Processing (ICIIP), 2011 International Conference on (pp. 1-6). IEEE.

[12] Sonam Nikhar1, A. M. Karandikar2, K.,2016,Prediction of Heart Disease Using Data Mining, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 03 Issue: 02.

[13] Jenn-Long Liu, Yu-Tzu Hsu and Chih-Lung Hung (2012, June). Development of evolutionary data mining algorithms and their applications to cardiac disease diagnosis. InEvolutionary Computation (CEC), 2012 IEEE Congress on (pp. 1-8). IEEE.

[14] Geeta Yadav, Yugal Kumar and G. Sahoo (2012, November). Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers. In Computing and Communication Systems (NCCCS), 2012 National Conference on (pp. 1-8). IEEE.

[15] Ilayaraja, M., and Meyyappan, T. (2013, February). Mining medical data to identify frequent diseases using Apriori algorithm. In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on(pp. 194-199). IEEE.

[16] Sivagowry, S., Durairaj, M., & Persia, A. (2013, February). An empirical study on applying data mining techniques for the analysis and prediction of heart disease. In Information Communication and Embedded Systems (ICICES), 2013 International Conference on (pp. 265-270). IEEE

[17] K. Vasantha Kokilam and Dr. D. Pon Mary Pushpa Latha. (2012, December). A review on evolution of data mining techniques for protein sequence causing genetic disorder diseases. In Computational Intelligence & Computing Research (ICCIC), 2012 IEEE International Conference on (pp. 1-6). IEEE.

[18] Syed Umar Amin, Kavita Agarwal and Dr. Rizwan Beg, (2013, April). Genetic neural network based data mining in prediction of heart disease using risk factors. InInformation & Communication Technologies (ICT), 2013 IEEE Conference on(pp. 1227-1231). IEEE.

[19] Girija, D. K., Shashidhara, M. S., & Giri, M. (2013, October). Data mining approach for prediction of fibroid disease using neural networks. In Emerging Trends in Communication, Control, Signal Processing & Computing Applications (C2SPCA), 2013 International Conference on (pp. 1-5). IEEE.

[20] Juliet Rani Rajan and Dr.C.Chilambu Chelvan (2013, December). A survey on mining techniques for early lung cancer diagnoses. In Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on (pp. 918-922). IEEE.

[21] Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y. L., Castro, A. F., Feldman, H. I., ... & Coresh, J. (2009). A new equation to estimate glomerular filtration rate. Annals of internal medicine, 150(9), 604-612.

[22] Ahalya, C. S., Abin, K. O., & Kanchana, V. (2017, May). Building up an information archive for putting away pesticide data. In Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017 2nd IEEE International Conference on (pp. 2125-2128). IEEE.

[23] Devasia, T., Vinushree, T. P., & Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In Data Mining and Advanced Computing (SAPIENCE), International Conference on (pp. 91-95). IEEE.