Chronic Kidney Disease and Stage Detection Using Machine Learning Classifiers

Sadaf Farheen, Shafiya S[*], Chandini A H, Nagana Devi G J, Akshatha M

Department of CSE, Vidya Vikas Institute of Engineering and Technology, Mysuru, Karnataka

DOI: https://doi.org/10.21467/proceedings.1.60

* Corresponding author email: sshafiya289@gmail.com

Abstract

Data mining has been a current trend for attaining diagnostic results. Huge amount of unmined data is collected by the healthcare industry in order to discover hidden information for effective diagnosis and decision making. Data mining comes up with a set of tools and techniques which when applied to this processed data, provides knowledge to healthcare professionals for making appropriate decisions and enhancing the performance of patient management tasks. Patients with similar health issues can be grouped together and effective treatment plans could be suggested based on patient's history, physical examination, diagnosis and previous treatment patterns. 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. The proposed system predominantly focuses on detecting chronic kidney diseases using naïve bayes and artificial neural network technique C4.5. naïve bayes is a technique used for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from finite set. The stage prediction is done using C4.5 algorithm. Decision trees can be generated using C4.5 algorithm. The decision trees generated by C4.5 are used for classification.

Index Terms- Artificial Neural Network (ANN), C4.5, Chronic Kidney Disease (CKD), Classifiers, Data mining, Naïve Baye's.

1 INTRODUCTION

The processes of extracting useful knowledge from huge data is known as Data Mining. The domains of data mining include image mining, opinion mining, web mining, text mining, graph mining and so on. Some of its applications include anomaly detection, financial data analysis, medical data analysis, social network analysis, market analysis etc. Health organizations use data mining as there isan enormous requirement of analytical methodology inorder to predict and find unknown patterns and information in health data. Which is why it is playing a vital role for discovering new trends in healthcare industry. Data Mining is in particular used in medical field when there is no availability of evidence favoring a particular treatment option is found. Since large amount of complex data is being generated by healthcare industry with regard to patients, diseases, hospitals, medical equipment's, claims, treatment cost to name a few, which requires processing and analysis for knowledge extraction.

Data mining comes up with a set of tools and techniques which when applied to this processed data, provides knowledge to healthcare professionals for making appropriate decisions and enhancing the performance of patient management tasks. Patients with similar health issues can be grouped together and effective treatment plans could be suggested based on patient's history, physical examination, diagnosis and previous treatment patterns.

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 diseases like Chronic Kidney Disease (CKD) using Classification algorithms like Naive Bayes and Artificial Neural Network(ANN) C4.5 predicts stages of Chronic kidney disease(CKD).

2 LITERATURE SURVEY

In these days, health care industries are providing many benefits like fraud detection in health insurance, availability of healthcare facilities to patients at cheap prices, identification of smarter treatment methodologies, and construction of serviceable healthcare policies, effective medical management, better customer relation, advanced patient care and hospital infection control. Disease detection is one of the powerful areas of research in medical. The present lifestyle of people, working environment and diet may give rise to many diseases, one of which includes chronic kidney disease. Chronic Kidney disease (CKD) is prevailing nowadays and has become a global health issue which must be timely detected and diagnosed. Kidneys are important organs of human body that eradicate toxic and unwanted waste from blood causing smooth functioning of body organs. CKD is a condition that describes loss of kidney function over time making it difficult for them to filter poisonous wastes from the body. Researchers in their recent study have addressed the use of data mining techniques for CKD detection.

Veenitha Kunwar et.al, 2016 developed a 'chronic kidney disease analysis usingdata mining classification techniques' shows a potential use of data mining techniques. Chronic Kidney Diseasehas been predicted and recognized using data mining classifiers: ANN and Naive Bayes. Performances of these algorithms are compared using Rapid miner tool. The obtained results showed that Naïve Bayes is the most accurate classifier with 100% accuracy when compared to ANN having 72.73% accuracy.

Anima Singh et.al developed a'Leveraging Hierarchy in Medical Codes for PredictiveModeling' shows Electronic health records (EHRs) contain information that can be used to make clinically useful predictions about the future trajectory of a patient's health. Important parts of this information are recorded using ICD-9 codes. The methods that leverage hierarchy by incorporating the data during feature construction are compared, using a learning algorithm that addresses the structure in the ICD-9 codes when building a model, or doing both.

Chih-Yin Ho et.al, 2012 developed a 'Ultrasonography Image Analysis for Detection and Classification of Chronic Kidney Disease' present the main goal of this study is to provide a consistent and stable indicator to detect and identify Different stages of CKD. Data collection of renal ultrasound images for evaluating performance of the proposed system and the trained CKD reference indicators will be discussed. through observing and comparing the size of kidneys, thickness of renal pelvis and parenchyma, and fibrosis condition gathered to reference indicators, a physician can consistently diagnose the kidney disease identify the stage of CKD under the support of capacities measurements.

Ruey kei chw et.al, 2013 developed a 'intelligent systems on the cloud for the early detection of chronic kidney disease' to develop a profitable intelligent model for detecting CKD for evaluating the severity of a patient with or without CKD. Neural Network models developed for CKD detection may effectively and feasibly supply medical staff with the ability to make precise diagnosis and treatment to the patient. Hlaudi Daniel Masethe et.al, 2014 developed a 'Prediction of Heart Disease using Classification Algorithms' present objective of the research is to predict possible heart attacks from the patient dataset using data mining techniques determines which model gives the highest percentage of correct predictions for the diagnoses. Ju-Hsin Tsai, 2008 developed a 'Data Mining for DNA Viruses with Breast Cancer and its Limitation' explores the possibility that viruses play a role in development of breast tumors. In order to overcome the difficulty, and approaches that are used in ANN (Artificial Neural Network) and AHCTs (Agglomerative hierarchical clustering techniques) to accomplish the research objective. The prediction is done for breast cancer using ANN technique whereas area of concern is CKD.

3 EXISTING SYSTEM

Present system of CKD and its stage detection is manual approach where doctor personally goes through the report and come up with the conclusion. It becomes more expensive due to the need of consulting more doctors. It lacks user satisfaction as the results are not accurate because it may vary on doctor's experience. It is Less Efficient as the whole existing system approach is manual. There is no automation for chronic kidney disease prediction.

4 PROPOSED WORK

Chronic kidney disease (CKD) has become a global health issue and is an area of concern. Itis a condition where kidneys become damaged and cannot filter toxic wastes in the body. Our work predominantly focuses on detecting life threatening diseases like Chronic KidneyDisease (CKD) and its stages using Classification algorithms. Proposed system is anautomation for chronic kidney disease and its stage detection using classification technique "naïve bayes" and artificial neural network technique "C4.5".

4.1 System Design

The system flow is as shown in figure-1 the dataset is taken from UCI repositoryand stored in the database and it is preprocessed to remove noise that is irrelevantdata, then Naïve Byes and C4.5 algorithms are applied to form a model and model isanalyzed to detect CKD and its stages.

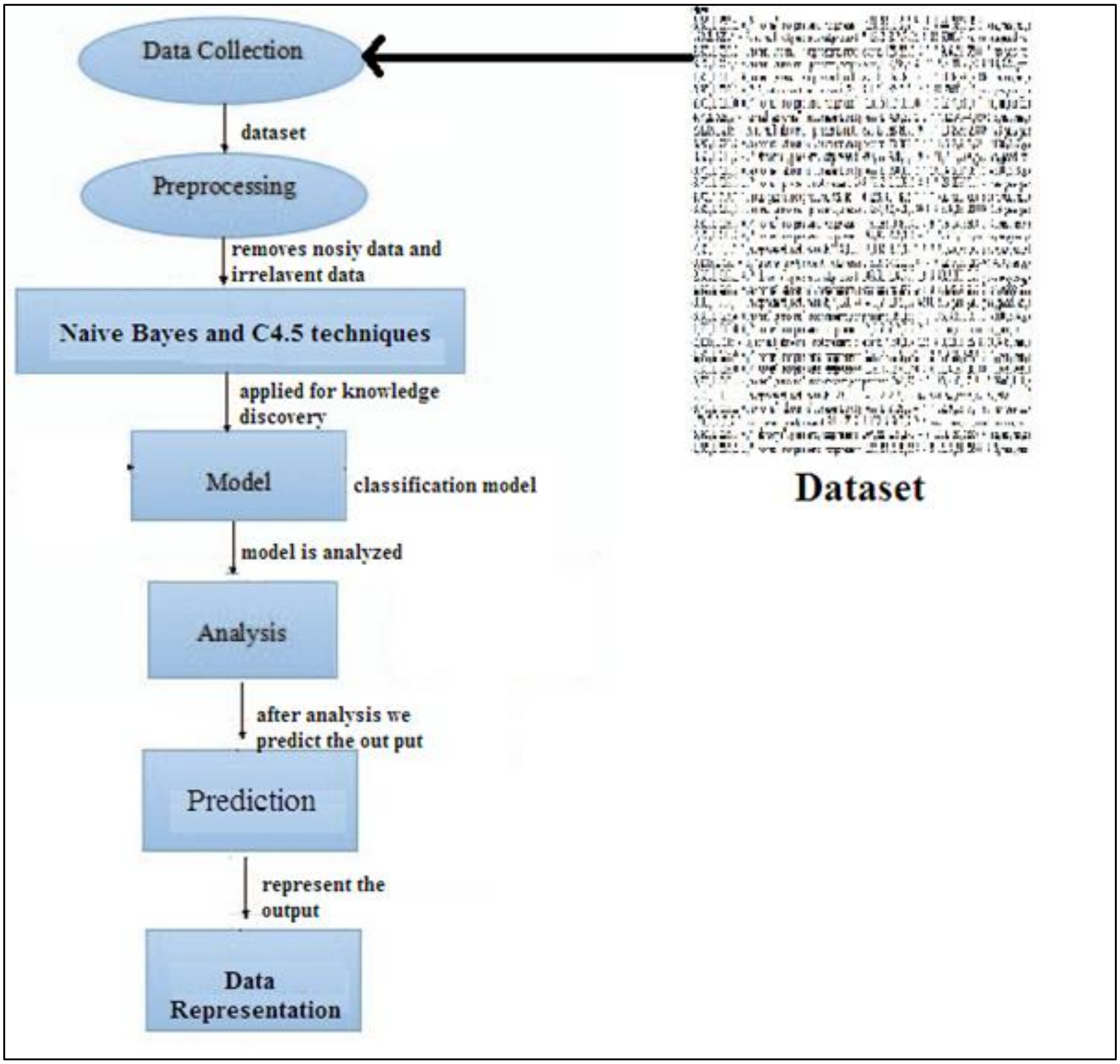

Figure-1: Flow diagram

5 Methodology

5.1 Naïve Bayes Algorithm Steps for the detection of CKD

Step 1: Scan the dataset (storage servers)

Retrieval of required data for mining from the servers such as database, cloud, excel sheet etc.

Step 2: To calculate the attribute value probability. [n, n_c, m, p]

For each attribute the calculation of the probability of occurrence using the following formula. (Mentioned in the next step). For each class (disease) the formulae are applied.

Step 3: use the formula

- P (attribute value (ai)/subject valuevj)=(n_c + mp)/(n+m)

Where:

- n = is the count of training dataset for which v = vj
- n_c = count of training dataset for which v = vj and a = ai
- p = a prior estimate for P(aijvj)
- m = is the sample size

Step 4: Multiply the probabilities by p

For each class multiply the results of each attribute with p and final results are used for classification.

Step 5: By comparing the values, classify the attribute values to one of the predefined class.

The prediction of CKD is shown using Naïve Bayes Algorithm considering following illustration. Table-1

Attributes (Constraints) – S1, S2, S3 [m=3]

Subject (Disease) – CKD, NOT CKD [p=$\frac{1}{2}$= 0.5]

Training Dataset

Table-1: Old patients training dataset for CKD detection

Patient Name	**S1(X,Y,Z)**	**S2 (A,B,C)**	**S3 (P,Q,R)**	**Disease (subject)**
Anil	X	A	P	
Ajay	X	B	Q	
Arun	Y	B	P	
Kumar	Z	A	R	
Naveen	Z	C	R	

New Patient data – Neel Aja Constraints (S1 -X, S2-A, S3-R) Disease – CKD / NOT CKD

P= [n_c + (m*p)]/ (n+m)

Table-2: naïve bayes algorithm applied for new patient

CKD	NOT CKD

X
P=[n_c + (m*p)]/(n+m)
n=2, n_c=2,m=3,p=0.5
p=[2+(3*0.5)]/(2+3)
p=0.7
	X
	P=[n_c + (m*p)]/(n+m)
	n=2, n_c=0,m=3,p=0.5
	p=[0+(3*0.5)]/(2+3)
	p=0.3

A
P=[n_c + (m*p)]/(n+m)
n=2, n_c=2,m=3,p=0.5
p=[2+(3*0.5)]/(2+3)
p=0.7
	A
	P=[n_c + (m*p)]/(n+m)
	n=2, n_c=0,m=3,p=0.5
	p=[0+(3*0.5)]/(2+3)
	p=0.25

R
P=[n_c + (m*p)]/(n+m)
n=2, n_c=1,m=3,p=0.5
p=[1+(3*0.5)]/(2+3)
p=0.5
	R
	P=[n_c + (m*p)]/(n+m)
	n=2, n_c=1,m=3,p=0.5
	p=[1+(3*0.5)]/(2+3)
	p=0.5

CKD – 0.7 * 0.7 * 0.5 * 0.5 (p)=0.1225 NOT CKD – 0.3 * 0.25 * 0.5 * 0.5 (p)= 0.0375

Since CKD > NOT CKD, so this new patient is classified to CKD.

5.2 Disease stage detection using C4.5 Algorithm steps:

Step 1: Scan the dataset (storage servers)

Step 2: For each attribute a, calculate the gain [number of occurrences]

Step 3: Let a_best be the attribute of highest gain [highest count]

Step 4: Create a decision node based on a_best – retrieval of nodes [patient] where theattribute values matches with a_best.

Step 5: Recur on the sub-lists [list of patient] and calculate the count of outcomes [Stages]termed as sub

nodes. Based on the highest count the new node is classified.

Attributes (Features) – F1, F2, F3 [m=3]

Subject (stages) – S1, S2 [p=$\frac{1}{2}$=0.5]

Training Dataset

Table-3: old patients training dataset for CKD stage detection

Name	F1(X,Y,Z)	F2(A,B,C)	F3(P,Q,R)	Stage (subject)
Anil	X	A	P	S1
Kumar	X	B	Q	S1
Ajay	Y	B	P	S2
Naveen	Z	A	R	S1
Akash	Z	A	Q	S2

New Patient Features – Akul F1-X, F2-A, F3-R Which Stage - ?

Feature Count (X) in the dataset = 2
Feature Count (A) in the dataset = 3
Feature Count (R) in the dataset = 1
Sort ();
Reverse ();

Table-4: feature count for more number of occurrences

Feature	**Count**
A	3
X	2
R	1

A – S1 (2) & S2 (1);

This algorithm is based on single attribute values.

Output

Table-5: newpatient stage detection using C4.5 algorithms

Stage	Priority
S1	2

6 CONCLUSIONS

The idea in this paperis a medical sector application which helps the medical practitioners indetecting thedisease (CKD)and its stage. It is an automation for disease detection which identifies the disease stage from the clinical database in an efficient, economical and faster way. The chronic kidney disease and its stage detection is successfully accomplished by applyingthe Naïve Bayes algorithm and C4.5 algorithm respectively. These classificationtechniques come under data mining technology. These algorithms take attributes asinput and predicts the disease based on old patients data.

REFERENCES

[1] M. Abdelaal, et al. “Using data mining for assessing diagnosis of breast cancer”. InComputer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on (pp. 11-17). IEEE.

[2] R. K Chiu, et al. “Intelligent systems on the cloud for the early detection of chronic kidney disease”. InMachine Learning and Cybernetics (ICMLC), 2012 International Conference on(Vol. 5, pp. 1737-1742). IEEE

[3] K. R. Lakshmi,et al.“Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability”. International Journal of Advances in Engineering & Technology (IJAET)(2014), 7(1), 242-254.

[4] L. Xun, et al. “Application of radial basis function neural network to estimate glomerular filtration rate in Chinese patients with chronic kidney disease”. In Computer Application and System Modeling (ICCASM),2010 International Conference on (Vol. 15, pp. V15-332). IEEE.

[5] Ravindra, et al. “Discovery of significant parameters in kidney dialysis data sets by Kmeans algorithm”. InCircuits, Communication, Control and Computing (I4C), 2014 International Conference on (pp. 452-454). IEEE.

[6] S. Ahmed, et al. “Diagnosis of kidney disease using fuzzy expert system” InSoftware, Knowledge, Information Management and Applications (SKIMA), 2014 8th International Conference on (pp. 1-8). IEEE.

[7] VeenitaKunwar et.al, “Chronic kidney disease analysis using data mining classification techniques” 2016 6th International Conference - Cloud System and Big Data Engineering.

[8] Anima Singhet.al, “Leveraging Hierarchy in Medical Codes for Predictive Modeling” Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, Pages 96-103 Newport Beach, California — September 20 - 23, 2014

[9] Chih-Yin Ho,et.al, “Ultrasonography Image Analysis for Detection and Classification of Chronic Kidney Disease”2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems

[10] Ruey Kei Chiu, et.al “Intelligent Systems Developed for the Early Detection of Chronic Kidney Disease”,Advances in Artificial Neural SystemsVolume 2013 (2013), Article ID 539570, 7 pages

[11] Hlaudi Daniel Masethe et.al, “Predictionof Heart Disease using Classification Algorithms” Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA

[12] Ju-Hsin Tsai, “Data Mining for DNA Viruses with Breast Cancer and its Limitation” Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou, ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria