IJCSE International Journal of Computer Sciences & Engineering Open Access

Research Paper Volume-3, Issue-8 E-ISSN: 2347-2693

Performance Analysis of Machine Learning Algorithms for Predicting Chronic Kidney Disease

N.Radha[1] and S.Ramya[2*]

[1] *Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore,*
www.ijcseonline.org

Received: Jul /19/2015 Revised: Jul/30/2015 Accepted: Aug/25/2015 Published: Aug/30/ 2015

Abstract— chronic kidney disease refers to the condition of kidneys caused by conditions, diabetes, glomerulonephritis or high blood pressure. These problems may happen gently for a long period of time, often without any symptoms. It may eventually lead to kidney failure requiring dialysis or a kidney transplant to preserve survival time. So the primary detection and treatment can prevent or delay of these complications. The aim of this work is to reduce the diagnosis time and to improve the diagnosis accuracy through classification algorithms. The proposed work deals with classification of different stages in chronic kidney diseases using machine learning algorithms. The experimental results performed on different algorithms like Naive Bayes, Decision Tree, K-Nearest Neighbour and Support Vector Machine. The experimental result shows that the K-Nearest Neighbour algorithm gives better result than the other classification algorithms and produces 98% accuracy.

Keywords: Chronic Kidney Disease (CKD), Machine Learning (ML), End-Stage Renal Disease (ESRD), Cardiovascular disease, data mining, machine learning,

I. Introduction

Data mining is a used for the healthcare industry to enable health systems systematically. It uses data for analytics to identify incompetence and best practices that increase the care and reduce costs. Medical treatment is facing a challenge of knowledge discovery from the growing volume of data. Nowadays huge data are collected continuously through health examination and medical treatment. Classification rules are typically useful for medical problems that have been applied mainly in the area of medical diagnosis. Moreover, various machine-learning (ML) techniques have been applied to the field of medical treatments over the past few years.

Chronic kidney disease (CKD) is a worldwide common health problem, with predictable lifetime risk of >50%, higher than that for invasive cancer, diabetes and coronary heart diseases. CKD is a long term disorder caused by damage to both kidneys [1], [2]. There is no single cause and the damage is typically permanent and can lead to ill health. In some cases dialysis or transplantation may become essential. Diabetes mellitus is also becoming more common in one cause of CKD. Chronic kidney disease is become more frequently in older people and consequently is likely to increase in the population as a whole. People with chronic kidney diseases are at higher risk of cardiovascular disease and they should be recognized early so that appropriate preemptive measures can be taken [3,4].

CKD is defined as the presence of kidney damage, revealed by the abnormal albumin excretion or decreased kidney function. The disease is quantified by measured or estimated by Glomerular Filtration Rate (GFR) that persists for more than 3 month of the CKD patients. The glomerular filtration rate (GFR) is the best indicator of how well the kidneys are working. The National Kidney Foundation published treatment guidelines for identified five stages of CKD based on diminishing GFR measurements. The guidelines mention different actions based on the stage of kidney disease [5].

A GFR of 90 or above is considered as normal. Even with a normal GFR, it may be at increased risk for developing CKD if the patients have diabetes, blood pressure in high, or a family history of kidney disease. The risk increases with age over 65 are more than twice as likely to develop CKD as people between the ages of 45 and 65.

The remaining paper is organized as follows: Section II deals with literature survey about chronic kidney diseases. In section III methodologies used for classifying chronic kidney diseases are discussed. Section IV deals with experimental and its results. Section V gives prediction of chronic kidney diseases with various performances and its future works.

II. LITERATURE SURVEY

Miguel A. et al. [6] proposed an approach for the management of alarms related to monitoring CKD patients within the eNefro project. The results proof the pragmatism of Data Distribution Services (DDS) for the activation of emergency protocols in terms of alarm ranking and personalization, as well as some observations about security and privacy.

Christopher et al. [7] discussed a contextualized method and possibly more interpretable means of communicating risk information on complex patient populations to time-constrained clinicians. Dataset was collected from American Diabetes Association (ADA) of 22 demographic and clinical variables related to heart attack risk for 588 people with type2 diabetes. The method and tool could be encompasses to other risk-assessment scenarios in healthcare distribution, such as measuring risks to patient safety and clinical recommendation compliance.

Srinivasa R. Raghavan et al. [8] explored reviews the literature on clinical decision support system, debates some of the difficulties faced by practitioners in managing chronic kidney failure patients, and sets out the decision provision techniques used in developing a dialysis decision support system.

Ricardo T. Ribeiro et al. [9] proposed a method, called clinical based classifier (CBC), discriminates healthy from pathologic conditions. A large multimodal feature database was specifically built for this study. It containing chronic hepatitis, 34 compensated cirrhosis, and 36 decompensated cirrhosis cases, all validated after histopathology examination by liver biopsy. The CBC classification outperformed the nonhierarchical one counter to all scheme, achieving better accuracy.

Mitri F.G. et al. [10] presented an ultrasound-based modality complex to stiffness and free from speckle noise and owns some advantages over the conventional ultrasound imaging in terms of the quality.

Chih-Yin Ho et al. [11] presented a computer-aided diagnosis tool based on analyzing ultrasonography images and the system could detect and classify various stages of CKD. The dataset was collected thousands of ultrasonic images from patients with kidney diseases, and the selected typical CKD images were applied to be pre-analyzed and trained for assessment. The calculated changeover locations are reference indicators could be responsible for physicians an auxiliary and objective computer-aid diagnosis tool for CKD identification and classification.

Al-Hyari et al. [12] proposed a new clinical decision support system for identifying patients with CRF. Some data classification algorithms including Artificial Neural Networks, Decision Tree and Naive Bayes are developed and applied to diagnose patients with CRF and determine the evolution stage of the disease. The dataset containing 102 instances is collected from patients' records and used for this study. The attained results showed that the developed decision tree algorithm is the most accurate CRF classifier (92.2%) when compared to all other algorithms used in this study.

Kuo-Su Chen et al. [13] established a detection system based on computer vision and machine learning techniques for simplifying diagnosis of CKD and different stages of CKD. The proposed system required average time of 0.016 seconds for feature extraction and classification of each testing case. The results presented that the system could produce reliable diagnosis based on noninvasive ultrasonography methods and which could be measured as the most proper clinical diagnosis and medical treatment for CKD patients.

Anne Rogers et al. [14] discussed to discover patients' experiences disclosure of CKD in primary care settings. The dataset contains purposive sample of 26 patients, with a mean age of 72 years were cross-examined using constant relative techniques. This study challenges the assumptions characteristic in extensive health policy objectives that are increasingly built on the notion of responsible patients and the ethos of the active support of self-management for pre-conditions.

Mohammed Shamim Rahman et al. [15] described the effect of chronic kidney disease (CKD) on morbidity and mortality following Trans catheter aortic valve implantation (TAVI) including patients on hemodialysis, often excluded from randomized trials. There are 118 consecutive patients underwent TAVI 63 were considered as having (CKD) and 55 not having (No-CKD) significant pre-existing CKD.The result shows TAVI is a safe, suitable treatment for patients with pre-existing CKD, though carefulness must be trained, particularly in patients with pre-existing diabetes mellitus and elevated pre-operative serum creatinine levels as this confers a greater risk of AKI development, which is associated with increased short term post-operative mortality.

III. METHODOLOGY

A. *Naïve Bayes*

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions among predictors. A Naive Bayesian model is easy to build, with no complex iterative parameter assessment which makes it especially useful for very large datasets. Even though it's simple, the Naive Bayesian classifier often does unexpectedly well and is widely used because it often outperforms more refined classification methods. Bayes theorem delivers a way of calculating the posterior probability, P(c|x), from P(c),P(x) and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (*x*) on a given class (*c*) is independent of the values of other predictors. This hypothesis is called class conditional independence [16].

$$\mathrm{p(c|x)} = \frac{p(x|c)p(c)}{p(x)} \quad \text{---- (1)}$$

P (c|x) is the posterior probability of given predictor.
P(c) is the prior probability of class.
P (x|c) is the likelihood which is the ability of predictor given
class.
P(x) is the prior probability of predictor.

B. *Decision Tree*

Decision tree builds classification or regression models in the form of a tree like structure. It breakdowns a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally established. The last result is a tree with decision nodes and its leaf nodes. Decision nodes have two or more branches. Leaf node represents a classification or decision. The uppermost decision node in a tree which resembles to the finest predictor called root node. Decision trees can switch both categorical and numerical data values. The core algorithm for building decision trees is called ID3 by J. R. Quinlan which employs a top-down and greedy search over the space of possible branches with no backtracking.

Algorithm

- Start with single node N, with training data D.
- If all the data in D belongs to same class, then N becomes leaf. Otherwise attribute 'A' is selection method based on splitting criterion.
- The instance in 'D' is partitioned accordingly.
- Apply algorithm recursively to each subset in 'D' to each subset in 'D' to form decision tree.

The algorithm uses Entropy and Information Gain to construct a decision tree [17].

Entropy

ID3 algorithm uses entropy to calculate the similarity of a sample. If the sample is completely similar the entropy is zero and if the sample is an equally divided it has entropy of one.

$$E(S) = \sum_{i=0}^{c} -p_i \log 2 p_i \qquad \text{----- (2)}$$

Information Gain

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Creating a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

$$Gain(T,X) = Entropy(T) - Entropy(T,X) \qquad \text{--- (3)}$$

C. K- Nearest Neighbour

K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique. A case is classified by a majority vote of its neighbours, in the case being assigned to the class most common between its K nearest neighbours measured by a distance function. If K = 1, then the case is simply allocated to the class of its nearest neighbour [18].

Algorithm

- Training set: $(x_1,y_1),(x_2,y_2),,,,,(x_n,y_n)$.
- Assume X:= $(x:^{(1)},x:^{(2)},\ldots,x:^{(d)})$ is a dimensional feature vector of real numbers for all i.
- Y: is a class label {1…C}, for all i.
- Find the closest point X_j to X_{new} using distance measures.
- Classify by Y^{knn}= majority vote among the K points.

Distance Measures

There are three different measures are used for calculate the distances.

$$Euclidean - \sqrt{\sum_{i-1}^{k} (x_i - y_i)^2} \qquad \text{------- (4)}$$

$$Manhattan = \sum_{i=1}^{k} |x_i - y_i| \qquad \text{------ (5)}$$

$$Minkowski = \left(\sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{1/q} \qquad \text{-------- (6)}$$

D. Support Vector Machine

A Support Vector Machine (SVM) performs classification by finding the hyper plane that maximizes the margin between the classes. The vectors (cases) that define the hyper plane are the support vectors [19].

Algorithm

- Define an optimal hyper plane: maximize margin.
- Extend the above definition for non-linearly separable
  Problems: have a penalty term for misclassifications.
- Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space.
  For this type of SVM, training involves the minimization of the error function.

$$\frac{1}{2} w^T w + C \sum_{i=1}^{k} \xi_i \qquad \text{------- (7)}$$

IV. EXPERIMENTAL RESULT

A. Dataset

The dataset for diagnosis of chronic kidney disease is obtained from medical reports of the patients collected from different laboratories in Coimbatore. There are 1000 instances with 14 different attributes related to kidney disease like PID (patients ID), Age, Gender, Weight, Serium-albumin, Serium- sodium, Blood urea nitrogen, Serium creatinine, Serium uric acid, Sodium urine, Urine urea nitrogen, Urine creatinine, Urine uric acid and Kidney failure. The records are classified as Low, Mild, Moderate, Normal and Severe.

B. Classification using R

R tool is by academicians who persistently provide libraries for new and developing statistical techniques. It is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. The machine learning algorithms such as Decision tree, Naïve bayes, Support vector machine and K-nearest neighbour are implemented as classification task for accurate diagnosis of CKD based on different performance evaluation measures.

C. Performance evaluation

Diagnosis tests include different types of information, such as symptoms and medical tests. Doctor’s conclusion of medical treatment rest on diagnosis tests which makes the accuracy of diagnosis is important in medical care. Providentially, the attributes of the diagnosis tests can be measured for a given disease condition .The best probable test can be chosen based on these attributes. Sensitivity, specificity, kappa and accuracy are widely used statistics to describe a diagnostic test.

Kappa: Cohen's kappa measures the pact between two different measures which classify N items into C mutually exclusive categories.

$$k = \frac{\Pr(a) - Pra(e)}{1 - \Pr(e)} \quad \text{----- (8)}$$

Where; Pr(a) is observed agreement, and Pr(e) is the hypothetical probability of chance agreement.

Sensitivity: also called the true positive rate, measures the proportion of positives which are correctly identified. The probability of positive test give that the patient is ill.

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad \text{---- (9)}$$

Specificity: also called the true negative, measures the proportion of negatives which are correctly identified. The probability of negative test given that the patient is well.

$$Sepcificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad \text{---- (10)}$$

Accuracy: is the amount of true results, one or the other true positive or true negative, in a population. It measures the degree of precision of a diagnostic test on a condition.

$$Accuacy = \frac{No.of\ Correct\ assesments}{No.of\ all\ assesments} \quad \text{---- (11)}$$

D. Result

Kappa Statistic is used to evaluate the accuracy of any particular measuring cases which is used to distinguish between the reliability of the data collected and their validity. The Kappa score for the K-Nearest Neighbour is 0.9822. The K-Nearest Neighbour provides better result.

The Sensitivity and Specificity measures are used to calculate the true positive rate and true negative rate. A sensitivity and Specificity value for each classifier is given in Table.1.

The value of Sensitivity for Decision tree it is 0.9482, Naïve Bayes obtained 0.7244, Support Vector Machine attained 0.7673 and K-Nearest Neighbour gained 0.9856. And the value of Specificity for Decision tree value is 0.8067, Naïve Bayes obtained 0.4491, Support Vector Machine attained 0.6421 and K-Nearest Neighbour gained 0.976.

Table.1 Performance Measures for various Classifiers

Clas/PM	DT	NB	SVM	KNN
Kappa	0.7328	0.5160	0.7944	0.9822
Specificity	0.8067	0.4491	0.6421	0.9764
Sensitivity	0.9482	0.7244	0.7673	0.9856
Accuracy	78.6%	61.8%	83.9%	98%

*clas-Classifier, PM-Performance Measure, DT-Decision Tree, NB-Naive Bayes, SVM-Support Vector Machine, KNN-K Nearest Neighbour.

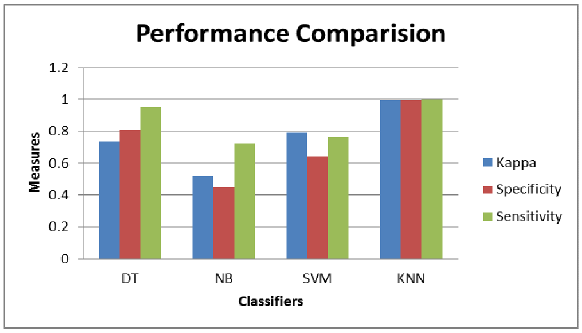

Fig.1 Performance Comparision Chart

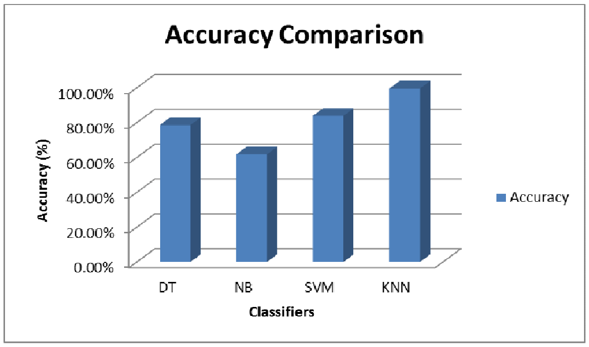

Fig.2 Accuracy Comparison Chart

V. CONCLUSION

Accurate prediction of chronic kidney disease is one of the emerging topics in medical diagnosis. Even though some approaches using real-time features shows very good performance in terms of accuracy. This work proposes a classification model to predict the chronic kidney disease using various machine learning algorithms.

All the four classification algorithms have been considered for diagnosis of chronic kidney disease. From the above results, the objective is to find the better model for chronic kidney disease. The K-Nearest Neighbour is the better model for diagnosis of chronic kidney disease it attains the accuracy of 98%. It correctly classified the 980 instances from 1000 instances. Thus finally it is observed that KNN is better algorithm for chronic kidney diagnosis.

REFERENCES

[1] John R, Webb M, Young A and Stevens PE, “Unreferred chronic kidney disease: a longitudinal study”, American Journal of Kidney Disease, Vol.5, Issue- 3, **2004**, pp.**825-35**.

[2] Coresh J, Astor BC, Greene T, Eknoyan G and Levey AS, “Prevalence of chronic kidney disease and decreased kidney function in the adult US population: Third National Health and Nutrition Examination Survey”, American Journal Kidney Disease, Vol.1, Issue- 4, **2003**, pp.**1-12**.

[3] De Lusignan S, Chan T, Stevens P, O’Donoghue D, Hague N and Dzregah B, et al. “Identifying patients with chronic kidney disease from general practice computer records” ,Oxford Journals of Family Practice,Vol.3, Issue- 22, **2005,** pp.**234-241**.

[4] Hallan SI, Coresh J, Astor BC, Asberg A, Powe NR and Romundstad S, et al. “International comparison of the relationship of chronic kidney disease prevalence and ESRD risk”, Journal American Society of Nephrology,Vol.17, Issue-8, **2006**, pp.**2275-2284**.

[5] Levin A, Coresh J, Rossert J, et al.“Definition and classification of chronic kidney disease: a position statement from kidney disease”, The New England Journal of Medicine, **2002**, pp.**36-42**.

[6] Miguel A. Estudillo-Valderrama, Alejandro Talaminos-Barroso and Laura M. Roa,“A Distributed Approach to Alarm Management in Chronic Kidney Disease”, IEEE journal of biomedical and health informatics,Vol.18, Issue-6, **2014**, pp. **1796-1803**.

[7] Christopher A. Harle, Daniel B. Neill and Rema Padman, “Information Visualization for Chronic Disease Risk Assessment”, IEEE Computer Society, **2012**, pp.**81-85**.

[8] Srinivasa R. Raghavan, Vladimir Ladik, and Klemens B. Meyer,“Developing Decision Support for Dialysis Treatment of Chronic Kidney Failure”, IEEE transactions on information technology in biomedicine, Vol. 9, Issue-2, **2005**, pp. **229-238**.

[9] Ricardo T. Ribeiro, Rui Tato Marinho, and J. Miguel Sanches, “Classification and Staging of Chronic Liver Disease from Multimodal Data”, IEEE transactions on biomedical engineering, Vol. 60, Issue- 5, **2013,** pp.**1336-134.**

[10] Mitri F.G. et al, “Vibro-acoustography imaging of kidney stones in vitro Vibro-acoustography”, IEEE Transactions on Biomedical Engineering **2011**.

[11] Chih-Yin Ho, Tun-Wen Pai, Yuan-Chi Peng and Chien-Hung Lee, “Ultrasonography Image Analysis for Detection and Classification of Chronic Kidney Disease”, IEEE conference published on Intelligent and Software Intensive Systems (CISIS),**2012**, pp.**624 – 629**.

[12] Al-Hyari and Al-Taee, “Clinical decision support system for diagnosis and management of Chronic Renal Failure”, IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), **2013**, pp.**1-6**.

[13] Kuo-Su Chen, Yung-Chih Chen and Yang-Ting Chen, “Stage diagnosis for Chronic Kidney Disease based on ultrasonography”, IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), **2014**, pp. **525 – 530**.

[14] Anne Rogers, Anne Kennedy, Thomas Blakeman and Christian Blickem, “Non-disclosure of chronic kidney disease in primary care and the limits of instrumental rationality in chronic illness self-management”, ELSEVIER Social Science & Medicine 131, **2015**, pp.**31-39**.

[15] Mohammed Shamim Rahman, Rajan Sharma and Stephen J.D. Brecker,“Transcatheter aortic valve implantation in patients with pre-existing chronic kidney disease”, ELSEVIER International Journal of Cardiology Heart & Vasculature , Vol.5, **2015**, pp. **9–18**.

[16] Eibe Frank, Ian H. Witten,” Data Mining – Practical Machine Learning Tools and Techniques”, Elsevier, **2005**.

[17] Han, J., Kamber, M. Kamber. “Data mining: concepts and techniques”. Morgan Kaufmann Publishers, **2000**.

[18] N. Bhatia et al, “Survey of Nearest Neighbour Techniques”, International Journal of Computer Science and Information Security, Vol. 8, , Issue- 2, **2010**.

[19] John Shawe-Taylor, Nello Cristianini, “Support Vector Machines and other kernel- based learning methods”, Cambridge University Press, UKS, 2000.