International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Chronic Kidney Disease Prediction Using Machine learning

Pooja Mane, Naresh Thoutam, Neha Tiwari, Gauri Mandlik and Nutan Pandey

UG Student, Sandip institute of Technology and Research Center, India

.

A B S T R A C T

The Chronic Kidney disease is the most important health issue in Nowadays. Chronic Kidney diseases cause morbidity and increase of death rates in India and other low- and middle-income countries. The chronic diseases account to about 60 of all deaths worldwide. 80 of chronic disease deaths worldwide also occur in low- and middle-income countries. In India, probably the amount of deaths thanks to chronic disease found to be 5.21 million in 2008 and seems to be raised to 7.63 million in 2020 approximately 66.7. There are approximately 1 million cases of Chronic Kidney Disease (CKD) per year in India and in 2020 it gives a count of 15-17 per cent. Chronic kidney disease is also called renal failure. It is a Dangerous disease of the kidney which produces gradual loss in kidney functionality. CKD may be a slow and periodical loss of kidney function over a period of several years. A person will develop permanent kidney failure. If CKD is not detected and cured in early stage then patient can show following Symptoms: Blood Pressure, anaemia, weak bones, poor nutrition health and nerve damage, decreased immune response because at advanced stages dangerous levels of fluids, electrolytes, and wastes can build up in your blood and body. Hence it's essential to detect CKD at its early stage but it's unpredictable as its Symptoms develop slowly and are not specific to the disease.

Keywords: Artificial Intelligence, CKD, electrolytes, anaemia.

1 INTRODUCTION

Kidney disease is considered a major problem for people of age 60 and above. The major cause is the degeneration of the kidney is that the minimize the rate of glomerular _filtration. This problem, when lasting more than three Months, is generally considered as chronic kidney disease (CKD) [1]. CKD is ranked as the 10th major cause of death in the world. Hypertension, diabetes, and aging are considered leading causes of CKD, in addition to other factors such as high blood pressure, coronary artery Disease and anemia. If the problem can be detected in early stages, then it is considered feasible to save kidney function for the longer survival of the patient. Early diagnosis of CKD can facilitate its treatment and help avoid costly treatment procedures such as dialysis and transplants. With machine learning techniques, it is possible to analyze lab records and other information on patients for the early detection of CKD [2]. Low-level data can be transformed into high-level knowledge through the knowledge discovery in databases (KDD) [2]. This transformation can help practitioners better understand CKD patterns for its early diagnosis. This study analyzes CKD using machine learning Techniques using a CKD dataset from the UCI machine learning data Warehouse. CKD is detected using the Apriority association technique for 400 instances of chronic kidney patients with 10- fold-cross-validation Testing and the results are compared across a number of classificationsAlgorithms including Zero, One, naive Bayes, J48, and IBk (K-nearest neighbor). The dataset is preprocessed by completing and Normalizing missing data. The most relevant features are selected from the dataset to improve accuracy and reduce training time for machine learning techniques. A set of experiments is conducted using various WEKA-implemented machine learning techniques to detect CMD Based on the CKD dataset from the UCI machine [2]. The results are compared for detection accuracy across different machine learning Techniques.

2. Literature Review:

There are many researchers who work on prediction of CKD with the assistance of the many different classification algorithms. And people researchers get expected output of their model. Gunarathne W.H.S.D et.al. [1] Has compared results of various models. And eventually they concluded that the Multiclass Decision forest algorithm gives more accuracy than other algorithms which is around 99% for the reduced dataset of 14 attributes. S.Ramya and Dr.N.Radha [2] worked on diagnosis time and improvement of diagnosis accuracy using different classification algorithms of machine learning. The proposed work deals with classification of various stages of CKD consistent with its gravity. By examine different algorithms like Basic Propagation Neural Network, RBF and RF. The analysis results indicate that RBF algorithm gives better results than the opposite classifiers and produces 85.3% accuracy. S.Dilli Arasu and Dr. R. Thirumalaiselvi [3] have worked on missing values during a dataset of chronic renal disorder. Missing values in dataset will reduce the accuracy of our model also as prediction results. They find solution over this problem that they performed a recalculation process on CKD stages and by doing in order that they got up with unknown values. They replaced missing values with recalculated values. As if salekin and john stankovic [7] they use novel approach to detect CKD by using machine learning algorithm. They get result on dataset which having 400 records and 25 attributes which provides results of patient having CKD or not CKD. They use k-nearest neighbors, random forest and neural network to urge results. For feature reduction they use wrapper method which recognizes CKD with high accuracy. Pinar Yildirim [8] searches the effect of sophistication imbalance once we train the info by using development of neural network algorithm for creating medical decision on chronic renal disorder. During this proposed work, a comparative study was performed using sampling algorithm. This study reveals that the performance of classification algorithms is often improved by using the sampling algorithms. It also reveals that the training rate may be a crucial parameter which significantly effect on multilayer perceptron. Sahil Sharma, Vinod Sharma, and Atul Sharma [9], has assessed 12 different classification algorithm on dataset which having 400 records and 24 attributes. That they had compared their calculated results with actual results for calculating the accuracy of prediction results. They used assessment metrics like precision, sensitivity, accuracy and specificity. They find that the choice tree technique gives correctness up to 98.6%, sensitivity of 0.9720, and precision of 1 and specificity of 1. Dataset and Attributes: during this paper CKD dataset [4] is downloaded from UCI archive. The given dataset includes 400 patients' records with 25 attributes. All this 25 attributes are main attributes which are associated with CKD disease. Out of 25 attributes we only use 14 attributes to create our predictive model.

3. The Proposed Work:

The main objective of this study is to investigate machine learning techniques in combination with feature selection techniques for effective CKD detection in terms of detection accuracy. The study proposes various prediction models using classification algorithms with different techniques offered by the WEKA tool and compares them for correctly classified instances. The identified classification technique can provide predicted values for early CKD diagnosis.

3.1. The Proposed Framework:

The proposed framework for developing prediction machine learning models and their comparison are depicted in Fig. 1. The main objective of the present research is to propose a machine learning technique to predict CKD using associative and classification algorithms. The proposed technique generates classification association rules (CARs) to determine techniques with a high percentage of correctly classified instances, and identified classifiers can facilitate early CKD diagnosis. A comparative analysis of the proposed technique is performed using other state-of-the-art techniques. Fig. 1 briefly details various stages:

i. Dataset selection stage: The dataset is selected to predict CKD for data analysis and effective knowledge. Enough data are required to implement a machine learning technique for a selected dataset. In this set of experiments, CKD data are obtained from the UCI machine learning repository.

ii. Preprocessing and transformation stage: The dataset is prepared in attribute-relation file format with 16 attributes. The given dataset is converted into a binomial format to implement associative techniques. In addition, missing records, duplicate records, and unnecessary fields are removed for a standard data format.

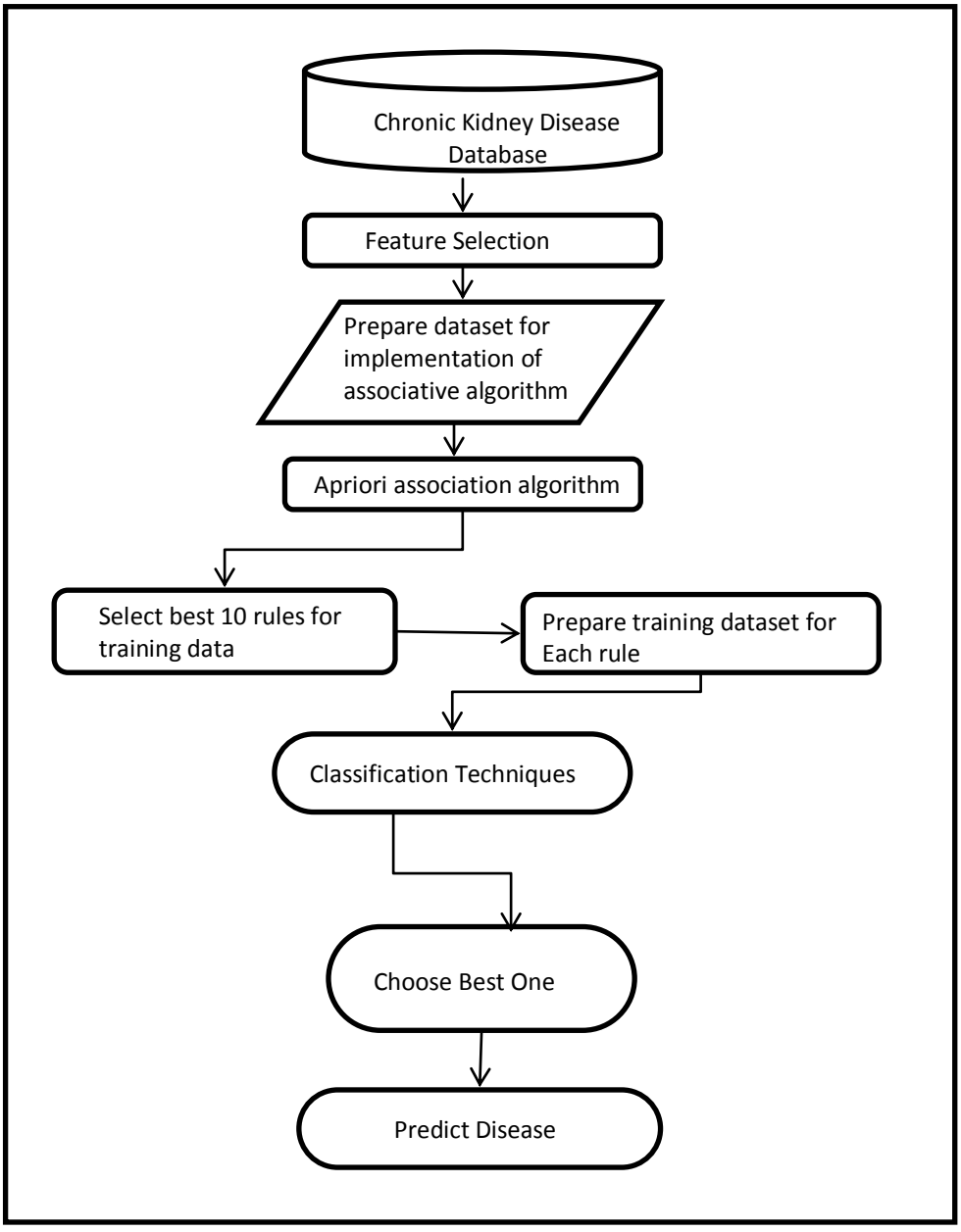
iii. Feature selection stage: The most promising characteristic of the CKD dataset are selected using the WEKA tool for better results. Feature evaluators and search methods are used for this purpose. The correlation-based feature selection subset evaluator is used as the feature evaluator, and the greedy stepwise search method is used. The selected characteristic include blood pressure, red blood cells, pus cell, serum keratinize, hemoglobin, hypertension, diabetes mellitus, appetite, and pedal edema for better results from the CKD dataset.

iv. Selection of associative rules: The Apriority association algorithm is implemented, and 10 best rules are selected to prepare the training dataset to implement different classification algorithms.

v. Implementation of classification algorithms: The five classifiers are trained using the dataset selected based on association rules including k-nearest neighbor, naive-Bayes, ZeroR, OneR, and J48.

vi. Performance evaluation stage: k-nearest neighbor, naïve Bayes, ZeroR, OneR, and J48 are trained and tested using the identified CKD dataset, and the performance of each classifier is measured for correctly classified instances of the identified dataset.

vii. The Disease prediction system: The recognized best classifier helps to form an intelligent CKD prediction system (ICKDPS) for the accurate prediction of other chronic diseases such as heart disease.



4.2. Benchmark Dataset

A dataset with a total of 400 instances with 16 selected attributes is used. The dataset is obtained from the Machine Learning Repository [14]. The attribute "class" is a measurable field with the value "ckd" and indicates an individual with CKD, and "nonckd" indicates an individual with no CKD. Table 1 shows the attributes, descriptions, and values for the CKD dataset. The dataset has 250 "ckd" and 150 "nonckd" instances.

Table 1: Description of attributes

Description of attributes Attributes	Description
Age	Range [2 -90] In the year
Blood pressure	Range [50 - 180] In mm Hg
Red Blood Cell	having two nominal value “normal” or “abnormal”
Pus Cell	having two nominal value “normal” or “abnormal”
Bacteria	having two nominal value Bacteria is “present” and “not present”
Serum Creatinine	Numerical value in mgs/dl
Hemoglobin	The numerical value in gms
Hypertension	having two nominal value “yes” and “no”
Diabetes Mellitus	having two nominal value “yes” and “no”
Coronary Artery Disease	having two nominal value “yes” and “no”
Appetite	having two nominal value Appetite is “good” and “poor”
Pedal Edema	having two nominal value Pedal Edema is “yes” and “no”
Anemia	having two nominal value Pedal Edema is “yes” and “no
Class	having the class value “ckd” represent Chronic Kidney Disease and “nonckd” represent Chronic Kidney Disease not present

5. Results & Discussion

A set of experiments is conducted using the identified benchmark dataset with different classification techniques implemented in WEKA. The results are compared for correctly classified instances. The evaluation of results is based on the following criteria:

1. Incorrectly classified instances, correctly classified instances, kappa statistic, and mean absolute error rate for different classifiers with and without the Apriori association algorithm using 10-fold-cross-validation testing are compared. The results are shown in Tables 2 and 3.
2. The results are compared for the accuracy of the CKD dataset from the UCI Machine Learning database, as shown in Table 4.
3. Four patient sample datasets are tested to predict CKD using the best classification technique, as shown in Table 5.

The results of the proposed framework are calculated using the WEKA tool. Table 2 compares different classifiers without the Apriori association algorithm with the 10-fold cross-validation testing option. Table 3 compares different classifiers for the Apriori association algorithm on the CKD dataset. The error rate plays no role in classification and is used for numeric prediction. The J48 algorithm uses the decision tree for classification, and Fig. 2 shows a sample decision tree created using J48.

Table 2: Comparison of Results for Classifiers on Chronic Kidney Disease Dataset

Classifiers	Correctly Classified Instance (%)	In-Correctly Classified Instance (%)	Kappa Statistic	Mean absolute error
ZeroR	62.5	37.5	0	0.4689
OneR	87.5	12.5	0.7468	0.125
J48	96	4	0.9153	0.0649
IBK	94.5	5.5	0.886	0.0491
NaiveBayes	96.5	3.5	0.9267	0.0397

Table 3: Comparison of Results for Classifiers Using Apriori on Chronic Kidney Disease Dataset

Classifiers	Correctly Classified Instance (%)	In-Correctly Classified Instance (%)	Kappa Statistic	Mean absolute error
ZeroR	56	44	0	0.4929
OneR	92	8	0.8316	0.08
J48	98.33	1.67	0.966	0.0186
IBK	99	1	0.9798	0.0109
NaiveBayes	98.33	1.67	0.9663	0.014

Table 4: Comparison of Results from Previous Studies

Author	Tool Used	Techniques	Accuracy (%)
Ramya and Radha [7]	WEKA	K-Mean	86
Sinha & Sinha [15]	WEKA & Orange	SVM KNN	73 78
Khan and Westin [16]	WEKA	Naive Bayes J48 KNN	90.4 82.5 84.1
Abeer & Ahmad [17]	WEKA	SVM	93.14
Jena & Kamila [18]	WEKA	Naive Base SVM J48	95 62 99
Vijayarani and Dhayanand [19]	MATLAB	Naive Bayes SVM	70.96 76.32
Kumar [20]	WEKA	SMO Naive Bayes RBF MLPC	97 95 98 98
The proposed technique	WEKA	IBk with Aprior Algorithm	99

As Table 4 compares the results from previous studies and finds the proposed method to show 99% accuracy using the IBk classifier with the Apriori association algorithm. Table 5 shows the four patient samples to test the proposed approach for the CKD risk level.

Table 5: Sample Data for Predicting Chronic Kidney Disease Risk Level

Attributes	Sample 1	Sample 2	Sample 3	Sample 4
Age	57	38	46	31
Blood pressure	152 {High}	75 {Low}	137 {High}	110 {Normal}
Red blood Cell	Abnormal	Normal	abnormal	Normal
Pus Cell	Abnormal	abnormal	Normal	Normal
Bacteria	Present	not present	present	not present
Serum Creatinine	3.8 {High}	0.5 {Low}	1.7 {High}	0.9 {Normal}
Hemoglobin	8 {low}	13.5	9.5	12.5
		(Normal)	(Low)	(Normal)

6. Conclusion and Future Scope

This study investigates various machine learning techniques, particularly classification and association techniques, to predict CKD. The study analyzes the effects of using feature selection techniques in combination with classification techniques. Classification techniques implemented in WEKA are used to benchmark the CKD dataset. The results are computed using 10-fold cross-validation with and without the feature selection technique. The results are compared for correctly classified instances, kappa statistic, and mean absolute value with and without the feature selection technique. The benchmark dataset is prepared using the Apriori association algorithm. The extracted data are further used to validate ZeroR, OneR, J48, IBk, and naive Bayes implemented in WEKA. The results note that the best result can be achieved using IBk with the Apriori associative algorithm for 99% accuracy. Future research should analyze different supervised and unsupervised machine learning techniques and feature selection techniques with additional performance metrics for better CKD prediction.

References

[1] Patil PM, "Review on Prediction of Chronic renal disorder using data processing Techniques", International Journal of Computer Science and Mobile Computing, Vol.5, No.5, (2016), pp.135-141.

[2] Dulhare UN & Ayesha M, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), (2016), pp.1-5.

[3] Gopika M, "Machine learning Approach of Chronic renal disorder Prediction using Clustering Technique", International Journal of Innovative Research in Science, Engineering and Technology, Vol.6, No.7, (2017), pp.14488 14496.

[4] Dr. S. Vijayarani1 and S.Dhayanand, "Kidney disease prediction using SVM and ANN algorithms", International Journal of Computing and Business Research (IJCBR) ISSN (Online), vol. 6, no. 2, (2015), pp. 2229-6166.

[5] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic renal disorder (ckd)," in 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2017, pp. 291–296.

[6] Kayaalp F, Basarslan MS & Polat K, "A hybrid classification example in describing chronic kidney disease", IEEE Electric Electronics, computing, Biomedical Engineering's' Meeting (EBBT), (2018), pp.1-4.

[7] Guneet Kaur, "Predict Chronic renal disorder using data processing in Hadoop, International Conference on Inventive Computing and Informatics, 2017.

[8] Wibawa MS, Maysanjaya IMD & Putra IMAW, "Boosted classifier and features selection for enhancing chronic renal disorder diagnose", IEEE 5th International Conference on Cyber and IT Service Management (CITSM), (2017), pp.1-6

[9] R. Devika, S. V. Avilala and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic kidney disease prediction using Naive Bayes, KNN and Random Forest," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 679-684.
[10]J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools within the prediction of chronic renal disorder progression," Journal of translational medicine, vol. 17, no. 1, p. 119, 2019.
[11]A. Q. Ansari and N. K. Gupta, "Automated Diagnosis of Coronary heart condition Using Neuro-Fuzzy Integrated System".
[12] the connection between machine learning and data mining, (2018).
[13] N. Sharma and Er. Rohit Kumar Verma, "Prediction of kidney disease by using processing Techniques", International Journal of Advanced Research in computing and Software Engineering, Volume 6, Issue 9, September 2016, ISSN: 2277 128X.