Chronic Kidney Disease Prediction: A Review

[1]Sujata Drall, [2]Gurdeep Singh Drall, [3]Sugandha Singh,

[4]Bharat Bhushan Naib

[1]Research Scholar, C.S.E, PDMCE, India
[2]Research Scholar
[3]Prof. & Head. CSE, PDMCE, India
[4]Assistant Prof., CSE, PDMCE
[1]drallsujata@gmail.com, [2]gurdeepdrall@gmail.com, [3]sugandha_engg@pdm.ac.in,
[4]bharat_engg@pdm.ac.in

Abstract

Chronic kidney is a condition when kidney gradually losses its capability to filter out waste fluids from the blood. Generally Glomerular filtration rate is measured for diagnosis of chronic kidney disease. It's a long term disease and can be caused by Diabetes, High B.P, Smoking and much more factors. Complicated Chronic kidney can lead to Anaemia, Heart Disease, and High level of potassium. Various researches have been carried out in the field of chronic kidney disease prediction using various data mining and machine learning approaches. Data mining helps to examine big database to recognise patterns and solve problems. In this paper we have discussed and reviewed various classification methodologies which are used for Chronic Kidney Disease prediction. These researches will help to come up with new machine learning approaches and techniques which are more effective and efficient in performing the prediction task. Machine learning is the new science in which we make computers to automatically learn from past experiences

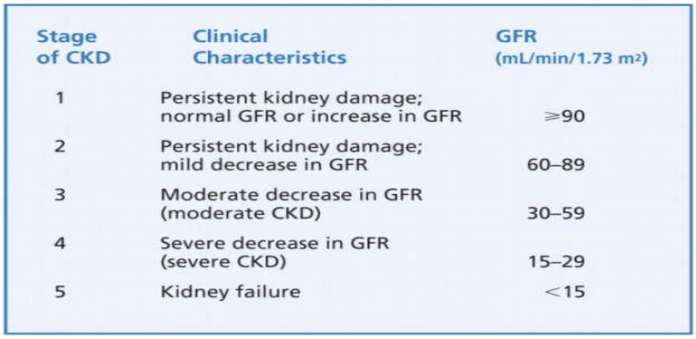
Keywords: Chronic Kidney Disease, Symptoms, Classification algorithms, Machine learning, Data mining, Predictive models

1. Introduction

Chronic Kidney is progressive loss of kidney function over a period of time and if not cured leads to permanent loss of kidney [9]. Most of the time this disease gets undetected until the condition gets severe. Several researches have been done in the field of Disease prediction using machine learning approach and data mining. Data mining is the process of converting raw data to useful important data [1]. Hidden patterns and correlations are recognised according to different business requirements by applying machine learning. Machine learning can be categorised as supervised and un-supervised [28].It gives machine the ability to learn from past experiences and patterns and predict the similar patterns and outcomes. Supervised learning can defined as when we map an input to a desired output. Machine learning algorithms are provided to support future predictions. There are various supervised machine learning algorithms like Logistic regression, multi-class classification, support vector machine, K-nearest neighbour, Naïve Bayes we train the data using information which is not labelled. In this algorithm we divide the data into two groups based on similarity and reducing the dimensionality. Most common unsupervised learning approaches are clustering algorithms. There are various clustering algorithms like Hierarchical Clustering, K-means clustering and many more [28], [10]. Kidneys are the organs that filter out blood and produce waste material which is then excreted via urine. In healthy person kidneys keep a balance of water and minerals. Kidneys also produce Renin which is used by body to maintain Blood pressure. Kidney also makes vitamin D, which is required for keeping bones health.

Diabetes, High Blood pressure, decreased glomerular filtration rate (GFR) can lead to chronic kidney diseases. Various data mining and machine learning approaches have been used for chronic kidney disease prediction. Researches have used multiple machine learning algorithm's like Support

vector machine, Linear regression, Naïve Bayes, Multiclass forest, Decision Tree, Random Forest and many more for the occurrence of Kidney and other diseases.



Stage of CKD	Clinical Characteristics	GFR (mL/min/1.73 m²)
1	Persistent kidney damage; normal GFR or increase in GFR	⩾90
2	Persistent kidney damage; mild decrease in GFR	60–89
3	Moderate decrease in GFR (moderate CKD)	30–59
4	Severe decrease in GFR (severe CKD)	15–29
5	Kidney failure	<15

Figure 1. Stages of chronic kidney disease

Figure 1, describes various stages of chronic kidney diseases on basis of glomerular filtration rate (GFR) value. People in stage 1 chronic kidney disease damage have GFR value at normal or greater than 90ml/min. This stage goes undetected as kidneys do good job even when the GFR value is not 100ml/min. People in stage 2 of CKD have little decrease in GFR which lies in range 60-89 ml/min. in stage 3 CKD kidneys suffers from moderate damage. At this stage complications are likely to develop in kidney. The range of GFR in stage 3 lies from 30-59 ml/min. in CKD Stage 4 a person is likely to suffer from severe kidney damage and may suffer from high blood pressure, heart disease. The GFR value in stage 4 lies in range 15-29 ml/min. The last stage of chronic kidney disease is stage 5. The GFR value goes is 15ml/min or less. In this stage Kidneys lose all their ability and are no longer able to remove waste fluids and toxins from blood.

2. Related Work

Perera, K. D. M et al. in 2017 carried out evaluation on machine learning classification methods for predicting Chronic Kidney Disease through Data analytics [2]. For making this prediction model they initially downloaded the data of 400 people with their disease related attributes from UCI have used machine learning and different classification algorithms like multiclass neural networks, and other multiclass classification algorithms. In their work same data was used by all the classification algorithms and results for the occurrence of chronic kidney were predicted. In their research Multiclass decision forest predicted the occurrence of chronic kidney by an accuracy of 99.1%, Multiclass Decision Jungle predicted the occurrence of CKD with an accuracy of 96.6%, Multiclass Logistic Regression performed with an accuracy of 95% and Multiclass neural networks performed with an accuracy of 97.5%. As per their experiment we can easily conclude that Multiclass Decision Forest algorithm came up with a highest accuracy of 99.1% in prediction of disease. This research can be used in developing prototypes model for predicting various diseases by using the disease specific data and applying these algorithms on it.

Kahandawaarachchi K.A.D.C.P.et al. in 2017 made a Dietary prediction for Kidney disease patients by using machine learning algorithms and acknowledging blood potassium level [3]. They have suggested diet plans by using blood potassium levels of different patients. In their diet plan they have marked people with potassium level between 3.5-5.0 on the safe zone , people with potassium levels between 5.1-6.0 in the Caution Zone and people with blood Potassium level higher than 6.1 in the Danger zone and prevail more chances of having chronic kidney disease. For their research they have collected data of 400 people from UCI repository. In their research they have used four Multi class classification algorithms each with different prediction accuracy of chronic kidney disease. In their research Multiclass Decision Forest algorithm performed with an accuracy of 99.17%, multiclass decision jungle algorithm performed with an accuracy of 97.50%, Multiclass Logistic regression performed with an accuracy of 89.10% and Multiclass Neural networks performed with an

accuracy of 82.50%. Their experiment analysis shows that Multiclass decision Forest classification algorithm predicted with highest accuracy of 99.17% for the occurrence of the chronic kidney disease. As per their research a suitable diet plan will help in treatment of patient suffering from CKD.

Yildirim P. in July 2017 conducted Kidney disease prediction on imbalanced data by using multilayer perceptron classification [4]. In their research they have focused on effect of imbalance data on developing neural network classifiers. A neural network is a group of neurons grouped together in layers which convert input data to output data by applying non-linear functions on it. In their research Multilayer perceptron algorithm performed with precision of 99.8%.

Radha, N in 2016 performed diagnosis of chronic kidney using machine learning algorithm [5]. Their work dealt in determining kidney failure by using classification algorithms. Their work looks forward to increase the accuracy of disease prediction and decrease the occurrence of chronic kidney diseases. They have also focused on stages on different stages of kidney basis on severity. Among the classification algorithms they have used back propagation neural networks, random forests and Radial basic function. In their work the data used is collected is real data of 1000 patients with 15 attributes related to kidney disease. This data belongs to different laboratories in South India. As per their study the main attribute which contributes to Chronic kidney disease is Glomerular filtration rate and the value of GFR determines the severity of chronic kidney. They have built this classification model in R Programming language. The test results of various algorithms are calculated and compared on the basis of sensitivity, specificity, and accuracy. The Back propagation neural network performed with an accuracy of 80.4%, the radial basis function neural network performed with an accuracy of 85.3% and the random forest algorithm performed with an accuracy of 78.6%. The results are compares and it can be seen that Radial basis function performed with highest accuracy of 85.3% among the three algorithms and thus was adopted as the best algorithm for predicting chronic kidney.

Ayesha, M in 2016 performed chronic kidney disease prediction using Naïve Bayes classifier [6]. Naïve Bayes belong to supervised learning of machine learning. In their work data set was downloaded and pre-processed by replacing missing values, noisy data. All the missing numerical attributes are replaced by Median value of that attribute. Feature selection with OneR classifier was used. Different stages of disease were predicted by calculating GFR value, and by using those predicted stage value action rules were generated. Data set used is downloaded from UCI repository. Number of attributes was reduced in order to increase the percentage. OneR attribute evaluator with Naïve Bayes selected 5 attributes from a total of 25 attributes, thus reducing the percentage of attribute by 80% and with an overall accuracy of 97.5%.

Briggs, D in 2017 used Random forest and decision tree classification model for predicting incompatibility of kidney transplantation [7]. Their research deals with kidney transplantation and this approach can be used in other organ transplantation field. They have worked on data of 80 people collected from university hospital convent and Warwickshire, UK for prediction outcome in kidney transplantation. They have identified key factors which lead to organ rejection. The donor specific igG and the number of leucocytes mismatch of both the recipient and donor can cause kidney transplantation failure. They have used classification algorithms like Decision tree and Random forest for predicting the transplant rejection while keeping all the cause factors in notice. A set of 14 parameters has been used for the data set of 80 people. Data of 60 people was used for training the machine and the rest 20 datasets were tested for prediction accuracy. Both of the prediction models predicted the outcome of kidney transplant with an accuracy of 85%.

Radha, N in 2015 used machine learning approach for predicting chronic kidney disease [8]. In their work they have used different classification algorithms like Naïve Bayes, KNN and Support vector machine and have compared the results of these algorithms on basis of disease prediction accuracy. Medical data for this research was taken from different labs of south India. The data is real data and consist of record of 1000 people with their 14 chronic kidney disease related attributes like blood urea, uric acid and many more. In selected classification algorithms, KNN performed with an

accuracy of 98%, SVM performed with an accuracy of 89.9%, Naïve Bayes performed with an accuracy of 61.8% and Decision tree performed with an accuracy of 78.6%, thus we can draw a conclusion that KNN proved to be the best prediction algorithm among the all taken in this research. This prediction approach can be used for other disease as well by taking their attributes and data set.

3. Methodology

Various Methodologies are used by Data Miners and Data scientists. Analytics Solutions Unified Method for data Mining/Predictive Analytics – Developed by IBM [19], Cross Industry Standard Process for data Mining (CRISP-DM) [20]. In our paper we will take in consideration - Cross Industry Standard Process for data Mining (CRISP-DM) as our Research Methodology. This is one of the leading methodologies used by Data scientists and Data Miners.

Figure 2. Cross industry standard process for data mining

Figure 2 describes various phases of Cross Industry Standard Process for data Mining-(CRISP-DM).

3.1. The Phases in CRISP-DM methodology are as Follows:

i.Business Understanding: Here we look after the Business purpose and what are its demands, thus designing a model to fulfil the purpose. Our main focus is to find CKD status of a patient with highest efficiency in least possible time. This will help Doctors to anticipate the Disease more precisely and in less time.

ii.Data Understanding: This begins with gathering of data as per business requirements. We look at the business plan and search for the kind of data required to achieve our goal. Data can be collected from various health centres, online websites and some data repositories which provide data free of cost.

iii.Data Preparation: Once data is available, exploratory data analysis is done, it is Pre-processed. The Nominal values are transformed to Real Numbers. As data taken is real, thus contains missing values which are then replaced by the Mean value of selected attribute. This does not bring change into the original data. Correlation between the attributes is found in order to identify highly co related disease causing attributes.

iv.Modelling: In this phase few attributes are selected. Highly Correlated attributes are chosen to make the machine learn in a more appropriate and precise manner and obtain higher accuracy. The Prediction classification model is to be created.

v.Evaluation: In this phase, model is evaluated using on basis of different factors such as Precision and Recall, Accuracy. Highest accuracy classification algorithm is chosen the best for classification model.

vi.Deployment: The classification model for predicting chronic kidney disease can be created in any language programming language.

4. Classification Algorithms

The most widely used classification algorithms are :

i. Naive Bayes: Naïve Bayes are probabilistic classifiers, which are based on Bayes Theorem [11]. In Naïve Bayes each attribute is considered independent of the value of any other feature. For variable evaluation Naïve Bayes uses the method of Maximum likelihood. This algorithm is fast and can be used for making real time predictions. Naïve Bayes algorithm can be used for Text classification, Sentiment Analysis.

ii. K-Nearest Neighbor: This algorithm is inspired by human reasoning. When a new situation occurs, KNN scans through all the past experiences called as data points and looks up the closest experience to find a solution [12]. For a new data point, we can determine the most likely class for prediction by looking up to the closest classes. It stores all the available classes and find out the new class based on the majority votes of neighbour classes. This algorithm is fast, easy and tune to different predictive problems.

iii. Support Vector Machine: SVM is a supervised machine learning algorithm [13]. SVM are commonly used for in classification problems. SVM model is a representation of the data point in space mapped so that different points can be categorized into separate categories. This algorithm is based on finding a hyperplane that divides a dataset into 2 categories and the goal is to decide which class a new data point will belong to. Hyperplane can be taken as a line that linearly separates a set of data. More the margin between our data point, the more we are accurately classified.

iv. Multiclass Decision Forest*:* This is a group of learning methods used for classification. Multiple decision trees are built up in this algorithms and the most popular output class is chosen [14], [15]. Voting is done in the form of clustering in which each decision tree gives an output of non-normalized probabilities. The clustering process adds up all these frequencies to get probabilities of each. The tree which has the most prediction confidence has more weight in decision in the final decision.

v. Logistic Regression*:* Equations are used in Logistic regression for representation. Input variables are taken with their coefficient values to predict the output value. Logistic regression uses 0 and 1 as output values rather than any numeric value [16], [17]. Logistic regression is used where prediction can be done with either positive or negative, 0 or 1. The maximum likelihood estimation algorithm finds out the regression coefficient which accurately predicts the probability of dependent variable. Logistic Regression can be used to predict several outcomes with yes or no, 0 or 1 value.

vi. Multiclass Decision Jungle*:* This is one of the widely used classification algorithm. It is a supervised learning algorithm. The basic idea behind this algorithm is to map all the possible decision paths in the form of a tree. It is an extension of decision forest. Decision Jungle consist of directed acyclic graphs. Directed acyclic graphs in decision jungle stores less amount of memory space, it allows the tree branches to merge thus taking less space [18]. DAG in decision tree allows multiple paths from a node to every leaf. When large amount of data is gives to a decision jungle the decision tree will grow exponentially in depth.

vii. Multilayer Perceptron*:* This classification algorithm belongs to feedforward artificial neural network.

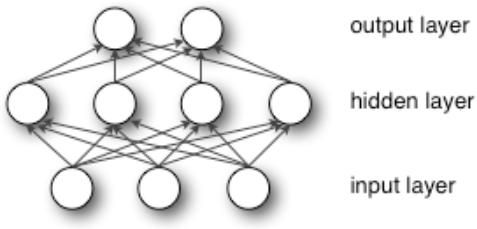


Figure 3. Multilayer perceptron model

Figure 3 describes the basic structure of multilayer perceptron model. It consists of an input layer, an arbitrary hidden layer and an output layer. [21] The input layer receives the signal, the hidden layer act as a computational engine capable of approximating any condition and thus giving the results to the output layer. It is a form of supervised learning and they train the input set and learn from the correlation predicted from the input and output and use them for prediction on test data.[22] Multilayer perceptron model consist of 3 or more than 3 layers and all the nodes in the layers are fully interconnected.

viii. Multiclass Neural Networks*:* It is a supervised machine learning algorithm. It consist of interconnected layers and the input layer being the first layer which is connected to the output layers by acyclic graph with edges and nodes having some weight [23],[24]. The input and output layers are fully connected to the hidden layer. The number of nodes in input layers depends upon the type of input data. The user can set the number of hidden layers which act as basic computational layers. The default value for these hidden layers is 100 nodes.

ix. Radial Basis Neural Network*:* This is a type of neural network. The radial basis neural network consists of an input layer. RBF neurons are in hidden layer and the output nodes. The input layer is N-dimensional, The RBF neurons store a prototype value, and it compares the input value with prototype value and gives an outputs value of either 0 or 1. This value depends upon the distance between the input value and the prototype value [25]. When the distance is large the response goes more towards 0. The output layer consists of a set of nodes. The neurons response values can also be termed as activation value. The shape of the RBF neurons response in a bell shaped curve. The prototype value with RBF neurons is also called as neurons centre as it form the centre part of the bell curve.

5. Comparison Of Various Prediction Algorithms

Table1: Comparison of classification algorithms

S.no	Data Source	Classification algorithms	Performance
1	[26]	Multiclass Logistic Regression	95%
		Multiclass Decision Jungle	96.60%
		Multiclass Decision Forest	99.10%
		Multiclass Neural Networks	97.50%
2	[26]	Multiclass Decision Forest	99.17%
		Multiclass Decision Jungle	97.50%
		Multiclass Logistic Regression	89.10%
		Multiclass Neural Networks	82.50%
3	[26]	Multilayer Perceptron	99.80%
4	Laboratories in Coimbatore, India	Back Propagation Neural Network	80.40%
		Radial Basis Function Neural Network	85.30%
		Random Forest	78.60%
5	[26]	Naïve Bayes	97.50%
6	[27]	Random forest	85%
		Decision tree Classification	85%
7	Laboratories in Coimbatore, India	Naïve Bayes	61.80%
		KNN	98%
		SVM	89.9%

In Table 1 various classification algorithms have been used by researchers for predicting chronic kidney disease and its other forms on the basis of severity, GFR value, Diet, Kidney transplant complications. We have made a comparison of their prediction accuracy and tried to find out the best one.

6. Conclusion

There exist various classification algorithms and techniques that can be used for prediction of chronic kidney diseases. These techniques can also be used for prediction of other diseases. Researches have been done in different aspects related to chronic kidney disease. Prediction models were developed to find out the best chronic kidney disease prediction algorithm, to make a dietary prediction of potassium level in food for Chronic kidney disease patient in order to maintain a safe potassium zone in their body, prediction models have also worked with decreasing the number of disease related attributes and increasing the overall accuracy . UCI machine learning repository is widely used as a data source by researchers. The data available is real and contain dataset of 400 people with their respective 25 disease related attributes. Thus, UCI machine learning repository can be considered as a trusted source of data. In this paper we have gone through different research papers and have compared various classification algorithms used. Prediction of Chronic kidney disease is made using different classification algorithms and the highest prediction results are obtained by KNN, Naïve Byes and Multiclass decision forest and multilayer perceptron algorithms. Among them Multiclass perceptron achieved highest accuracy of 99.80%. Thus multilayer perceptron algorithm can be said as the best disease prediction algorithm among them.

References

[1] https://searchsqlserver.techtarget.com/definition/data-mining

[2] Gunarathne, W. H. S. D., K. D. M. Perera, and K. A. D. C. P. Kahandawaarachchi. "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)." In Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference on, pp. 291-296. IEEE, 2017.

[3] Wickramasinghe, M. P. N. M., D. M. Perera, and K. A. D. C. P. Kahandawaarachchi. "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms." In Life Sciences Conference (LSC), 2017 IEEE, pp. 300-303. IEEE, 2017.

[4] Yildirim, Pinar. "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction." In Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual, vol. 2, pp. 193-198. IEEE, 2017.

[5] Ramya, S., and N. Radha. "Diagnosis of chronic kidney disease using machine learning algorithms." International Journal of Innovative Research in Computer and Communication Engineering 4, no. 1 (2016): 812-820. IJIRCCE,2016.

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

[7] Shaikhina, Torgyn, Dave Lowe, Sunil Daga, David Briggs, Robert Higgins, and Natasha Khovanova. "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation." Biomedical Signal Processing and Control (2017).

[8] Radha, N., and S. Ramya. "Performance Analysis of Machine Learning Algorithms for Predicting Chronic Kidney Disease." (2015).

[9] https://www.medicalnewstoday.com/articles/172179.php

[10] https://www.mathworks.com/discovery/unsupervised-learning.html

[11] http://www.statsoft.com/textbook/naive-bayes-classifier

[12] https://www.analyticsvidya.com/blog/2018/03/introduction-

k-neighbours- algorithm-clustering/

[13] https://en.wikipedia.org/wiki/Support_vector_machine

[14] https://docs.microsoft.com/en-us/azure/machine-learning/

studio-module-reference/multiclass-decision-forest

[15] https://en.wikipedia.org/wiki/Random_forest

[16] https://machinelearningmastery.com/logistic-regression-for-

machine-learning/

[17] https://www.quantinsti.com/blog/machine-learning-logistic-

regression-python/

[18] https://docs.microsoft.com/en-us/azure/machine-learning/

studio-module-reference/multiclass-decision-jungle

[19] ftp://ftp.software.ibm.com/software/data/

sw-library/services/ASUM.pdf

[20] https://en.wikipedia.org/wiki/Crossindustrystandard_

process_for_data_mining

[21] https://en.wikipedia.org/wiki/Multilayer_perceptron

[22] https://deeplearning4j.org/multilayerperceptron

[23] https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-neural-network

[24] https://developers.google.com/machine-learning/crash-course

/multi-class-neural-networks/one-vs-all

[25] http://mccormickml.com/2013/08/15/radial-basis-function-network-rbfn-tutorial/

[26] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[27] https://www.datascience.com/blog/supervised-and-nsupervised-machine-learning-algorithms

[28] Sin, Katrina, and Loganathan Muthu. "APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS--A LITERATURE REVIEW." ICTACT Journal on soft computing 5, no. 4 (2015).

[29] Tang, Jiexiong, Chenwei Deng, and Guang-Bin Huang. "Extreme learning machine for multilayer perceptron." IEEE transactions on neural networks and learning systems 27, no. 4 (2016): 809-821. IEEE,2016.

[30] Vafeiadis, Thanasis, Konstantinos I. Diamantaras, George Sarigiannidis, and K. Ch Chatzisavvas. "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory 55 (2015): 1-9.

[31] Madni, Hussain Ahmad, Zahid Anwar, and Munam Ali Shah. "Data mining techniques and applications—A decade review." In Automation and Computing (ICAC), 2017 23rd International Conference on, pp. 1-7. IEEE, 2017.

[32] Lian, R. J. (2014). Adaptive self-organizing fuzzy sliding-mode radial basis-function neural-network controller for robotic systems. IEEE Transactions on Industrial Electronics, 61(3), 1493-1503.

[33] Wang, Tong, Huijun Gao, and Jianbin Qiu. "A combined adaptive neural network and nonlinear model predictive control for multirate networked industrial process control." IEEE Transactions on Neural Networks and Learning Systems 27, no. 2 (2016): 416-425. IEEE,2016.