Detection of Chronic Kidney Disease Using Artificial Neural Network

Chakrapani
B-Tech Student, Department of Computer Science and Engineering, Gaya College of Engineering, Gaya- Bihar, India

Sumit Raj
B-Tech Student, Department of Computer Science and Engineering, Gaya College of Engineering, Gaya-Bihar, India

VibhavPrakash Singh
Assistant Professor, Department of Computer Science andEngineering, Gaya College of Engineering, Gaya-Bihar, India

DhrubJyotiKalita
Assistant Professor,Department of Computer Science and Engineering, Gaya College of Engineering, Gaya-Bihar, India

Abstract
Disease classification/detection is a crucial and challenging problem, because it helps in early diagnosis of disease by supporting the pathologists and doctors in their decision. Machine Learning technique is one of the emerging field can be used in the health sectors for the diagnosis of different diseases. This paper presents an effective approach for the diagnosis of chronic kidney disease(CKD) using artificial neural network (ANN) with back propagation algorithm, where first we fill the missing values of the dataset using mean, mode and median of attributes. Further, we have trained the NN classifier and evaluate the detection performances on separate test dataset. From the comparative analysis with other variants of classifiers like SVM, K-NN, Classification and Regression tree it is found that the recognition accuracy of ANN is significantly encouraging.

Keywords:CKD, Pre-processing, Supervised learning, ANN, Classification.

Introduction

CKDis the lasting damage to kidney that can get worse over the time. If the kidney damage up to last stage than it may stop working.Generally people suffer with this disease with their age, but recently from 5 years children and youth also suffering from CKD disease. The main task of the kidney is to filter out waste products and excess fluid from body which is then passed out through urine.But in CKD kidney loose their functionality due to which some excess amount of urine mix with blood and also some protein mix with urine.There are some symptoms which shows kidneys are beginning to fail likemuscle cramps, nausea and vomiting, appetite losses, swelling in your feet and ankles, too much urine or not enough urine, trouble catching your breath, trouble sleeping, fever and vomiting[8].From last 15 years data it has been noticed that increase in number of patient which are suffering from CKD disease. And more than 60% patients not receiving medical attention [9].Therefore, early diagnosis and detection of this disease can help the patients in recovery on the right time.

Data mining and machine learning can be used as an informative tool to extract the useful information which helpspathologists and doctors inpromtdecisionsmaking[11].Today'ssome researchers are working on CKD by applying different computational techniques for the prediction and diagnosis of this disease. Boukenze et al. [2] has given analysis on CKD dataset , by applying different data mining techniques i.e., support vector machine (SVM), Decision tree (C4.5) and Bayesian Network in WEKA data mining tool. They marked Decision tree(C4.5) is best as compared to others.Nishanthet al. [3] has used linear discriminat analysis (LDA), and K-nearest neighbor (K-NN) algorithm. Also they have used two methods to remove the attributes i.e., Omit One Method and Four Attribute Combination Method. They have used only 18 important attributesfrom 24 attributes for prediction of CKD disease. Through analysis they found LDA is better than K-NN and having accuracy above than 98%.Vijayaet al. [5]hasgiventhe analysis of CKD

dataset using SVM and ANNand they reported 76.32% and 87.70% recognition accuracy.Aroraetal. [6]hasgivenanalysis to predict the CKD using WEKA data mining tool.They used Naïve Bayes, J48(Decision Tree) and Sequential Minimal Optimization(SMO) for SVM using Linear Kernel algorithm.Theyfound that Naïve Bayeshas 95.5556% accuracy, J48 has 99% accuracy and SMO has 97.75% accuracy.Lakshmi et al. [4] has used three data mining techniques i.e., ANN, Decision tree and Logical regression .They found that ANNhas 93.852% accuracy, Decision tree has 78.456% accuracy and Logical regression has 74.745 % accuracy. Jayalakshmi et al [7]analyzed the results obtained by applying different data mining techniques in MATLAB tool. They discuss about techniques used like ANN, Naïve Bayes, SVM, MBPN, Ada boost classifier, LDA, K-NN etc.Charleonnan et al[1] has given analysis of chronic kidney disease using KNN, SVM,and Decision Tree classifiers and they trained and tested the model using 5-fold cross validation. Theyhave also compared the accuracy, sensitivity and specificity between different classifiers, and finally concluded that SVM is appropriate for the prediction of CKD.

In this paper, we have used the mean, mode and median based pre-processing techniques for the missing values. Further, we have used the feed-forward neural network using backpropagation and Levenberg-Marquardt(L-M) training algorithm. Also,the classification performance of several machine learning techniques such as K-NN,SVM, classification tree, regression tree are compared. This paper work isorganized as follows; Section 2 presents the methods and model, section 3 sheds some light on experimental analysis and finally, section 4 is concluding section with further discussions.

Methods and Model

In this paper, we have used the techniques of data mining and machine learning for the early diagnosis of the CKD. The proposed framework is shown in Fig. 1, in which datasets are firstly pre-processed by data mining statistical techniques. To fill the missing values of dataset we have used the

Table 1 Filling of missing values using mean, mode and medianthree different statistical methods like mean, median and mode. These values are calculated only for the missing value attributes. Table 1 gives the glimpse of the applied statistical methods and corresponding filled values. For thenominal attributes, we have taken mode and median and for the numerical type of attributes we have taken the mean of the values.

Table 1: Filling of missing values using mean,mode, and median

S.No	Attribute	Statistical methods	Filled missing values
1	age	Mean	51
2	bp	Mean	76
3	sg	Median	1.02
4	Bgr	Mean	148
5	Bu	Mean	57
6	Sc	Mean	3.0
7	Sod	Mean	137
8	Pot	Mean	4.6
10	Hemo	Mean	12.5
11	Pcv	Mean	38.8
12	Wc	Mean	8413
13	Al	Mean	0
14	Su	Mean	0
15	Rbc	Mean	Normal
16	Pc	Mean	Normal
17	Pcc	Mean	NotPresent
18	Ba	Mean	NotPresent
19	Htn	Mean	No
20	Dm	Mean	No
21	Cad	Mean	No
22	Appet	Mean	Good
23	Pe	Mean	No
24	Ane	Mean	No

Table 2: Classification Techniques [8, 12]

S.N	Classifier	Descriptions
1	K-NN	In this classifier the representative of every class is selected. The classification is performed by assigning each tuple to the class to which it is most similar.
2	SVM	It is a robust classifier based on the concept of support vector that separate tuples with a linear decision surface and maximizes the margin of separation between the classes to be classified.
3	Decision Tree	Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

4	Neural Network (NN)	NN trained repeatedly, weight updated repeatedly to minimize error or gradient and increase performance of network using sigmoid function i.e., network propagated back again and again through back propagation technique. We can retrain our model again and again for best training,validation and testing.In NN three layer defined the architecture of this classifier. These layers are: Input Layer- Number of neurons contain input layer is equal to number of features in data, Output layer- NN has exactly one output layer, and Hidden Layer- There is no hard and fast rule for selection of hidden layers. We generally select one hidden layer when our data is smaller and number of hidden layer can be more for large data. If we increase hidden layer means our time complexity increases but our accuracy can also increase .

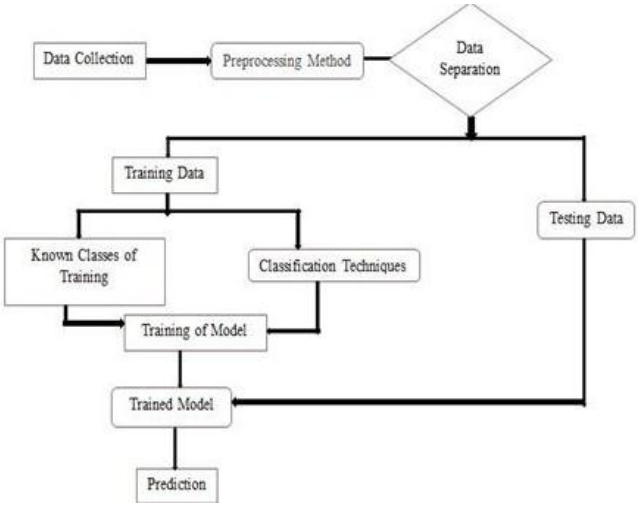


Figure 1: Working diagram

After pre-processing of the dataset, data is divided into two sections i.e training and testing as shown in Figure 1. Further, training data with their known target classes is used for the traning of the classifier, and after the traning of classifier seperate test data is fed into trained classifier. This trained classifier detects the class of sample query.For the performance analysis of different classifiers same process are repeated. In this study, we have evaluated the detection performance on four supervised machine learning techniques, which descriptions are given in Table 2

Results analysis and discussion

Result analysis and discussion of the proposed work is divided into two sections, where section 1gives the details of the used dataset , and section 2 sheds light on the further discussions of results.

Dataset details

Table 3: Details of benchmark dataset

	Attribute	Representation	Information Attribute	Description
1	Age	Age	Numerical	Years
2	Blood Pressure	Bp	Numerical	Mm/Hg
3	Specific Gravity	Sg	Nominal	1.005,1.010,1.015, 1.020, 1.025
4	Albumin	Al	Nominal	0,1,2,3,4,5
5	Sugar	Su	Nominal	0,1,2,3,4,5
6	Red Blood Cell	Rbc	Nominal	Normal,Abnormal
7	Pus Cell	Pc	Nominal	Normal,Abnormal
8	Pus CellClumps	Pcc	Nominal	Present,NotPresent
9	Bacteria	Ba	Nominal	Present,NotPresent
10	Blood Glucose Random	Bgr	Numerical	mgs/dl
11	Blood Urea	Bu	Numerical	mgs/dl
12	Serum Creatinium	Sc	Numerical	mgs/dl
13	Sodium	So	Numerical	mEq/L
14	Potassium	Pot	Numerical	mEq/L
15	Haemoglobin	hemo	Numerical	Gms
16	Packed Cell Volume	Pcv	Numerical	cells
17	White Blood cell Count	Wc	Numerical	Cells/cumm
18	Red Blood Cell count	Rc	Numerical	Millions/cmm
19	Hypertension	Htn	Nominal	Yes,No
20	Diabetes Mellitus	Dm	Nominal	Yes,No
21	Coronary Artery Disease	Cad	Nominal	Yes,No
22	Appetite	appet	Nominal	Good,Poor
23	Pedal Edema	Pe	Nominal	Yes,No
24	Anemia	Ane	Nominal	Yes,No
25	Class	Class	Nominal	CKD, NotCKD

Table 4: Class distribution

Class	Distribution
CKD	250(62.5%)
NotCKD	150(37.5%)

In this paper, we have taken the CKD data set from UCI repository [10]. This dataset contains 400 samples of patients having 24 attributes for each sample. Last attribute of this dataset is the class value, which are used as target class for the training and testing of the classifier. In Table 3,we have given the details about the attributes, where Nominal value shows that data is separated into discrete categories and numerical shows the random data.In Table 4 ,we have given the and distribution of CKD dataset, where 250 samples belong to the class of CKD and 150 samples belongto the class of healthy cases (NotCKD).

Discussion

For the analysis of this dataset, we have applied hold-out method for sample selection, where randomly 72 % samples of this dataset is used for training and 28% samples are used for the testing (As shown in Figure 2). It means, we have selected 289 samples fortraining of K-NN, SVM,Classification Tree and Regression Tree and 111 samples for testing.

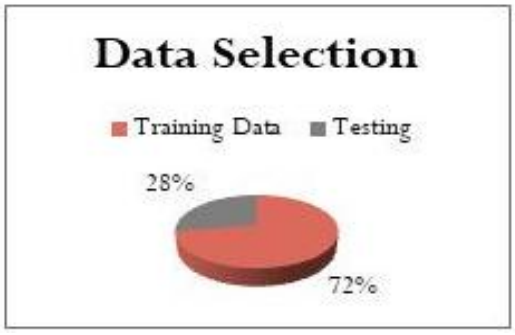
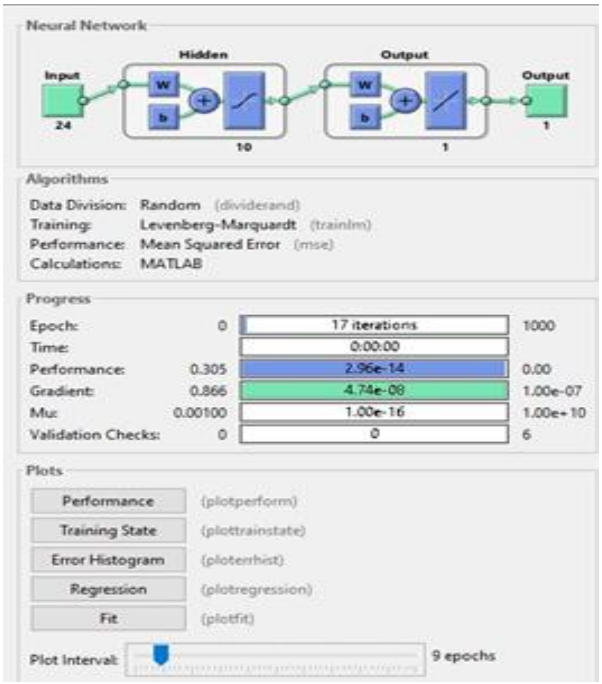


Figure 2:Data partition

For the traning and validation of NN we have taken same 289 samples, in which 203 samples are used for training,43 samples for validation and 43 for testing. In order to evaluate the performance of proposed work, we have constituted a confusion matrix, which holds TP (true positive), TN (true negaive), FN (false negative), FP (False positive) values. Where TP is the number of positive samples correctly predicted, TN is the number of negative samples correctly predicted, FN is the number of positive samples wrongly predicted, and FP is the number of negative samples wrongly predicted as positive. Using these definitions, which can calculate the recognition accuracy of the classifier.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \qquad (1)$$

For the training of NN , we have taken 1-hidden layer and defined 10-neurons in 1 hidden layer. We can increase the number of neurons but our data set is small and if we increase the number of neurons then our time for training and prediction will increase.Wehave selected Levenberg-MaRquardttraining algorithm, this algorithm take more memory but less time and training stops automatically whenperformance stops improving.In Figure 3a,we have shown the trained neural network by selecting samples randomly and calculating performance by mean squared error(MSE). Epoch generally message that how many times we updated our weight and bias to get performance(tends to zero) as well as gradient or error (tends to zero).This model is trained randomly ,and we get best result and after 17iteration. In Figure 3b, there is graph between MSE and Epoch. In this graph blue line shows the training of NN, which reflects minimum MSE at 17 epochs, Green line shows for the validation of samples given after training, and red line shows the performance on the testing dataset.



In Figure 4 ,Regressiom R, measures correction between output and target. As R value tends to one means close relationship and as tends to zero means no any relationship. From the figure, we can see that proposed NN based framework performance is significantly encouraging for the traning, validation and test dataset. Further, overall accuracy of the traning is 99.981 % confirms the effectiveness of this classifier. Furthermore, for the final testing we have evaluated all the samples which are not given at thetraning time. These samples are act as real time sample for the analysis. We can see from Table 3 that NN using LM traning algorithm gives 99.91% recongnitin accuracy on the test data. Same testing set is also compared with K-NN, SVM, Classification Tree and regression tree classifires. From the results as given in Table 5, it is cleared that the diagnosis accuracy of the NN is significantly encouraging than

other classifiers. Here out of 111 samples only one sample is wrongly mis-classified.

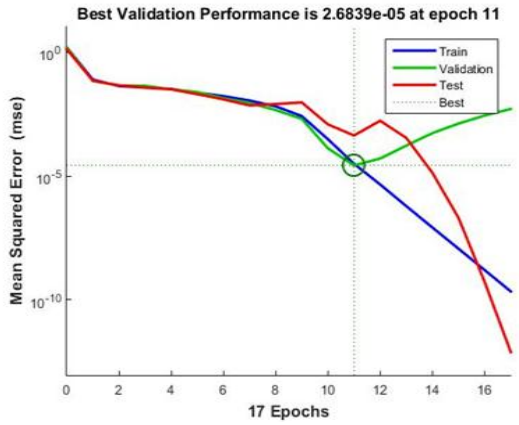


Figure 3(a-b): Training, MSE and Epochrelation

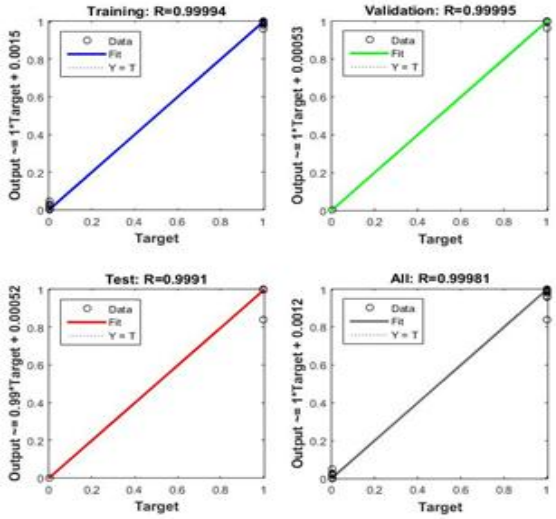


Figure.4:Performance analysis of ANN

Table 5: Comparative performance analysis

Applied Techniques	Accuracy(%)
KNN	72.90
SVM	95.49
Classification Tree	99.09
Regression Tree	95.49
ANN	99.19

Conclusions

CKD classification and detection can be used as fast interpretating and analyzing tool for providing the second opinion to the doctors and pathologists. This paper presented an approach for the prediction of chronic kidney disease using data mining and machine learning techniques.After the experimental analysis it was found that the classification and detection accuracy of mean, mode and median based pre-processing techniqueswithneural network was significantly encouraging than K-NN, SVM, Regression Tree and Classification Tree. Therefore, we can used this framework for the better prediction of chronic kidney disease.

References

[1] Charleonnan,Anusorn,ThipwanFufaung,TippawanNiyomwong, WandeeChokchueypattanakit, SathitSuwannawach, and NitatNinchawee. "Predictive analytics for chronic kidney disease using machine learning techniques." In *2016 Management and Innovation Technology International Conference (MITicon)*, pp. MIT-80. IEEE, 2016.

[2] Boukenze, Basma, HajarMousannif, and AbdelkrimHaqiq. "Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease." *IJDMS*8, no. 3 (2016): 1-9.

[3] Nishanth, Anandanadarajah, and TharmarajahThiruvaran. "Identifying important attributes for early detection of Chronic Kidney Disease." *IEEE reviews in biomedical engineering* 11 (2018): 208-216.

[4] Lakshmi,K.R.,Y.Nagesh, and M.Veera Krishna. "Performance comparison of three data mining techniques for predicting kidney dialysis survivability." *International Journal of Advances in Engineering & Technology* 7, no. 1 (2014): 242.

[5] Vijayarani, S., S. Dhayanand, and M. Phil. "Kidney disease prediction using SVM and ANN algorithms." *International Journal of Computing and Business Research (IJCBR)* 6, no. 2 (2015).

[6] Arora, Milandeep, and Er Ajay Sharma. "Chronic Kidney Disease Detection by Analyzing Medical Datasets in Weka." *International Journal of Computer Application* 6, no. 4 (2016): 20-26.

[7] Jayalakshmi, V., and LipsaNayak. "A Survey on Chronic Kidney Disease Detection Using Novel Methods."

[8] Singh, VibhavPrakash, SubodhSrivastava, and Rajeev Srivastava. "An efficient image retrieval based on fusion of fast features and query image classification." *International Journal of Rough Sets and Data Analysis (IJRSDA)* 4, no. 1 (2017): 19-37.

[9] www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/

[10] https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

[11] Singh, VibhavPrakash, SubodhSrivastava, and Rajeev Srivastava. "Effective mammogram classification based on center symmetric-LBP features in wavelet domain using random forests." *Technology and Health Care* 25, no. 4 (2017): 709-727.

[12] https://www.saedsayad.com/decision_tree.htm