JES
Journal of Engineering Sciences

Vol 13, Issue 04, APRIL /2022
ISSN NO:0377-9254

Identification Of Chronic Kidney Disease (CKD) Using Artificial Neural Networks Algorithms

Lakkoju Sairam Yashwanth [1], Chengalvala Anudeep[2], Gorijilli Venkata Sai Mukesh[3], Dr. Kompella Subhadra[4]

#1, #2,#3, Student, Department of CSE, GITAM (deemed to be University), Gandhi nagar Rushikonda Visakhapatnam530045 Andhra Pradesh, INDIA.

#4Assistant Professor, Department of CSE, GITAM (deemed to be University), Gandhi nagar Rushikonda Visakhapatnam530045 Andhra Pradesh, INDIA.

ABSTRACT_ The field of biosciences have advanced to a larger extent and have generated large amounts of information from Electronic Health Records. This have given rise to the acute need of knowledge generation from this enormous amount of data. Data mining methods and machine learning play a major role in this aspect of biosciences. Chronic Kidney Disease(CKD) is a condition in which the kidneys are damaged and cannot filter blood as they always do. A family history of kidney diseases or failure, high blood pressure, type 2 diabetes may lead to CKD. This is a lasting damage to the kidney and chances of getting worser by time is high. 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.

1.INTRODUCTION

Data mining plays a vital role in health care domain, nowadays. There is an increased need for an efficient analytical methodology to detect unknown and valuable information in health data. It produces huge amount of data about patients, diseases, diagnosis and medicines so on. By applying data mining techniques in health care domain, the administrators can improve the quality of service by discovering latent, potentially useful trends required by medical diagnosis. In the health care industry, the data mining is mainly used for predicting the diseases from the datasets [1].

Classification is one of the most widely used methods of data mining in healthcare organization. The classification technique predicts the target class for each data points. The classification methods such as Decision Tree, Support Vector Machine, K-Nearest Neighbour, Naïve Bayes and Neural Network. Decision Tree is widely used by many researchers in healthcare field like skin diseases and chronic kidney disease etc. The K-Nearest Neighbour is used to analyse the relationship between cardiovascular disease, hypertension and the risk factors of various chronic diseases in order to construct an early warning system. Multilayer Neural Network is used for diagnosis of various chest related diseases such as Lung Cancer, Asthma, and Pneumonia etc.

Chronic kidney diseases have become a major public health problem. Chronic diseases are a leading cause of morbidity and mortality in India. Chronic kidney diseases account for 60% of all deaths worldwide. Eighty percentage of chronic disease deaths worldwide occur in low- and middle-income countries [2]. The National Kidney foundation determines the different stages of chronic kidney disease based on the presence of kidney damage and glomerular filtration rate (GFR), which is measure a level of kidney function. There are five stages of chronic kidney disease.

Vol 13, Issue 04, APRIL /2022
ISSN NO:0377-9254

2. LITERATURE SURVEY

Lambodar Jena et. al [3] presented prediction of chronic kidney disease using Naive Bayes, Multilayer Perceptron, Support Vector Machine, J48, Conjuctive Rule and Decision Table. The performance of these algorithms is measured by classification accuracy, time taken to build model, time taken to test the model, mean absolute error, Kappa statistics and ROC Area. From the experimental result, the Multilayer Perceptron algorithm gives better than the other five algorithms with the classification accuracy of 99.7%.

Abeer et. al [4] presented a new clinical decision support system for diagnosing patients with chronic renal failure. The aim of this work is to improving performance of a previously reported Chronic Renal Failure diagnosis system which was based on Artificial Neural Network, Decision Tree and Naïve Bayes. Data Mining Classifier like Support Vector Machine and Logistic Regression is to compare the performance of these algorithms on the basis of its diagnostic accuracy, sensitivity and specificity. From the experimental result it is observed that the performance of the Support Vector Machine is better than the other algorithm with the accuracy of 93.1%.

Jerlin Rubini et. al [5] proposed a new chronic kidney disease dataset with three classifiers such as radial basis function network, multilayer Perceptron, and logistic regression. The obtained result of this experiment shows in terms of prediction accuracy, error rate, sensitivity, specificity and F-score. Accuracy of these three classifiers is evaluated. The performance of multilayer Perceptron is better than the other two algorithms with the accuracy of 99.7%.

Ruey Kei Chiu et. al [6] developed a best-fitting neural network model to detect various severity levels of chronic kidney disease. Back-Propagation Network (BPN), Generalized Feed Forward Neural Networks (GRNN), and Modular Neural Network (MNN) are used. The performance of these algorithms is measured by accuracy, sensitivity, and specificity. From the experimental result, the Back-Propagation Network provides better than the other two algorithms with the accuracy of 94.7%.

Jayalakshmi et. al [7] discussed a research to increase the performance of the network in terms of accuracy. The accuracy was increased by using three key concepts: missing data replacement, data pre-processing and introducing the Performance Vector (PV) in the search direction. The results of the network have been tested using Pima Indian Diabetes Dataset. The experimental system improves the performance more than 7% than the standard Gradient Descent method. The accuracy rate of improved Gradient Descent method is above 99%.

Koushal Kumar et. al [8] compared the performance of all three neural networks on the basis of its accuracy, time taken to build model, and training data set size. Learning vector quantization (LVQ), two layers feed forward Perceptron trained with back propagation training algorithm and Radial basis function (RBF) networks for diagnosis of kidney stone disease. The multilayer Perceptron with two hidden layers and back propagation algorithm is the better model for diagnosis of kidney stone disease. Its accuracy is 92% to diagnosis the kidney stone disease.

Rajalakshmi et. al [9] presented an efficient and effective model of forecast and classification of functional abnormalities of kidney using Associative Neural Network (ASNN) and Polynomial Neural Network (PNN). From the experimental result, the Associative Neural Network is better than the other classifier

algorithm with the squared correlation coefficient of 0.98%.

Vijayarani et. al [10] presented to predict kidney diseases by using Support Vector Machine (SVM) and Artificial Neural Network (ANN). The aim of this work is to compare the performance of these two algorithms on the basis of its accuracy and execution time. From the experimental results it is observed that the performance of the ANN is better than the other algorithm with the accuracy of 87%.

Shweta Kharya et. al [11] described various review and technical articles on breast cancer diagnosis and prognosis. The current research is being carried out using the data mining techniques to enhance the breast cancer diagnosis and prognosis. Among the various data mining classifiers and soft computing approaches, Decision tree is found to be better predictor with 93.62% accuracy on benchmark dataset (UCI machine learning dataset) and also on SEER dataset.

Abhinandan Dubey et. al **[**12] discussed the automated detection of diseases using Machine Learning Techniques. The K-Means Clustering Algorithm with a single mean vector of centroids, to classify and make clusters of varying probability of likeliness of suspect being prone to CKD. The results are obtained from a Real Case Data-Set from UCI Machine Learning Repository.

3.Methods and Methodology

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

Vol 13, Issue 04, APRIL /2022
ISSN NO:0377-9254

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

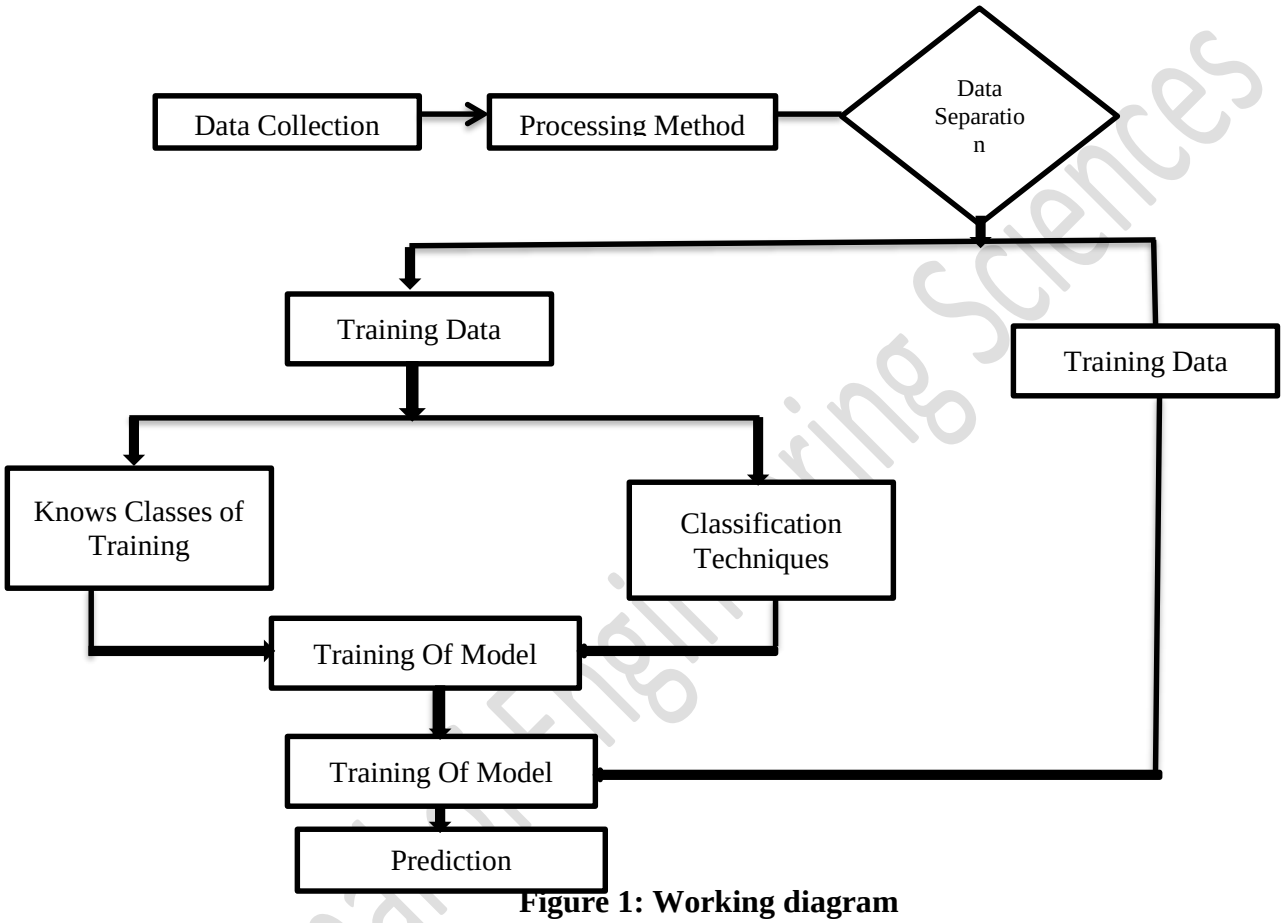

Figure 1: Working diagram

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

3.1 ALGORITHMS

Support vector machine:

Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyper plane which separates the data into classes. In

machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyper plane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyper plane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyper plane as possible, while still being on the correct side of it.

So when new testing data is added, whatever side of the hyper plane it lands will decide the class that we assign to it.

How do we find the right hyper plane?

Or, in other words, how do we best segregate the two classes within the data?

The distance between the hyper plane and the nearest data point from either set is known as the margin. The goal is to choose a hyper plane with the greatest possible margin between the hyper plane and any point within the training set, giving a greater chance of new data being classified correctly.

Both algorithms generate model from train dataset and new data will be applied on train model to predict it class. SVM algorithm is giving better prediction accuracy compare to ANN algorithm.

Naïve Bayes Classifier Algorithm

It would be difficult and practically impossible to classify a web page, a document, an email or any other lengthy text notes manually. This is where Naïve Bayes Classifier machine learning algorithm comes to the rescue. A classifier is a function that allocates a population's element value from one of the available categories. For instance, Spam Filtering is a popular application of Naïve Bayes algorithm. Spam filter here, is a classifier that assigns a label "Spam" or "Not Spam" to all the emails.

Naïve Bayes Classifier is amongst the most popular learning method grouped by similarities that works on the popular Bayes Theorem of Probability- to build machine learning models particularly for disease prediction and document classification. It is a simple classification of words based on Bayes Probability Theorem for subjective analysis of content.

Decision tree:

A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision, based on certain conditions. In a decision tree, the internal node represents a test on the attribute, each branch of the tree represents the outcome of the test and the leaf node represents a particular class label i.e. the decision made after computing all of the attributes.

The classification rules are represented through the path from root to the leaf node.

Types of Decision Trees

Classification Trees- These are considered as the default kind of decision trees used to separate a dataset into different classes, based on the response variable. These are generally used when the response variable is categorical in nature.

Regression Trees-When the response or target variable is continuous or numerical, regression trees are used. These are generally used in predictive type of problems when compared to classification.

Decision trees can also be classified into two types, based on the type of target variable- Continuous Variable Decision Trees and Binary Variable Decision Trees. It is the target variable that helps decide

what kind of decision tree would be required for a particular problem.

Random forest:

Random Forest is the go to machine learning algorithm that uses a bagging approach to create a bunch of decision trees with random subset of the data. A model is trained several times on random sample of the dataset to achieve good prediction performance from the random forest algorithm. In this ensemble learning method, the output of all the decision trees in the random forest, is combined to make the final prediction. The final prediction of the random forest algorithm is derived by polling the results of each decision tree or just by going with a prediction that appears the most times in the decision trees.

For instance, in the above example - if 5 friends decide that you will like restaurant R but only 2 friends decide that you will not like the restaurant then the final prediction is that, you will like restaurant R as majority always wins.

K – nearest neighbor:

***K-nearest* neighbor's algorithm (*k*-NN)** is a non parametric method used for classification and regression In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

- In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.

- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

K-NN is a type of instant-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification.

Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor

The neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity of the *k*-NN algorithm is that it is sensitive to the local structure of the data.

Bagging classifier:

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

Each base classifier is trained in parallel with a training set which is generated by randomly drawing, with replacement, N examples (or data) from the original training dataset – *where N is the size of the original training set*. Training set for each of the base classifiers is independent of each other. Many of the original data may be repeated in the resulting training set while others may be left out.

Vol 13, Issue 04, APRIL /2022
ISSN NO:0377-9254

Bagging reduces over fitting (variance) by averaging or voting, however, this leads to an increase in bias, which is compensated by the reduction in variance though.

AdaBoost:

Adaptive boosting is a machine learning meat algorithm formulated. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the over fitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

Every learning algorithm tends to suit some problem types better than others, and typically has many different parameters and configurations to adjust before it achieves optimal performance on a dataset, AdaBoost is often referred to as the best out-of-the-box classifier.[2] When used with decision tree learning, information gathered at each stage of the AdaBoost algorithm about the relative 'hardness' of each training sample is fed into the tree growing algorithm such that later trees tend to focus on harder-to-classify examples.

Multilayer perceptron (MLP):

A **multilayer perceptron**(MLP) is a class of feed forward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to refer to *any* feed forward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons (with threshold activation); see § Terminology. Multilayer perceptrons are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer.

An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

Artificial neuron network (ANN):
An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output.

ANNs are considered nonlinear statistical data modelling tools where the complex relationships between inputs and outputs are modelled or patterns are found.

ANN is also known as a neural network.

An ANN has several advantages but one of the most recognized of these is the fact that it can actually learn from observing data sets. In this way, ANN is used as a random function approximation tool. These types of tools help estimate the most cost-effective and ideal methods for arriving at solutions while defining computing functions or distributions. ANN takes data samples rather than entire data sets to arrive at solutions, which saves both time and money. ANNs are considered fairly simple mathematical models to enhance existing data analysis technologies.

ANNs have three layers that are interconnected. The first layer consists of input neurons. Those neurons send data on to the second layer, which in turn sends the output neurons to the third layer.

Training an artificial neural network involves choosing from allowed models for which there are several associated algorithms.

To implement above all algorithms we have used python technology and 'student data' dataset.This dataset available inside dataset folder which contains test dataset with dataset information file.

4.RESULTS AND DISCUSSION

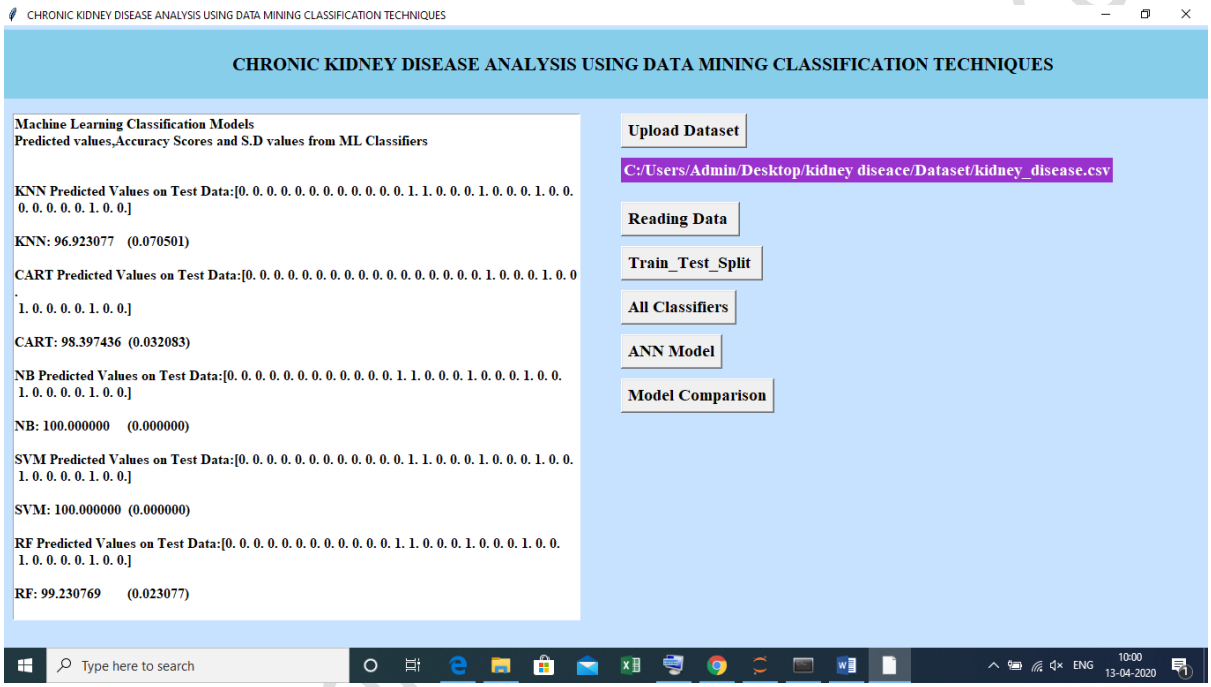

Fig 2:Accuracy Prediction

KNN Predicted Values on Test Data is 96.92%
CART Predicted Values on Test Data is 98.39%
SVM Predicted Values on Test Data is 100%
RF Predicted Values on Test Data is 100%
NB Predicted Values on Test Data is 97.41%
bagging Predicted Values on Test Data is 98.39%
MLP Predicted Values on Test Data is 100%
Now click on 'ANN Model'

Journal of Engineering Sciences
Vol 13, Issue 04, APRIL /2022
ISSN NO:0377-9254

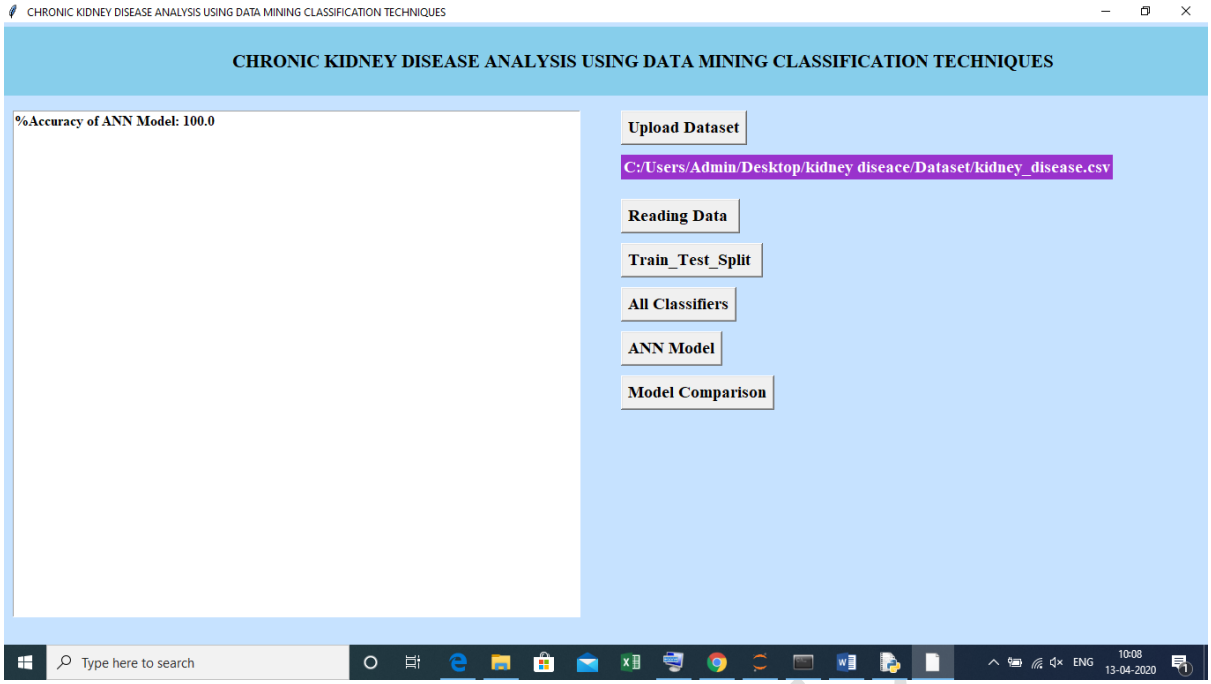

Fig 3:Accuracy of ANN

Now click on 'Model co
mparison' the comparison between the models

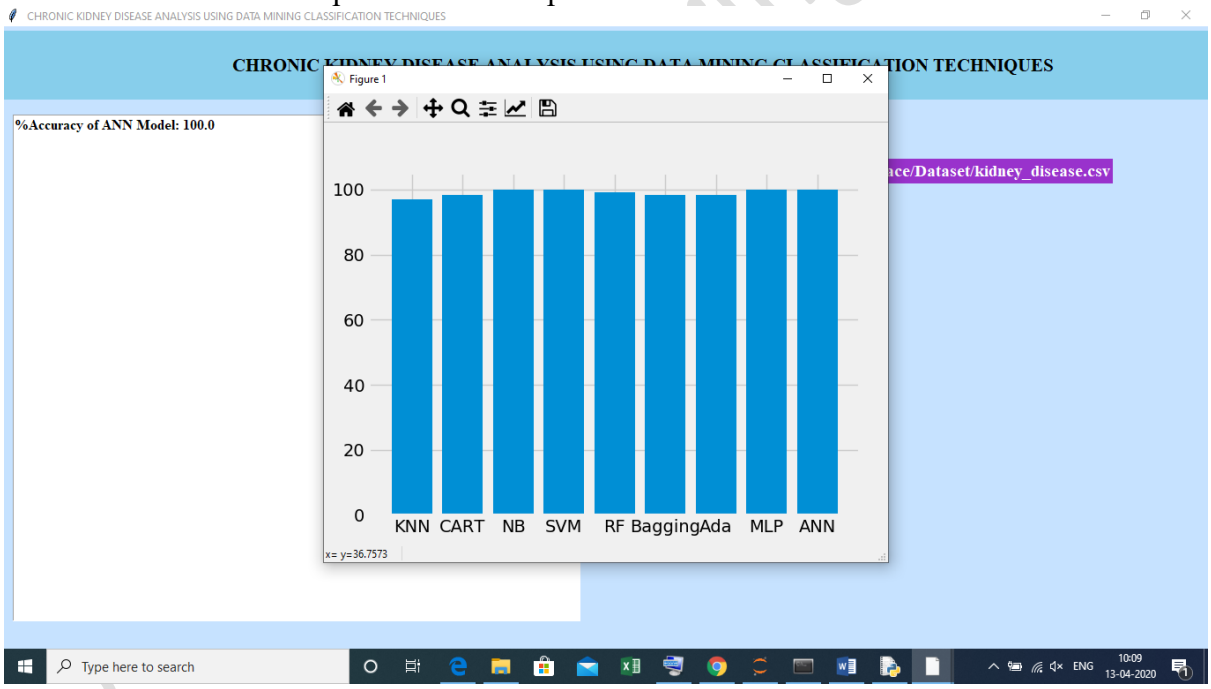

Fig 4:Prediction Graph

5.CONCLUSION

Prediction of chronic kidney disease is one of the essential topics in medical diagnosis. The proposed work is to classify the different stages of chronic kidney disease according to its severity. 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 techniques with neural 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. Mohammed Abdul Khaleel and Sateesh Kumar Pradham, "A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases", International Journal of Advanced Research in Computer Science and Software Engineering, 2013, Vol.3, No. 8, pp. 149-153.
3. Lambodar Jena and Narendra Ku. Kamila, "Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease", International Journal of Emerging Research in Management and Technology, 2015, Vol. 4, No. 11, pp. 110-118.
4. Abeer, Ahmad and Majid," Diagnosis and Classification of Chronic Renal Failure Utilising Intelligent Data Mining Classifiers", International Journal of Information Technology and Web Engineering, 2014, Vol.9, No.4, pp. 1-12.
5. Jerlin Rubini and Eswaran," Generating Comparative Analysis of Early Stage Prediction of Chronic Kidney Disease", International Journal of Modern Engineering Research, 2015, Vol. 5, No. 7, pp. 49-55.
6. Ruey Kei Chiu and Renee Yu-Jing," Constructing Models for Chronic Kidney Disease Detection and Risk Estimation", Proceedings of 22nd IEEE International Symposium on Intelligent Control, Singapore, 2011, pp. 166-171.
7. Jayalakshmi and Santhakumaran, "Improved Gradient Descent Back Propagation Neural Networks for Diagnoses of Type II Diabetes Mellitus", Global Journal of Computer Science and Technology, 2010, Vol. 9, No. 5, pp. 94-97.
8. Koushal Kumar and Abhishek, "Artificial Neural Networks for Diagnosis of Kidney Stones Disease", International Journal Information Technology and Computer Science, 2012, Vol. 7, No. 3, pp. 20-25.
9. Rajalakshmi, Neelamegam and Bharathi, "Diagnosis and Classification of Level of Kidney Function Using Associative Neural Network and Polynomial Neural Network", Research Journal of Pharmaceutical, Biological and Chemical Sciences, 2013, Vol. 4, No. 4, pp. 724-738.
10. Vijayarani and Dhayanand, "Kidney disease prediction using Support Vector Machine and Artificial Neural Network algorithms", International Journal of Computing and Business Research (IJCBR), 2015, Vol. 1, No. 3, pp.1765-1771.
11. Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2012, Vol. 2, No. 2, pp. 55-66.
12. Abhinandan Dubey, "A Classification of CKD Cases Using Multi Variate K-Means Clustering", International Journal of Scientific and Research Publications, 2015, Vol. 5, No. 8, pp. 1-5.
13. Yashpal Singh and Alok Singh Chauhan," Neural Networks in Data Mining", Journal of Theoretical and Applied Information Technology, 2005, Vol.5, No.6, pp. 37-42.
14. Baylor, Konukseven and Koku," Control of a Differentially Driven Mobile Robot Using Radial Basis Function Based Neural Networks", World Scientific and Engineering Academy and Society Transcations on Systems and Control, 2008, Vol. 3, No. 12, pp. 1002-1013.
15. Shubham Bind, Arvind Kumar Tiwari and Anil Kumar Sahani, " A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction", International Journal of Computer Science and Information Technologies, 2015,Vol. 6, No. 2, pp.1648-1655.