Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 4 (2016) pp. 175-184
© MIR Labs, www.mirlabs.net/jnic/index.html

Intelligent heart disease prediction system using random forest and evolutionary approach

M.A.Jabbar[1], B.L.Deekshatulu[2] and Priti Chandra[3]

[1] Associate Professor, MJCET, Hyderabad, India,
jabbar.meerja@gmail.com

[2] Distinguished fellow, IDRBT, RBI, Hyderabad,
deekshatulu@hotmail.com
[3] Senior Scientist, ASL, DRDO,Hyderabad
priti_murali@gmail.com

Abstract: **Heart disease is a leading cause of premature death in the world.Predicting the outcome of disease is the challenging task.Data mining is involved to automatically infer diagnostic rules and help specialists to make diagnosis process more reliable.Several data mining techniques are used by researchers to help health care professionals to predict the heart disease.Random forest is an ensemble and most accurate learning algorithm,suitable for medical applications.Chi square feature selection measure is used to evaluate between variables and determines whether they are correlated or not.In this paper ,we propose a classification model which uses random forest as classifier ,chi square and genetic algorithm as feature selection measures to predict heart disease. The experimental results have shown that our approach improve classification accuracy compared to other classification approaches,and the presented model can be successfully used by health care professional for predicting heart disease.**

Keywords: Heart disease, Random forest, Data mining, Feature selection, Chi square, Genetic algorithm
.

I. Introduction

Heart disease also called as coronary artery disease is a condition that affects the heart. Heart disease is a leading cause of death worldwide. Physicians generally make decisions by evaluating current test results of the patients. Previous decisions taken by other patients with the same conditions are also examined. So diagnosing heart disease requires experience and highly skilled physicians. Heart disease will become a leading cause of death by 2020. Heart disease diagnosis is an important yet complicated task. Today many hospitals collect patient data to manage health care of patients. This information is in different format like numbers, charts, text and images. But this database contains rich information but poorly used for clinical decision making [1]. The automation or decision support system would be extremely advantageous. Data mining can be used to automatically infer diagnostic rules and help specialists to make diagnosis process more reliable. The purpose of predictions in data mining is to discover trends in patient data in order to improve their health care [2].

Knowledge discovery in data bases is applied to extract useful patterns from the medical data sets using various data mining techniques. Data mining have shown a good result in prediction of heart disease and is widely applied for prediction of heart disease. Due to shortage of doctors and experts in medical field to predict heart disease, and because of neglecting the patient's symptoms, data mining is emerged as an analysis tool.

Random forest is an ensemble classifier which combines bagging and random selection of features. Random forest can handle data without preprocessing. Random forest algorithm has been used in prediction and probability estimation. Random forest consists of many decision trees and outputs the class, which is the mode of individual trees class [3].It is one of the most accurate classifier. It produces a highly accurate classification for many data sets especially for heart disease data set.

Feature selection is a process of identifying and removing redundant and irrelevant features and increasing accuracy. During the last decade, the motivation for applying the feature subset selection has been increased for model building .Feature subset selection methods are classified into four types.1) Embedded method 2) Wrapper method3) Filter method4) Hybrid method. Genetic algorithm is a randomized wrapper feature selection technique. Chi square test is a filter method used to determine the difference between expected frequency and observed frequency. Information gain and gain ratio are univariate filter based fast feature selection methods. These feature selection are independent of the classifier.

Major contributions of our paper are summarized as

1) We propose a new method which employs the random forest ensemble algorithm for prediction of heart disease.
2) Apply chi square and genetic algorithm to select best features.
3) Apply feature selection measures to improve the accuracy in predicting heart disease.

The rest of the paper is organized as follows .Section 2 presents related work. We will review various articles related to heart disease. Section 3 deals with literature review. Section 4 presents proposed approach. Experimental results are discussed in section 5.We will conclude in section 6.

II. Related work

In this section, we will review some articles related to heart disease.

Kemal polat et.al proposed hybrid method which uses fuzzy weighted preprocessing and artificial immune system [4].Their proposed medical decision making method consists of two phases. In the first phase fuzzy weighted preprocessing is applied to heart disease data set to weight the input data. Artificial immune system is applied to classify the weighted input. They applied their methodology on Cleveland heart disease data set which consists of 13 attributes. The method uses 10 fold cross validation.

Diagnosis of heart disease through neural network ensembles was proposed by Resul das et.al[5].Their method creates a new model by combining posterior probabilities from multiple predecessor models. They implemented the method with SAS base software on Cleveland heart disease data set and obtained 89.01% accuracy.

P.K.Anooj developed a clinical decision support system to predict heart disease using fuzzy weighted approach. The method consists of two phases. First phase consists of generation of weighted fuzzy rules, and in second phase fuzzy rule based decision support system is developed. Author used attribute selection and attribute weight method to generate fuzzy weighted rules. Experiments were carried out on UCI repository and obtained accuracy of 57.85% [6].

Robert Detrano et.al proposed probability algorithm for the diagnosis of coronary artery disease. The probabilities that resulted from the application of the Cleveland algorithm were compared with Bayesian algorithm. Their method obtained an accuracy of 77% [7].

Decision tree for diagnosing heart disease patients was proposed by Mai shouman et.al [8].Different types of decision trees are used for classification. The research involves data discretization, decision tree selection and reduced error pruning. Their method outperforms bagging and j48 decision tree. Their approach achieved 79.1% accuracy.

Diagnosis of heart disease through bagging approach was proposed by My chau Tu et.al [9].The proposed bagging algorithm is used to identify warning signs of heart disease. They made a comparison with decision tree. Their approach claimed an accuracy of 81.4%.

Andreeva used C4.5 decision tree for the diagnosis of heart disease. Feature extraction and specific rule inferring from heart disease data set is considered. Their proposed approach achieved an accuracy of 75.73% [10].

Diagnosis of CVD with Bayesian classifiers was proposed by Alaa Elsayad et.al [11].The researchers evaluated the performance of Bayesian classifier to predict the heart disease. Cleveland heart disease data set is used for their study. The model is implemented in SPSS work bench. Cleveland heart disease data set consists of 14 features. Their study evaluates two Bayesian network classifiers namely 1) tree augmented naïve bayes and 2) Markova blanket estimation (MBE).Classification accuracies are compared with SVM. The performance of classification model is evaluated using classification accuracy, specificity and sensitivity. MBE model achieved an accuracy of 97.92 where as TAN and SVM classifiers achieved an accuracy of 88.54 and 70.83 respectively.

Hlaudi Daniel Masethe et.al proposed prediction of heart disease using five different classifiers namely J48, Bayes, CART, Reptree and Bayes net[2].

Data set collected from south Africa is used for experimental analysis. Only 11 attributes are considered for modeling. Accuracy obtained by various algorithms are used as reliable indicators for prediction of heart disease.

Heart disease classification using nearest neighbor classifier with feature subset selection was proposed in[12].Their method achieved an accuracy of 97.5%
Feature analysis of coronary artery heart disease data set is proposed in [13].Their work is focused on integrating result of machine learning on different data sets targeting the coronary artery disease.

Heart disease prediction system using associative classification was proposed in [14].Authors proposed efficient associative classification algorithm for heart disease prediction using genetic algorithm. Their method uses gini index for class association rule generation. Gini index is used to improve classification accuracy as a informative attribute centered rule generation. Fitness of rule is evaluated using Z statistics. The experimental results showed that their approach achieved an accuracy of 88.9%.

M.A.Jabbar et.al [15] proposed a model for prediction of heart disease using random forest (RF) and feature subset selection. Authors proposed a new method which uses RF and feature subset selection chi square for disease prediction. Chi square metric is used to filter features in the data set. Cleveland data set is used for experimental analysis. Five metrics sensitivity, specificity, disease prevalence, negative predictive value and positive predictive value are used for analysis of classification model. Experimental results demonstrated in their approach that there is significant improvement in accuracy.

Early diagnosis of heart disease using computational intelligence techniques are proposed in [16]. Authors attempted to increase the accuracy of the naïve bayes classifier to classify heart disease data. They used a discretization method and genetic search to remove irrelevant features.

Genetic algorithm is used for optimization. Authors performed a comparison with other traditional algorithms.

Alternating decision trees for early diagnosis of heart disease was proposed by Jabbar et.al[17].Alternating decision tree is a new type of classification, which is a generalization of decision tree, voted decision trees and voted decision stumps. Principal component analysis is used as a feature selection measure and used to select best features .Heart disease data consists of 96 patient's records with 10 features. Their proposed approach achieved an accuracy of 91.66%.

M.A.Jabbar et,al proposed cluster based association rule mining for heart attack prediction[18].Authors proposed a model to analyze medical data set using association rule mining. Cleveland data set is used for experimental analysis. Medical data set is divided into partitions of equal size based on skipping fragments. Their approach reduces main memory requirement and is efficient in pruning heart disease prediction rules.

III. Literature review

This section reviews literature used in this paper.

A. Heart Disease

Heart disease also called as coronary heart disease (CHD), is a deposition of fats inside the tubes which supplies blood to the heart muscles. Heart disease actually starts as early as 18 years and patients only came to know about heart disease when the blockage exceeds about 70%.Theses blockages develop over the years and lead to rupture of the membrane covering the blockage due to pressure increases. If the chemicals released by broken membrane mixed with blood and lead to a blood clot, results to heart disease [19].

The reasons which increase blockage are called as risk factors. These risk factors are classified as modifiable and non modifiable risk factors. Non modifiable risk factors are age, gender, and heredity. These risk factors can't be modified and they will always keep causing heart disease.
Risk factors which can be changed by our efforts are called as modifiable risk factors. Some modifiable risk factors are 1) Food related 2) Habit related 3) Stress related 4) Bio chemical and miscellaneous risk factors. Atherosclerosis, coronary, congential, rheumatic, myocarditis, arrhymia and angina are the different types of heart diseases[20].Common symptoms of heart disease are listed in table 1[21].

Table 1: Symptoms of heart disease

Sl.no	Symptoms name
1	Chest pain
2	Strong compressing or flaming in the chest
3	Discomfort in chest area
4	Sweating
5	Light headedness
6	Dizziness
7	Shortness of breath
8	Pain spanning from the chest to arm and neck
9	Cough
10	Fluid retention

Major risk factors of Coronary heart disease are listed in table 2

Table 2: Risk factors of heart disease[22]

Sl.no	Risk factor
1	Diabetes
2	High blood pressure
3	High LDL
4	Low HDL
5	Not getting enough physical activity
6	Obesity
7	Smoking

Effective decision support system should be developed to help in tackling the menace of heart disease.

B. Random forest (RF)

Random forest algorithm is one of the most effective ensemble classification approach. The RF algorithm has been used in prediction and probability estimation.RF consists of many decision trees .Each decision tree gives a vote that indicate the decision about class of the object. Random forest item was first proposed by Tin kam HO of bell labs in 1995.
RF method combines bagging and random selection of features. There are three important tuning parameters in random forest1) No. of trees (n tree) 2) Minimum node size 3) No. of features employed in splitting each node 3) No. of features employed in splitting each node for each tree (m try). Random forest algorithm advantages are listed below.

1) Random forest algorithm is accurate ensemble learning algorithm.
2) Random forest runs efficiently for large data sets.
3) It can handle hundreds of input variables.
4) Random forest estimates which variables are important in classification.
5) It can handle missing data.
6) Random forest has methods for balancing error for class unbalanced data sets.
7) Generated forests in this method can be saved for future reference [23].

8) Random forest overcomes the problem over fitting.
9) In training data, RF is less sensitive to outlier.
10) In RF, parameters can be set easily and eliminates the need for tree pruning.
11) In RF accuracy and variable importance is automatically generated [24].

When constructing individual trees in random forest, randomization is applied to select the best node to split on. This value is equal to √A, where A is no. of attributes in the data set [25].However RF will generate many noisy trees, which affect classification accuracy and wrong decision for new sample. [25]

Following algorithm illustrates random forest method.

Algorithm Random forest

```
Step 1: From the training set, select a new bootstrap sample.
Step 2: Grow on a un pruned tree on this bootstrap sample.
Step 3: Randomly select (m try) at each internal node and determine best split.
Step 4: if each tree is fully grown. Do not perform pruning.
Step 5: Output overall prediction as the majority vote from all the trees.
```

B. Chi-square method

Feature selection is a preprocessing technique used to remove irrelevant and redundant features. Medical data is high volume in nature and consists of redundant features. Medical diagnosis is a complicated task, needs to be executed accurately and efficiently. Feature selection if applied on medical data set will give accurate results. In this paper, we consider chi square and genetic search as feature selection and ranking methods, which show good performance in various domains.

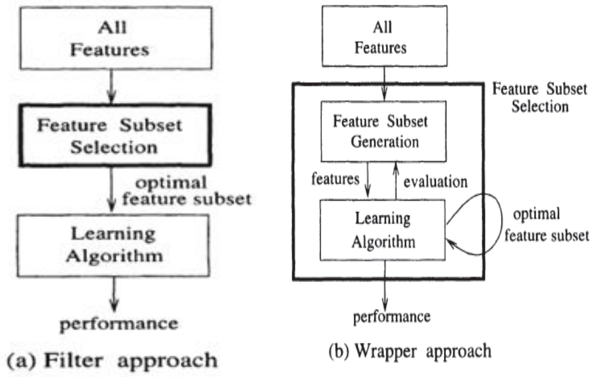

Figure 1 : Filter and wrapper feature subset selection measures

Chi square is a statistical test that is used to measure divergence from the distribution of feature occurrence which is independent of the class value [26].Chi square requires the following conditions to be satisfied.

1) Data must be quantitative
2) One or more categories of data required
3) Independent observations
4) Sample size should be adequate and simple
5) Data must be in frequency form
6) All observations must be read.

Chi square formula is represented as

$$X^2 = \sum \frac{(o-e)^2}{e}$$

Where O is observed frequency and e is expected frequency. Following example illustrates chi square hypothesis.

Example: A six sided die is thrown 264 times. Results are shown in the table. We want to know if the die is biased

[let $\chi^2_{0.05}$=11.07 for 5d Degree of freedom (df)]

Number appeared on the die	*1*	*2*	*3*	*4*	*5*	*6*
Frequency	40	32	28	58	54	52

Solution

Degrees of freedom can be calculated as the number of categories in the problem minus 1.Null hypothesis H0: The die is unbiased.

Expected frequency of each of the numbers =264/6=44

Observed frequency(O)	Expected Frequency(e)	$(O-e)^2$	$(O-e)^2/e$
40	44	16	0.3636
32	44	144	3.2727
28	44	256	5.8181
58	44	196	4.4545
54	44	100	2.2727
52	44	64	1.4545
χ^2=17.636			17.636

The number of degree of freedom (df) =n-1=5

The tabulated value of χ^2 for degree of freedom (df=5) at 5% level=11.07. Since calculated χ^2 is greater than tabulated χ^2 we reject the null hypothesis H0.i.e. We reject the hypothesis that the die is unbiased. Hence the die is biased.

Table 3 shows example data set weather data. This data set consist of 14 instances and 5 features. The last feature is class.

Table 3: Weather data set

No	Outlook	Temperature	Humidity	Windy	Play
1	SUNNY	HOT	HIGH	F	No
2	RAINY	MILD	NORMAL	F	yes
3	SUNNY	MILD	NORMAL	T	yes
4	OVER CAST	MILD	HIGH	T	yes
5	OVER CAST	HOT	NORMAL	F	yes
6	RAINY	MILD	HIGH	T	no
7	SUNNY	HOT	HIGH	T	no
8	OVER CAST	HOT	HIGH	F	yes
9	RAINY	MILD	HIGH	F	yes
10	RAINY	COOL	NORMAL	F	yes
11	RAINY	COOL	NORMAL	T	no
12	OVER CAST	COOL	NORMAL	T	yes
13	SUNNY	MILD	HIGH	F	no
14	SUNNY	COOL	NORMAL	F	yes

Ranking of the attributes for weather data set is based on chi is shown in table 4

Table 4: Ranking of attributes based on chi square

Rank	Name of the attribute	Chi square value
1	outlook	3.547
2	humidity	2.8
3	windy	0.933
4	temperature	0.57

C. Genetic Algorithm(GA)

Genetic algorithm (GA)represents general purpose search method based on natural selection and genetics.GA stimulate natural process based on law mark and Darwin principles[27].GA are implemented using computer simulation for optimization.GA are useful for searching very general spaces depending on some probability values for optimization[14][28].Each solution generated in GA is called a chromosome. Genetic algorithms have played a major role in many applications of the engineering science.

Driving force behind genetic algorithm is the use of three operators namely

1) Selection:

Selection operator is used to give preference to better chromosomes using objective function

2) Crossover:

Crossover operator takes more than one parent chromosome and produces a child from them.

3) Mutation:

Mutation operator is used to maintain diversity and to inhibit premature convergence. In mutation, a portion of the new individual bits are flipped.

Pseudo Code of Genetic Algorithm

```
Step 1: Randomly initialize population
Step 2: Compute fitness of population
Step 3: Repeat
Step 4: Select parents from population
Step 5: Perform crossover
Step 6: Perform mutation
Step 7: Compute fitness
Step 8: Until best individuals are selected and stop
```

Flow chart of Genetic algorithm is shown in figure 2

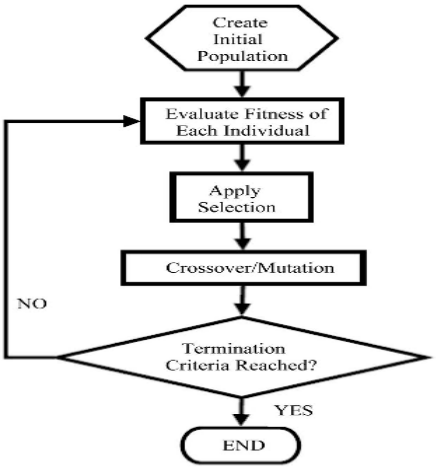
Create Initial Population
Evaluate Fitness of Each Individual
Apply Selection
Crossover/Mutation
NO
Termination Criteria Reached?
YES
END

Figure 2: Flow chart of genetic algorithm

IV. Proposed method

The literature survey presents various techniques for prediction of heart disease. Each method has its own advantages and their short comings. The proposed technique uses random forest algorithm for prediction of heart disease. Feature subset selection is a process that selects a subset of original attributes and reduces feature space [29].

We applied, Random forest with chi square and GA as feature selection measures on heart disease data set collected from various corporate hospitals in Hyderabad (Heart disease data set T.S) and also on heart stalog data set.

In our proposed work ,we used chi square and GA to select attributes and keep only attributes which contribute more towards the diagnosis of heart disease.

Confusion matrix is a table used to visualize the performance of an algorithm. Confusion matrix(Table 5) has two rows and two columns (for two class problems) that specify TP, FP, TN, FN.

Confusion matrix is used to compare actual classification of heart disease data set, with number of correct and incorrect predictions made by the model. The traditional classification matrix is shown below.

To evaluate the performance of our proposed model, we used following classification measures [30].

Table 5: Confusion Matrix

Prediction		Disease	
		+	-
	+	True positive	False positive
		TP	FP
	-	False Negative	True Negative
		FN	TN

1) Specificity=TN/ (FP+TN)
2) Sensitivity =TP/ (TP+FN)
3) Disease prevalence= (TP+FN)/ (TP+FP+TN+FN)
4) Positive predictive value(PPV): TP/ (TP+FP)
5) Negative Predictive value(NPV): TN/ (FN+TN)
6) Accuracy= (TP+TN)/ (TP+FP+TN+FN)

Where TP=> Positive tuples that are correctly labeled by the classifier.
TN=> Negative tuples that are correctly labeled by classifier.
FN=> Positive tuples that are incorrectly labeled by classifier.
FP=> Negative tuples that are incorrectly labeled by classifier.

Positive predictive value (PPV) is defined as probability that the heart disease is present when the diagnosis test is positive. Positive predictive value (NPV) is defined as probability that the heart disease is absent when the diagnosis test is negative [16].

Attributes for our heart disease data set T.S are listed in Table 6 and heart stalog attributes are shown in table 7.

Table 6: Heart disease data set attributes

1	Age	Numeric
2	Gender	Nominal
3	BP	Numeric
4	Diabetic	Nominal
5	Height	Numeric
6	Weight	Numeric
7	BMI	Numeric
8	Hypertension	Nominal
9	Rural	Nominal
10	Urban	Nominal
11	Disease class	Nominal

Table 7: Attributes of heart stalog data set

Sl.no	Attributes of heart disease
1	Age
2	Sex
3	Chest pain
4	Resting blood pressure
5	Serum cholestoral
6	Fasting blood sugar
7	Resting electro graphic results
8	Max.heart rate achieved
9	Exercise induced angina
10	Old peak
11	Slope
12	No.of major vessels
13	Thal
14	Class

Proposed algorithm:

Step 1: Load the heart disease data set

Step 2) Rank the features in descending order based on chi square and GA value. A high value of chi square indicates feature is more related to class.
Apply backward elimination algorithm .Back ward elimination algorithm starts from the full feature set, and iteratively removes one by one feature with low value.

In each iteration only one feature is removed, which mostly affects overall model accuracy, as long as the accuracy stops increasing. Least rank feature will be pruned. Chi square and GA is used to select high ranked features.

Step 4: Select the features with highest value.

Step 3) Apply Random forest algorithm on the remaining features of the data set that maximizes the classification accuracy.

Steps 4) Find the accuracy of the classifier.

Steps 1 to 4 deals with feature selection. High ranked features are selected for classification. From Step 3 to 4, RF classification will be applied to the selected feature subset. After applying classification, accuracy of the classifier will be calculated.

V. Experimental results

We carried out experiments using Hold out and Cross validation approach. In Hold out approach, we partitioned samples into two independent data set.75% of data set is used to train the classifier and to build the classifier. Remaining 25% data set is used for testing.

In 10-fold cross validation all the instances of the data set are used and are divided into 10 disjoint groups, where nine groups are used for training and the remaining are used for testing. The algorithm runs for 10 times and average accuracy of all folds is calculated.

To evaluate the performance of our approach, we used the measures listed in section 4.Accuracy comparison of Heart Disease data set-Cleveland [31] is shown in Table 8 and figure 3. Naïve bayes approach obtained an accuracy of 78.56% ,whereas decision table obtained an accuracy of 82.43%.The results are obtained using 10 cross validation.

Our approach obtained 7.97% improvement over C4.5 algorithm. Accuracy comparison for Heart Disease data set T.S is compared with Decision tree (DT) is shown in Table 9 and figure 4.Our approach obtained 100% accuracy, where as DT obtained an accuracy of 98.66%.Comparision of various parameters for heart disease data set T.S is listed Table 10 and Figure 5.

Specificity shows that the probability of testing the result of heart disease will be negative when the heart disease is not present. Positive predictive value (PPV) is the probability that the heart disease is present when the diagnosis test is positive.PPV value for DT is 98.39% where as our approach records 100%.Negative predictive value (NPV) recorded by our approach is 90% and positive predictive value is 75.8 for heart stalog data set which is shown in Table 11 and figure 6.

Clinically, the disease prevalence(DP) is the same as the probability of disease being present before the test is performed (prior probability of disease).

The above experimental results suggests that our proposed approach efficiently achieve high degree of dimensionality reduction and improve accuracy with predominate features. Overall our approach outperforms other approaches. This indirectly helps patient's no. of diagnosis tests to be taken for prediction of heart disease.

Table 8: Accuracy comparison for Heart Disease Data set (Cleveland data set)

Sl.no	*Approach*	*Accuracy*
1	PART C4.5	75.73
2	Naïve bayes	78.56
3	Decision table	82.43
4	Neural nets	82.77
5	**Our approach**	**83.70**

Table 9: Accuracy comparison for Heart Disease Data set (TS Data set)

Sl.no	*Approach*	*Accuracy*
1	Decision Tres(DT)	98.66
2	Our approach	100

Table 10: Comparison of various parameters for heart disease data set T.S

Sl.no	*Parameter*	*Our approach*	*Decision tree*
1	Sensitivity	100	100
2	Specificity	100	92.86
3	Disease prevalence	82.67	81.33
4	Positive Predictive Value(PPV)	100	98.39
5	Negative Predictive Value(NPV)	100	100

Table 11: Comparison of various parameters for heart disease data set –HEART STALOG

Sl.no	*Parameter*	*Our approach*	*Decision tree*
1	Sensitivity	85.8	80.18
2	Specificity	82.3	80.50
3	Disease prevalence	39.2	41.11
4	Positive Predictive Value(PPV)	75.8	74.17
5	Negative Predictive Value(NPV)	90.0	85.33

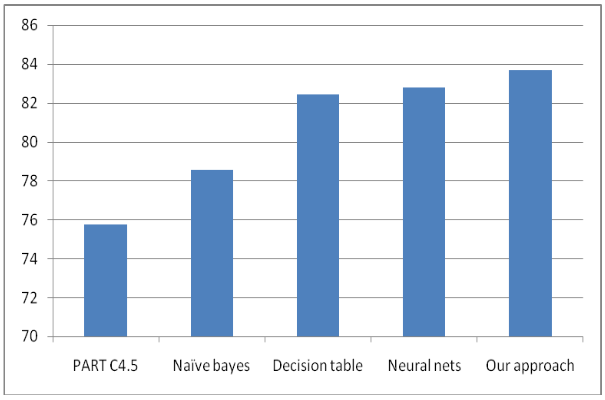

Figure 3 :Accuracy comparision of heart stalog data set by various approaches

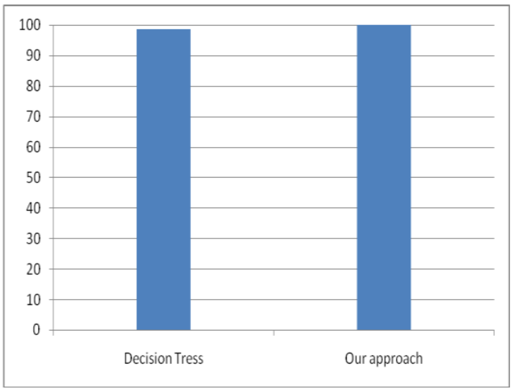

Figure 4 :Accuracy comparision of Heart disease data set-T.S

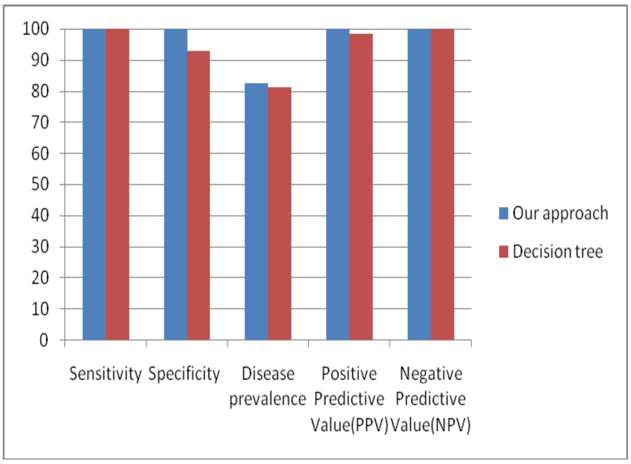

Figure 5 :Comparision of various parameters for Heart disease data set-T.S

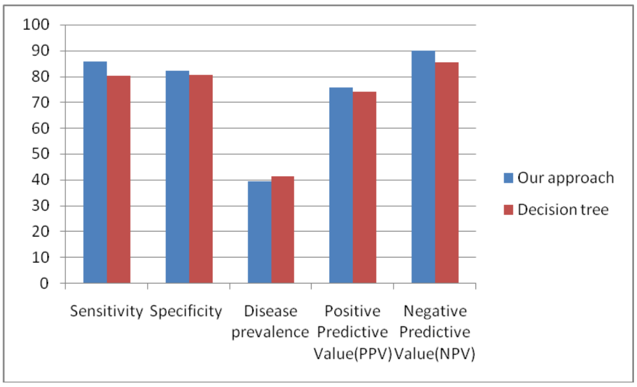

Figure 6: Comparision of various parameters for Heart stalog data set

Specificity of heart disease data set T.S obtaned by our approach is 100% where as it is 92.86% by decesion tree.Our approach obtained 1.8% improvement over decesion tree for heart stalog data set,which is shown in figure 5 and 6.

Table 12: Accuracy of heart disease using RF and GA

No. of trees in RF	*Before GA as feature selection (RF only)*	*After applying GA (RF+GA)*
50	80	84
10	80	82.96
100	80	82.96

Table 13: Parameters of GA

Sl.no	***Parameter name***	***Threshold value***
1	Crossover	0.6
2	Mutation	0.03
3	Population size	20
4	Max.Generation	20

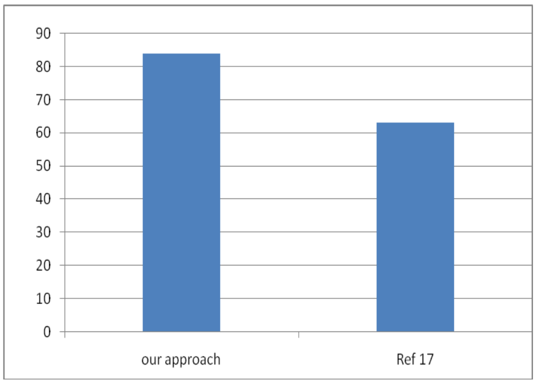

Figure 7 : Accuracy comparision

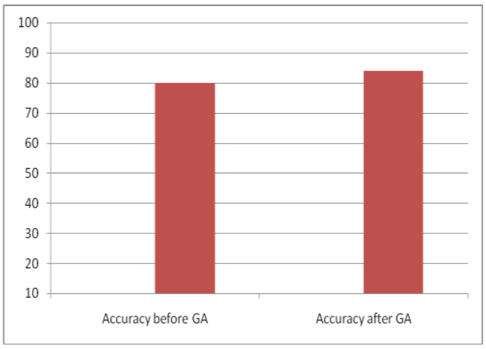

Figure 8 :Accuracy comparision before and after GA

Table 12 shows accuracy obtained using RF and GA .RF+GA model improved 4%accuracy than RF with out GA.We tested accuracy for various number of trees in RF.Table 13 shows various parameters of GA.Population size is limited to 20 and maximum generations are limited to 20.Crossover and mutaion probabilities are set to 0.6 and 0.03 respectively.
Table 14 and figure 9 shows the acuracy comparision for heart disease data by various methods.From the table it is clearly evident that our approach outperforms other models developed by researchers.

Table 14 :Accuracy comparision for heart disease data set

Name of the author/approach	***Accuracy***
Chen	80
Chaurasia	83.49
Abdullah	63.3
Decision tree	63.3
Anooj	57.85
Mi chau tu	81.4
Andreeva	75.73
Our Approach(RF+GA)	84

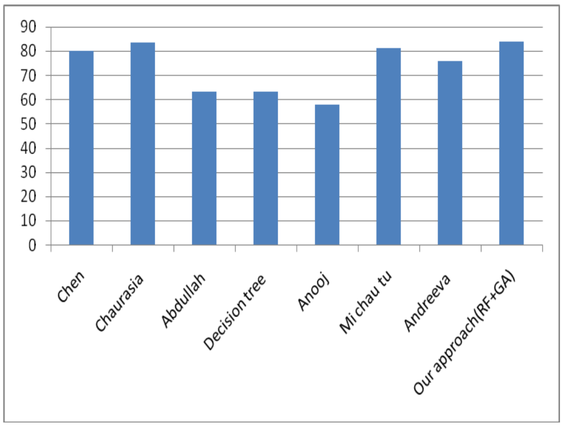

Figure 9 : Accuracy comparision for heart disesae data set by various approaches

VI. Conclusion

In this research paper, we developed efficient approach for prediction of heart disease using Random forest. Data mining plays an important role in the prediction of heart disease. We adopted feature selection using chi square and Genetic algorithm measures for heart disease classification.

Our proposed approach (Random forest and Chi square) achieved an accuracy of 83.70% for heart stalog data set. Applying Random forest has shown improved accuracy in prediction of heart disease. This research systematically tested using 10 fold cross validation to identify most accurate method. We compared our approach with other traditional classification algorithms.

Our approach outperforms traditional classification algorithms for effective classification of heart disease. This type of research can be successfully used in predicting the risk factors of heart disease and to help health care professionals for prediction of heart disease.

References

[1] P.Shrama,k.saxena,"heart disease prediction system evaluation using c4.5 rules and partial trees "AISC,Springer,pp 285-294(2016)

[2] Hlaudi Daniel Masethe,"Prediction of heart disease using classification algorithms", vol 11,pp1-4,WCECS2014

[3]Sheik abdullah,RR Rajalakshmi,"A data mining model for predicting the coronary heart disease using random forest classifier",IJCA,PP 22-25(2012)

[4] Kemal polat, S.Gunes, S.Tosun, " Diagnosis of heart disease using artificial immune recognition system and fuzzy weight preprocessing", pattern recognition, 39, pp2186-193(2006)

[5] Resul das ,Turkoglu,A Sengur," Effective diagnosis of heart disease through network ensembles", Expert System with Applications36,pp7675-7680(2009)

[6]PK Anooj," Clinical decision support system: Risk level prediction of heart disease using Weighted fuzzy rules", Journal of king saud university, CIS, 24, PP 27-40(2012)

[7] Detrano ,Janosi,W Stein burn,et.al," International application of new probability algorithm for the diagnosis of CAD". The American Journal of Cardiology, pp 304-310,64(5),(1989)

[8] Mai Shouman, Turner, Stocker," Using decision tree for diagnosing heart disease patients", In 9th Australian data mining conference, Australia vol 121,ACM(2011)

[9] Tu et.al," Effective diagnosis of heart disease through bagging approach" Biomedical Engineering and approach, pp 1-4, BMEI2009, IEEE (2009)

[10] Andreeva ," Data modeling and specific rule generation via data mining techniques", International conference on computer system and technologies" Comsystech 2006, pp 1- 6(2006)

[11] Alaa Elsayad,Mahmoud Fakhr,"Diagnosis of cardiovasular diseases with bayesian classifier",Journal of Computer Science,vol 11(2),pp274-282(2015)

[12] M.A.Jabbar,B L Deekshatulu,Priti chandra,"heart disease classification using nearest neighbor classifier with feature subset selection"annals computer science series , 11th tome,1st fasc,pp 47-54(2013)

[13]Randa El-Bay,"Feature analysis of coronary artery heart disease data sets",Procedia Computer science,Elsevier,vol 65,pp 459-469(2015)

[14]M.A.Jabbar,B L Deekshatulu,Priti chandra,"Heart disease prediction system using associative classification and genetic algorithm",ICECIT 2012,VOL 1,PP 183-192(2012)

[15] M.A.Jabbar,B L Deekshatulu,Priti chandra,"Prediction of heart disease using random forest and feature subset selection",Springer,AISC,IBICA 2015,pp 187-196(2015)

[16] M.A.Jabbar,B L Deekshatulu, Priti chandra, "Computaional intelligence technique for early diagnosis of heart disease'IEEE,ICETECH 2015,pp 1-6(2015)

[17] M.A.Jabbar,B L Deekshatulu,Priti chandra,“Alternating decision tree for early diagnosis of heart disease “IEEE,I4C2014,pp 322-328(2014)

[18] M.A.Jabbar,B L Deekshatulu,Priti chandra, “Cluster based association rule mining for heart disease prediction “, JATIT,Vol 32,No 2,pp 1-8(2011)

[19] Saaol times, Monthly magazine” Modifiable risk factors of heart disease”, pp 6-10, July (2015)

[20] Khan MG,“Heart disease diagnosis and therapy“, a practical approach,2nd Edition Springer,pp544(2015)

[21]Khan MG,“Heart disease diagnosis and therapy“, a practical approach,2nd Edition Springer,pp544(2015)

[22] M.A.Jabbar,B L Deekshatulu,Priti chandra ,”classification of heart disease using artificial neural network and feature subset selection”,GJCST,Vol13, issue 3,2013

[23**]** home.etf.rs/~vm/os/dmsw/Random%20Forest.pptx,last accessed 10/8/2015

[24] Jehad Ali et.al,“Random forest and decision trees“,IJCSI,Vol 9,No 3,pp272-278(2012)

[25] kahled fawagreh,mohamded medhat gaber,Eyad Elyan,“Random forest:freom early developments to recent advancements“,systems science and control engineering,2:1, pp602-609(2014)

[26] George forman,“An extensive empirical study of feature selection metrics for text classification“,Journal of Machine Learning Research 3,pp 1289-1305(2003)

[27] M.A.Jabbar,B L Deekshatulu,Priti chandra,“An evolutionary algorithm for heart disease prediction“,ICICP 2012,CCIS292,Springer,PP378-389(2012)

[28] M.A.Jabbar,B L Deekshatulu,Priti chandra,“prediction of risk scores for heart disease using associate classification and hybrid feature subset selection“,IEEE ,ISDA,pp 628-634(2012)

[29]Preecha sonwang et.al,“Computer network security based on SVM approach“,In 11th Intnl.conf on aontrol,automation,and systems.

[30] Med Calc, “www.medcalc.org” last accessed on (5/8/2015)

[31] UCI machine learning repository,” archive.ics.uci.edu/ml” Last accessed 15/08/2015

Author Biographies

Dr.M.A.Jabbar born in Telangana state, India. He obtained his B.E in computer science engineering from MGMCE, Nanded and M.Tech from JNTUH, Hyderabad. He obtained his Ph.D from JNTUH in data mining. He published more than 25 papers in international Journals and conferences. He is technical Committee member for many international Conferences . He is execom member in EEE computer society India council. Presently he is working as an associate professor in MJCET,Hyderabad.His research interests include data mining, Attack graphs,IDS, Big data,IOT.

Dr.BL Deekshatulu did his BSc (Electrical Engg) from BHU (1958) and ME (1960) and PhD (1964) from Indian Institute of Science (IISc), Bangalore. Dr Deekshatulu contributed in the areas of linear and non-linear systems, digital image processing and remote sensing (data processing and applications). At IISc, Deekshatulu developed grayscale and colour drum scanners besides a flying spot scanner for image digitization, introduced ME Course in Servo Mechanisms, image/photo-processing labs and initiated School of Automation. At NRSA, he developed wide spread applications of remote sensing, commercial version of the large format drum scanner, image analysis equipment, etc. He was the Chairman, Remote Sensing Application Missions (ISRO), IGBP and SCOPE. Deekshatulu is a recipient of number of awards that include Sir M Visveswaraya Award (1984), NRDC Invention Awards (1986, 1993), Dr Biren Roy Space Science Award (1988), Padma Shri (1991) and LTC award by INAE.

Dr.Priti Chandra obtained her Ph.D in artificial intelligence from HCU, Hyderabad..She published more than 40 papers in international journals and conferences. Presently she is working as senior scientist, ASL, DRDO, Hyderabad. Her research interest includes data mining, Artificial intelligence, optimization techniques, fault tolerant systems. She is a member in IEEE.