European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 4, 2020

AN EFFICIENT HYBRID CLASSIFICATION ALGORITHM FOR HEART PREDICTION IN DATA MININIG

Mr. N. Rajesh Pandian[1], Dr. M. Balasubramani[2]

[1]*Designation: Assistant Professor, email id: nrajeshpandian@psnacet.edu.in*

[2]*Designation: Assistant Professor, email id: manibala@psnacet.edu.in*

Abstract: Heart sickness is the major critical reasons of mortality in the globe today. The gauge of coronary illness is a most basic test in the clinical information investigation zone. A few strategies are proposed to discover the result of ailment at prior stage which is as yet getting looked at. The Data mining is generally used to separate the critical, significant and wanted information from the patient's datasets. The few characterization techniques are utilized in the customary strategies for the coronary illness forecast in which the information mining ascribes are taken care of it. In this paper, to obtain a optimal result and also for the prediction of heart disease in the earlier stage, the novel hybrid FA-KNN are proposed. The proposed FA-KNN is the hybrid combination of machine learning algorithm of K-nearest neighbour and the optimization method of Firefly Algorithm which has been analyzed that produces the best result for all types of health related datasets. Therefore, the outcome of FA-KNN provides the optimal results in terms of precision, accuracy and recall with the evaluation of conventional methods.

I. INTRODUCTION

One of the noticeable diseases that influence numerous individuals during mature age is coronary illness, and as a rule it in the long run prompts fatal confusions. Heart problems are more common in men than in ladies. As per measurements from WHO, it has been evaluated that 24% of passing's expected to non-communicable infections in India are achieved by heart hardships. Along these lines 33% of all worldwide passing’s are because of heart issues. Half of the peoples in the worlds are suffered because of heart diseases. Around 17 million individuals died because of cardiovascular illness (CVD) consistently around the world, and the disease is profoundly prevalent in Asia. The cardiovascular Heart Disease Database (CHD) is viewed as the true information base for coronary illness research.

The heart is the basic bit of human's body. Life is itself subject to amazing working of the heart. In the event that task of the heart isn't genuine; it will affect the other body segments of human, for example, cerebrum, kidney, and so forth. Coronary affliction is a problem that ramifications for the development of the heart. There is different segments which makes danger of Heart difficulty. It is difficult to recognize coronary ailment because of a couple of contributory risk factors, for instance, diabetes, hypertension, raised cholesterol, sporadic heartbeat rate and various factors.

A couple of methodologies in information mining and neural associations have been used to find the reality of coronary ailment among individuals. The seriousness of the sickness is grouped dependent on different techniques like Decision Trees (DT) [2], K-Nearest Neighbor Algorithm (KNN) [1], Naive Bayes (NB) and Genetic calculation (GA). The idea of coronary illness is unpredictable and thus, the ailment must be taken care of cautiously. Not doing so may influence the heart or cause unexpected passing. The perspective of clinical science and information digging are used for finding various types of metabolic

conditions. Information mining with gathering expects a tremendous part in the gauge of coronary disease and data assessment.

In this paper, the common process of data mining is carried out with the hybrid algorithm of machine learning and optimization techniques (i.e)., K-Nearest Neighbour algorithm and the Firefly Optimization method for the improvement of heart disease identification in the early stage. This approaches are highly concentrate on the Recall, precision and Accuracy of the disease prediction that are compared with the traditional methods.

The rest of the paper is ordered as follows, in Section II, the heart disease related works, existing approaches and techniques are explained. In Section III, the propose methodology and the workflow are discussed. Section IV discusses about the proposed method results and comparison results of proposed model with the traditional methods. Finally, Section V illustrated with a conclusion of current work and some ideas on future enhancement.

II. RELATED WORKS

Kiran Jyoti et al (2012) proposed an artificial neural network (ANN) based heart disease prediction system using fifteen attributes. In proposed ANN integrate an additional pair of attributes in order to improve the precision.

Mohammad et al (2012) have developed method for the breast cancer notwithstanding cardiovascular sickness prediction utilizing information handling methods. The C4.5 algorithm utilized to predict severity in earlier stage.

Ishtake et al (2013) have implemented decision tree based approach for a coronary illness prediction .It combines the features of Neural Network(NN), Decision trees and Naïve Bayes.

Shantakumar et al (2009) introduced a methodology for removing essential examples from the heart affliction information for the effective expectation of attack. The preprocessed cardiovascular illness data stockroom was bunched to remove data generally applicable to assault abuse K-implies cluster algorithmic guideline. The continuous things are mined adequately abuse MAFIA algorithmic standard. In view of the determined significant weightage, the continuous examples having worth bigger than a predefined edge were picked for the valuable forecast of attack.

Niti Guru et al (2007) proposed a framework that utilizes a NN for forecast of cardiovascular illness, pulse, and sugar. An assortment of 78 records with thirteen credits are utilized for preparing and testing.

Bharti et al (2015) presented a few methods that assume an essential job while distinguishing or predicting coronary illness namely PSO, hereditary calculation, Artificial Neural Network, and so on.). The most importantly the major thoughts of these three methods and assessed how they help to foresee heart illnesses.

Mamatha Alex et al (2019) introduced a different grouping strategies in coronary illness finding. It can proceed as dependably as in the analysis of coronary illness. Early identification of coronary illness is the essential advantages of this investigation. It tends to be finding effectively on schedule, giving therapy sensible cost.

Anjan Nikhil Repaka et al (2018) have proposed on coronary illness expectation approach utilizing NB (Naive Bayesian) grouping and AES (Advanced Encryption Standard) calculation. Usage results show that , the Naive Bayes by yielding an exactness of 89.77% contrasted with other methodology.

Senthilkumar Mohan et al (2019) have proposed another strategy for finding huge highlights by utilizing AI procedures. For coronary illness information base proposed cross breed arbitrary woodland with a direct model (HRFLM) accomplishes 88.7% exactness contrasted with different strategies.

Salam Ismaeel et al (2018) have proposed an Extreme Learning Machine (ELM) calculation model to foresee coronary illness by thinking about age, sex, serum cholesterol, glucose.

Yukti Sharma et al (2019) have examined the utilization of the conspicuous highlights of two information mining methods, to be specific, K-Means Clustering and Decision Tree in coronary illness forecast. The proposed classifier give best exactness and execution to the Heart illness recognizable proof framework.

III. PROPOSED METHODOLOGY

In this section, the proposed methodology is explained with its work flow which is discussed with the block diagrams and the algorithm of every process. The work flow diagram of proposed is shown in the figure 1 which initially set the heart disease classification dataset of the patients in UCI repository. It gives a simple to-utilize visual representation of the dataset, working condition and building the predictive analytics. The Machine Learning measure begins from a pre-processing data stage followed by KNN feature selection based on DT entropy, classification of modelling execution is done by using the Firefly optimization techniques, and the outcomes of proposed hybrid FA-KNN results with improved precision. The feature selection and modelling continue for different combinations of attributes.

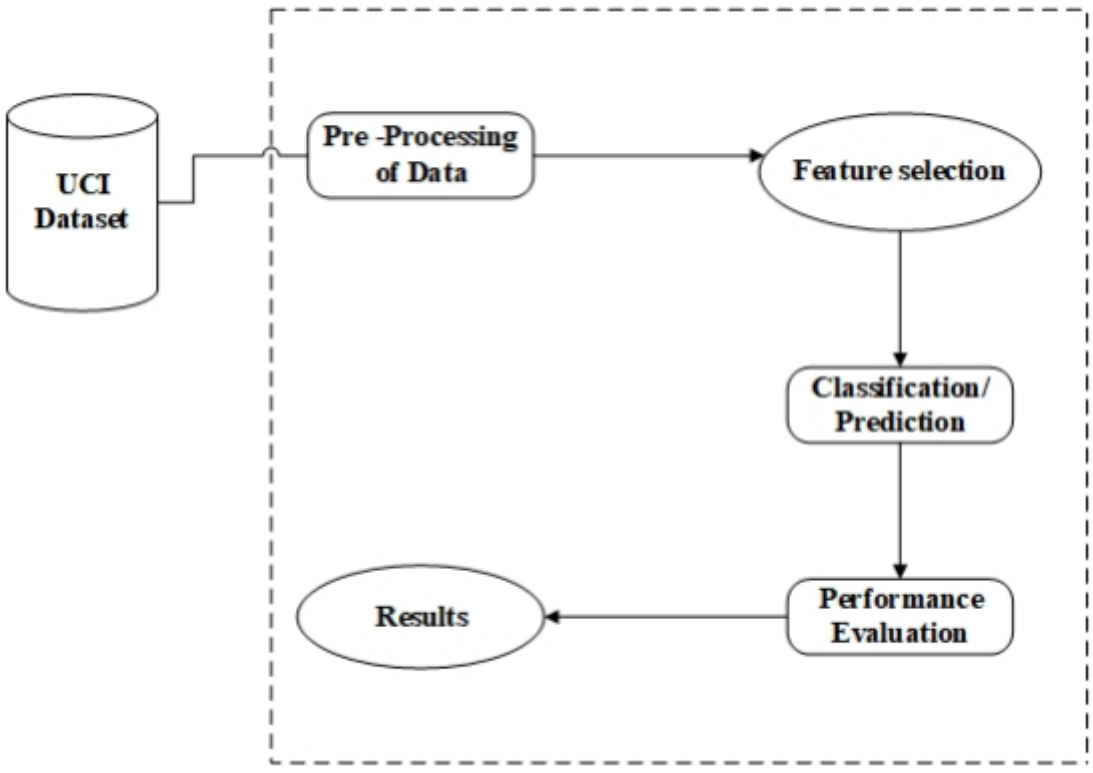

Figure 1: Proposed workflow

DATA PRE-PROCESSING

Heart illness is pre-prepared after dataset assortment. The dataset includes an entirety of 305 patient records, where 5 records are with some missing characteristics. Those 5 records have been eliminated from the dataset and the remaining 296 patient records are utilized in pre-getting ready. The multiclass variable and twofold grouping are presented for the characteristics of the given dataset. The multi-class variable is utilized to check the presence or nonattendance of coronary illness. In the case of the patient having coronary illness, the worth is set to 1, else the worth is set to 0 demonstrating the nonattendance of coronary illness in the patient. The field of "objective" identifies with the patient's essence of coronary illness. It is appraised whole number between 0 (no presence) and 4. Cleveland tests zeroed in on simply attempting to separate presence (values 1, 2, 3, 4) from need (esteem 0). Therefore from the dataset only 14 attributes are used which is shown in Table 1.

Table 1: UCI dataset attributes

1	Age	Age in years
2	Sex	Male or female
3	Cp	Chest pain type
4	Thestbps	Resting blood pressure
5	Chol	Cholestral
6	Restecg	Resting electrographic results
7	Fbs	Fasting blood sugar
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced gain
10	Oldpeak	ST depression induced by exercise relative to rest
11	Solpe	Peak exercise slope ST segment
12	Ca	Major vessels colored by floursopy
13	Thal	Defect type
14	Num	Diagnosis of heart disease

FEATURE SELECTION

Feature Selection is fundamental for the machine learning(ML) because of immaterial highlights influence the grouping techniques. The Feature determination techniques are the order exactness and decrease the model execution time. For include determination in this framework , this K-Nearest neighbor (KNN) algorithm is used which is a well-known FS algorithms that are used to select important features from the attributes.

KNN

The K-Nearest neighbor (KNN) is a clear, languid and non-parametric classifier. The KNN is supported when all the highlights are steady. The KNN is similarly called as case-based thinking and that has been used in various applications like example acknowledgment and measurable assessment. Characterization is gotten by recognizing the closest neighbor to choose the class of a dark model. The KNN is favored over other portrayal calculations in light of its high association speed and effortlessness.

1) Find the k number of tests in the dataset that is nearest to S

2) These k number of tests at that point vote to choose the class of S

The Accuracy of KNN dependent on the separation metric and K value. The Various conduct of estimating the separation between two examples is cosine, Euclidian separation. To survey the obscure example, KNN compute its K closest neighbors and allocate a class by dominant part casting a ballot.

Algorithm 1: KNN algorithm for Feature selection

Step 1: Let K is the no of nearest neighbours and T is the set of training samples

Step 2: For every test samples $z = x^2 y^2$

Step 3: Evaluate d(x^2, x), the distance between Z and each sample where (x,y) ∈ T

Step 4: Choose subset symbol T_Z, the k-closest set of training sample of Z

Step 5: Calculate $y^1 = argmax \sum(x_i, y_i)\, \varepsilon\, T_Z\, I(V = Y^1)$

Step 6: end

CLASSIFICATION

In the conventional method, the classifiers are utilized for exact prediction. In any case, it has been analysed that there is no single classifier that gives a best outcome for each dataset and not a single data mining procedures which give exact outcomes to cardiovascular infection related information. In this manner, an algorithm is required which can give optimal outcome.

For this classification, the proposed approaches will have a classifier which will hybridize with the optimization technique named as firefly Algorithm. This is motivated from the literature overview in which it is clarified that optimization method that can improve the components of the classifiers to improve their classification property. For example, in KNN technique if beginning loads of the organization are refreshed, the order precision is additionally going to differ. Therefore those weight esteems can be upgraded utilizing various optimization algorithms.

These findings motivate the proposed work to provide a hybrid model as best of classifier achieved (KNN) with optimization method of firefly optimization (fa-KNN).

Algorithm 2: Firefly Optimization Algorithm

Generate initial population of fireflies P = Xi(1; 2; . . . ; n) ;

Target work F(Xi), Xi = {x1; x2; . . . ; xm);

Light power Ii at Xi is dictated by F(Xi);

Calculte the firefly method parameters: γ;

Register a characteristic diminishment in view of directed fast reduce calculation

while (t Max Generations) do

Tune alluring parameter, β_(0) (t)

Assess fireflies light force utilizing target work using the below eq

$$Fitness\ (X) = \frac{m - |X|}{m} + \frac{n|R|\gamma_x(D)}{m\ \overline{\iota}}$$

Ranking light force of fireflies;

Find the current best solution;

For i = 1 to n do

For j = 1 to n do

If $Ii\ > Ij$ then

Compute attractiveness function using $\beta = \beta_0 e^{-\gamma T_{ij}^2}$

Differ the firefly position i in m-dimension using

$$X_i(t+1) = X_i(t) + \beta X_i(t) + \alpha(rand - \frac{1}{2})$$

Compute tanh function value, f(X_i^k)

f(X_i^k)=$\frac{\exp\left(2X_i^k\right)-1}{\exp(2X_i^k)+1}$ $i = 1, \ldots n; k = 1, \ldots . m$

If f , f(X_i^k)rand ≥ rand ≥ then X_i^k else $X_i^k = 0$ end if;

end if

end for

end for

t= t + 1;

end while

Post process results

The overall workflow of the proposed FA-KNN is explained with the step by step process which is shown in the algorithm 3.

Algorithm 3: FA-KNN approach

Step 1: Collection of input data from UCI to train the software.

Step 2: Read the input data.

Step 3: Applying KNN algorithms for feature selection

Step 4: Apply firefly optimization for selected classifier factor updation.

Step 7: Retrain the network to improve accuracy.

Step 8: Get final results from proposed FA-KNN model.

VI RESULT AND DISCUSSION

In this section, the comparisons results of several conventional classifiers are compared with the proposed algorithm of hybrid Fa-KNN are represented. The classification method is SVM, ANN, RF, KNN and Fa-KNN which is compared in terms of following parameters i.e. Accuracy, Precision and Recall respectively.

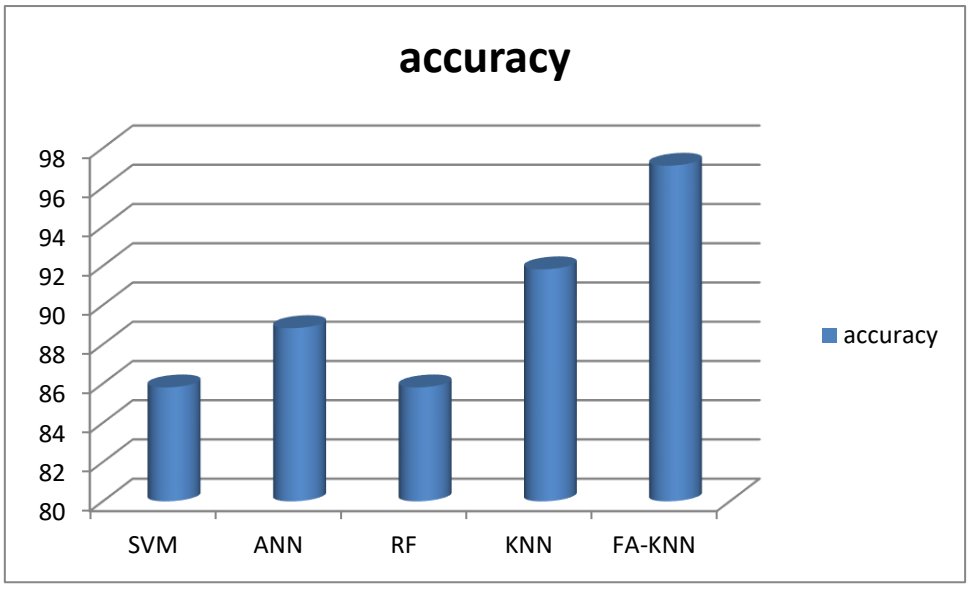

Figure 2: Accuracy value of several classifiers

According to the figure 2, it clearly shows that the comparison results for various classifiers with respect to accuracy. In the y-axis indicates the value of accuracy and x-axis given the several classifier methods namely SVM, ANN, RF, KNN and fa-KNN. The acquired outcomes shows that proposed classifier i.e. fa-KNN classifier has the maximum accuracy rate in contrast to all other classifiers as 97.1089.

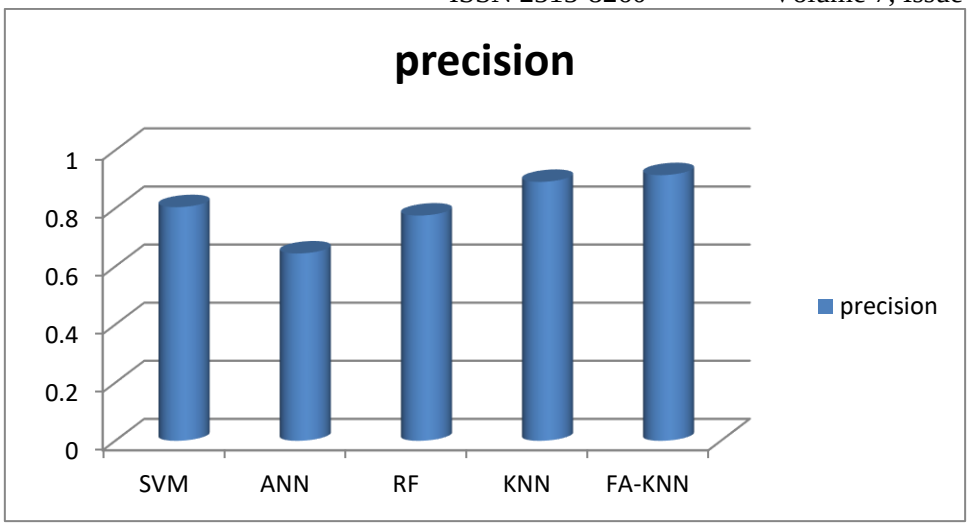

Figure 3: Precision results for different classifiers.

The figure 3 shows the precision values of various methods. It is observed that proposed Fa-KNN has the highest precision rate of 0.91453 which is the most efficient classifier than traditional classifiers.

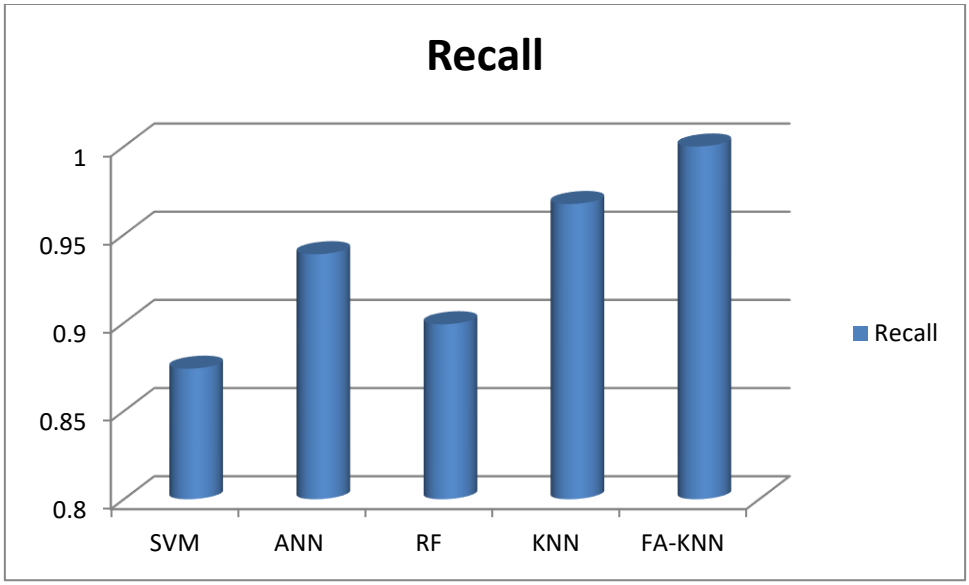

Figure 4: recall result of classifier methods

The recall value of different classifiers is charted in figure 3. According to the graph, the sequence of different classifiers for a recall rate can be represented and the result shows that Fa-KNN classifier has the highest recall rate and SVM classifier has the least recall. Therefore, it demonstrates that the proposed classifier is the most efficient one as it outperforms the other classifiers in terms of various parameters.

V CONCLUSION

Heart diseases prediction is the most common reason of death in the worldwide recently. It is important to predict the heart issues in earlier stage so that the disease can be prevented. Many previous studies are used for the heart disease prediction namely SVM, ANN, RF, KNN where the efficient

performance is achieved by using KNN. Even though, it is not enough for all type of dataset. Therefore, some novel hybrid classifier is presesnted in this paper for which the KNN classifier is hybridized with the optimization technique of firefly Algorithm. This firefly algorithm is used for finding the optimal results that can be achieved for the heart illness data. The proposed Fa-KNN classifier is compared with the conventional classifiers with respect to three parameters i.e. accuracy, precision and recall. From the obtained results of this paper, the proposed FA-KNN classifier is an efficient in terms of accuracy, precision and recall which gives the optimal solutions. These classification methods can be additionally improved by rising the number of attributes for the more exact prediction to be done. There are many probable improvements to enhance the scalability and precision of this prediction technique could be implemented in future.

REFERENCES

1. Khan, M., Ding, Q., & Perrizo, W. (2002). K-nearest neighbor classification on spatial data streams using p- trees. In Advances in Knowledge Discovery and Data Mining (pp. 517-528). Springer Berlin Heidelberg.
2. Lavanya, D., & Rani, K. U. (2012). Ensemble decision tree classifier for breast cancer data. International Journal of Information Technology Convergence and Services (IJITCS), 2(1), 17-24.
3. Picek, S., Golub, M.: On the Efficiency of Crossover Operators in Genetic Algorithms with Binary Representation. In: Proceedings of the 11th WSEAS International Conference on Neural Networks (2010)
4. Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). IBM New York.
5. Nidhi Bhatla Kiran Jyoti, An Analysis of Heart Disease Prediction using Different Data Mining Techniques, International Journal of Engineering Research & Technology (IJERT), 2012
6. Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin, A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining, International Journal of Applied Engineering Research, 2012.
7. Ma.jabbar, Dr.prirti Chandra, B.L. Deekshatulu, cluster based association rule mining for heart attack prediction, Journal of Theoretical and Applied Information Technology, 2011.
8. Ishtake S.H ,Prof. Sanap S.A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research,2013.
9. Dr. K. Usha Rani,analysis of heart diseases dataset using neural network approach,International Journal of Data Mining & Knowledge Management Process, 2011.
10. Y.S.Kumaraswamy and Shantakumar B.Patil "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network‖"; European Journal of Scientific Research, ISSN 1450-216X Vol.31 No.4, 2009.
11. Dilip Roy Chowdhury, Mridula Chatterjee & R. K. Samanta, An Artificial Neural Network Model for Neonatal Disease Diagnosis, International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (2): Issue (3), 2011.
12. Milan Kumari, Sunila Godara, Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST Vol. 2, Iss ue 2, June 2011