DIMENSION REDUCTION BASED ON ANT COLONY OPTIMIATION WITH EFFICIENT-KNN FOR PREDICTING OPTIMAL SOLUTION

[1]A. Jenita Mary, [2]K. Jayavani

[1]Assistant Professor, [2]Assistant Professor

[1]Department of Computer Applications, [2]Department of Computer Applications

[1]FSH, SRM IST, Chennai, India, [2]Sri Vijay College of Arts and Science, Dharmapuri, India.

Abstract : In the past years, medical data mining has become a popular data mining subject. Researchers have proposed several tools and several methodologies for developing effective medical expert systems. The applications of data mining techniques to medical data extract patterns which are useful for diagnosis, prognoses and treatment of diseases. This extraction of patterns allows doctors and hospitals to be more effective and more efficient. The huge volume of data is the barrier in the detection of patterns. Feature selection techniques mainly used in data preprocessing for data mining. Classification task leads to reduction of the dimensionality of feature space, feature selection process is used for selecting large set of features. The proposed algorithm reduces the dimension efficiently by E-KNN algorithm which is derived from K-Nearest Neighbor Algorithm. The classification further improves by ACO, the swarm intelligence technique computes optimal solution based on E-KNN feature selection method applied on heart disease dataset for better prediction. The accuracy of classification for whole feature set and the reduced feature subset are compared.

***IndexTerms* - Ant Colony Optimization-KNN, Feature Selection,Swarm intelligence,Heart disease.**

I. INTRODUCTION

The goal of data mining is to extract knowledge from huge amount of data. Data mining is an interdisciplinary field, whose fundamental lies on data analysis and pattern recognition. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data.

The World Health Organization has estimated that 12 million deaths occurs worldwide, every year due to the Heart diseases. Half the deaths in the United States and other developed countries occur due to cardio vascular diseases. It is also the chief reason of deaths in numerous developing countries. On the whole, it is regarded as the primary reason behind deaths in adults. The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in several illness, disability, and death. The diagnosis of diseases is a vital and intricate job in medicine. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous .Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. Efficient and accurate implementation of automated system needs a comparative study of various techniques available.

We propose a new and different approach to mine frequent patterns as discriminative features. It builds a Hierarchical structure that sorts or partitions the data onto nodes from the whole list. Then at each node, it directly discovers a discriminative pattern to further divide its examples into purer subsets that previously chosen patterns during the same run cannot separate. Since the number of examples towards leaf level is relatively small, the new approach is able to examine patterns with extremely low global support that could not be enumerated on the whole dataset by the batch method as given in Fig1, So in this paper had combined data mining techniques with ACO for better heart disease prediction. In this paper we use data mining to emphasize to discover knowledge that is not only accurate, but also comprehensible for the users. This paper aims to predict heart disease by reducing dimension with swarm intelligence technique, Ant Colony optimization and E-KNN.

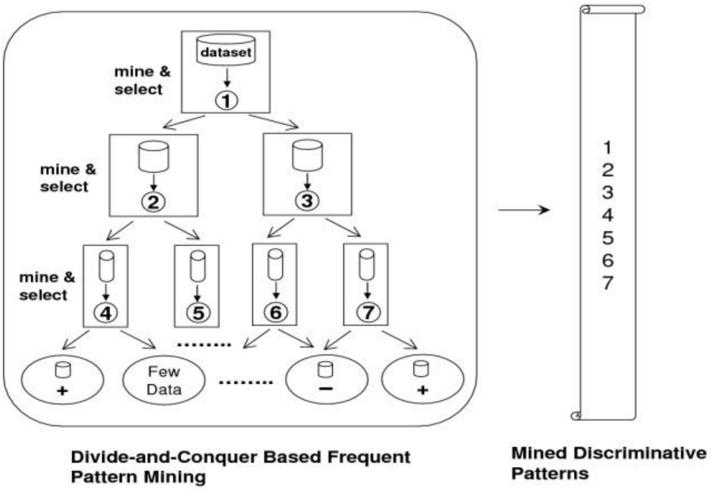


Fig.1

II. LITERATURE REVIEW

Since heart disease is something that cannot be detected by physical observation, but by analyzing different constraints that is associated with this disease. The diagnosis depends on the careful analysis of different clinical and pathological data of the patient by medical experts, which is a complicated process.

Yumin Chen et al. [1] proposed a new rough set approach to feature selection based on Ant Colony Optimization (ACO), which can adopt mutual information based feature significance as heuristic information. ?e paper also proposed a feature selection algorithm. ?is research approach started from the feature core, which changed the complete graph to a smaller one. To verify the efficiency of this algorithm, experiments are carried out on some standard UCI datasets. ?e results demonstrated that this algorithm could provide efficient solution to find a minimal subset of the features.

In 2016 Minal Zope, Sagar Birje, Lijo John, Amit Vasudevan , Nishant Salunkhe proposed an efficient algorithm hybrid with ANN (Artificial Neural Network) and K-mean technique approach for heart disease prediction. [2]

Aqeel Ahmed and Shaikh Abdul Hanan (2012) in their study combined 4 datasets of 920 records with 72 attributes to predict heart disease using different data mining technique. Only 13 attributes were shortlisted and used in their study, it was reported that Decision tree and SVM were most effective to predict heart disease.[3]

Kantesh kumar oad, et. al. (2014) used only 6 medical attributes to diagnose the CVD disease by using the fuzzy rule based expert system. The performance of the system matched with Neural Network and J48 Decision Tree Algorithms. [4]

Moloud Abdar, et al (2015)applied and compared data mining techniques to predict the rise of heart disease using five different algorithms such as C5.0, Neural Bayes, Support vector Machine, Neural Network and Logistic Regression with accuracy measures as: 93.02, 86.05, 88.37, 85.22 and 80.23 using 13 medical attributes. [5]

In 2014 Dr. Durairaj.M, Sivagowry.S describes a pre-processing technique and analyzes the accuracy for prediction after pre-processing the noisy data. It is also observed that the accuracy has been increased to 91% after pre-processing. Swarm Intelligence techniques hydride with Rough Set Algorithm are to be taken as future work for exact reduction of relevant features for prediction..[6]

In 2012, M. H. Mehta et al. [18] observed that in engineering field, many problems are hard to solve in some definite interval of time. These problems known as "combinatorial optimization problems" are of the category NP. These problems are easy to solve in some polynomial time when input size is small but as input size grows problems become toughest to solve in some definite interval of time. Long known conventional methods are not able to solve the problems and thus proper heuristics is necessary. Evolutionary algorithms based on behaviors of different animals and species have been invented and studied for this purpose. Particle swarm optimization is a new evolutionary approach that copies behavior of swarm in nature. However, neither traditional genetic algorithms nor particle swarm optimization alone has been completely successful for solving combinatorial optimization problems. So the authors present a hybrid algorithm in which strengths of both algorithms are merged and performance of proposed algorithm is compared with simple genetic algorithm.[7]

.Three different algorithms were proposed by Jyoti Soni et al those are Naïve Bayes, K-Nearest neighbor and Decision tree [9]. These algorithms were used to predict heart disease. Among these, Naïve Bayes performed better compared to other algorithms. The tool they used for medical data classification was named Tanagra. Moreover, those data were calculated using 10 fold cross validation [9]. Sometimes there were cases when there were kinds of attributes which didn't perform to predict heart disease or to classify data. In those cases, Genetic Algorithm was used to reduce the definite data size to obtain the best possible subset of attributes [8]

In 2008, Palaniappan, Set al. [10] proposes about healthcare industry which collects huge amounts of healthcare data which, unfortunately, are not ";mined"; to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited.

III. PROBLEM STATEMENT

After studying several research papers we observe that there are lot of work in the area of heart diseases detection because 17.7 million people die each year from CVDs, an estimated 31% of all deaths worldwide. K-Nearest Neighbors algorithm is a simplest among all machine learning algorithm. But the accuracy of the K-NN algorithm can be severely damaged by the presence of noisy or irrelevant features. The data was also sparse. Even though Ant Colony Optimization produces better Optimal solution, the dimensional rate if high which made prediction difficult. In order to get better prediction accuracy with reduced dimension can be achieved by using the hybrid of optimization method and a classification method by using ACO and K-NN respectively.

IV. RELATED WORK

K nearest neighbor(KNN) is a simple algorithm, which stores all cases and classify new cases based on similarity measure.KNN algorithm also called as 1) case based reasoning 2) k nearest neighbor 3)example based reasoning 4)instance based learning 5) memory based reasoning 6) lazy learning [4].KNN algorithms have been used since1970 in many applications like statistical estimation and pattern recognition etc.KNN is a non parametric classification method which is broadly classified into two types1) structure less NN techniques 2) structure based NN techniques. In structure less NN techniques whole data is classified into training and test sample data. From training point to sample point distance is evaluated, and the point with lowest distance is called nearest neighbor. Structure based NN techniques are based on structures of data like orthogonal structure tree (OST), ball tree, k-tree, axis tree, nearest future line and central line [5].Nearest neighbor classification is used mainly when all the attributes are continuous .Simple K nearest neighbor algorithm is given below. Steps 1) find the K training instances which are closest to Unknown instance Step2) pick the most commonly occurring classification for these K instances There are various ways of measuring the similarity between two instances with n attribute values. Every measure has the following three requirements. Let dist (A, B) be the distance between two points A, B then dist(A, B)) = 0 and dist(A, B) = 0, if and only if A = B. (Costiveness) dist(A, B) = dist(B, A) (Symmetry) dist(A, B) + dist(B, C) = dist(A, C) (the Triangle Inequality)

Ant based algorithms are optimization based and roused by natural foraging behavior of real ants. While looking for food, initially ants scrutinize the area surrounding its home in a random way. If a ant identifies a source for food, it assesses the capacity and nature of food and carry a tiny quantity of it to its home. Amid the arrival journey, the ant stores a trail of pheromone (which is a chemical substance) on the ground. The quantity of pheromone spread on the path relies on the measure level and food nature, which will control all the ants to food source. There arises a communication (indirect) among the ants by means of pheromone trails empowering it to locate the briefest ways between the home and food sources. This specific characteristic of real ants is utilized in the artificial ants to take care in solving optimization oriented problems. In ant colony optimization, artificial ants are built probabilistically incorporating dynamical forged pheromone trails. The focal part of the ACO algorithm is the pheromone methods including the state change control and refreshing guideline, which is utilized to probabilistically test the path. Figure 2a to 2d demonstrates a situation with a route from the home to the food source, where all ants pursue the pheromone trail. Abruptly, when an obstacle gets in their direction towards the food source, them immediately first ant arbitrarily select the alternate path (i.e., upper and lower path). When comparing with the upper path and lower path, the upper path seems to be better and shorter than the lower path tending towards reaching the food source in a shorter time and distance. As already discussed, ants spread pheromone on their traveling path, pheromone gets diminished after a short course of time. The shorter path has the ants pheromone stronger than the longer path. According to the ACO algorithm, the paths which have stronger pheromone are considered as the best path in finding the solution to the problem [18]. This natural behavior of ants can be utilized in finding solutions to machine learning algorithm stream [18].

Ant Colony Optimization Algorithm

Ant Colony optimization is a swarm intelligence algorithm. Swarm Intelligence (SI) can therefore be defined as a relatively new branch of Artificial Intelligence that is used to model the collective behavior of social swarms in nature, such as ant colonies, honey bees, and bird flocks. Although these agents (insects or swarm individuals) are relatively unsophisticated with limited capabilities on their own, they are interacting together with certain behavioral patterns to cooperatively achieve tasks necessary for their survival. The social interactions among swarm individuals can be either direct or indirect [6]. The typical swarm intelligence system has the following properties:

(1) It is composed of many individuals.

(2) The individuals are relatively homogeneous.

(3) The interactions among the individuals are based on simple behavioral rules that exploit only local information that the individuals exchange directly or via the environment. Examples in natural systems of SI include ant colonies, bird flocking, animal herding, bacterial growth, and fish schooling

The plan of an ACO algorithm implies the requirement of the subsequent features.

- A rule for pheromone updating, which precise how to fine-tune the pheromone trail(t)
- A probabilistic transition rule based on the value of the heuristic function (h) and on the contents of the pheromone trail (t) that is employed to iteratively erect a solution.

V. PROPOSED METHODOLOGY

Our proposed approach Ant colony optimization derived from KNN and particle swarm optimization to increase the classification accuracy of heart disease data set. We used Efficient Nearest Neighbor search as a goodness measure to prune redundant and irrelevant attributes, and to rank the attributes which contribute more towards classification. Least ranked attributes are removed, and classification algorithm is built based on evaluated attributes. This classifier is trained to classify heart disease dataset as either healthy or sick. Our proposed algorithm consists of two parts

1) First part deals with feature selection based on E-KNN

2) Part two deals with getting optimal solution based on ACO.

Algorithm for our proposed method is shown below.

A. ACO- Algorithm

```
1,Load Heart Database Xn
2. Normalize data Xn € [0,1]
3. Preprocessing the Data sets
Xn1 = {} Xn1 € Xn
4. Apply E-KNN Feature Subset
5. Attributes are ranked based on their value
6. Classify the data using E-KNN search
7. improve accuracy using nonconvex neighbour search
8. optimizing using ACO
9. repeat step 3 & 4
10. get optimization
```

B. E-KNN Algorithm

```
1. Get Heart set Di = {D1, D2, D3…. Dn}
2. Create Training Data set Tr = {T1, T2, T3 …. Trn}
 3. Get the Test data set Te = {Te1, Te2, ..Ten}
4. Compare Tr &Te 5. Select K value
5. Get similarity
6. Get dissimilarity
```

Our proposed method aims to reduce dimensions and select features based on E-KNN which is derived from KNN and get optimal solution based on ACO for disease prediction. This hybridization gives more accuracy with less attribute values.

5.1 Experimental Results

The input database used here was The Cleveland, Hungarian, Switzerland, and VA Long Beach data sets, which are taken from Data Mining Repository of University of California, Irvine (UCI).This is the most commonly used dataset used by the researchers for analyzing the heart disease for their research work. Attributes of heart disease and their corresponding data types is shown in Table 1.Our proposed method, E-KNN + ACO improves the classification accuracy with suitable fitness value by reducing dimensions and find the optimal solutions. The sample screenshots are given below.

5.2 Data set

To predict heart disease the dataset containing 270 instances is collected from UCI repository. This database contains 13 attributes extracted from a larger set of database.

1) Attribute Information:

- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholesterol in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved • exercise induced angina
- old peak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by fluoroscopy
- then: 3 = normal; 6 = fixed defect; 7 = reversible defect

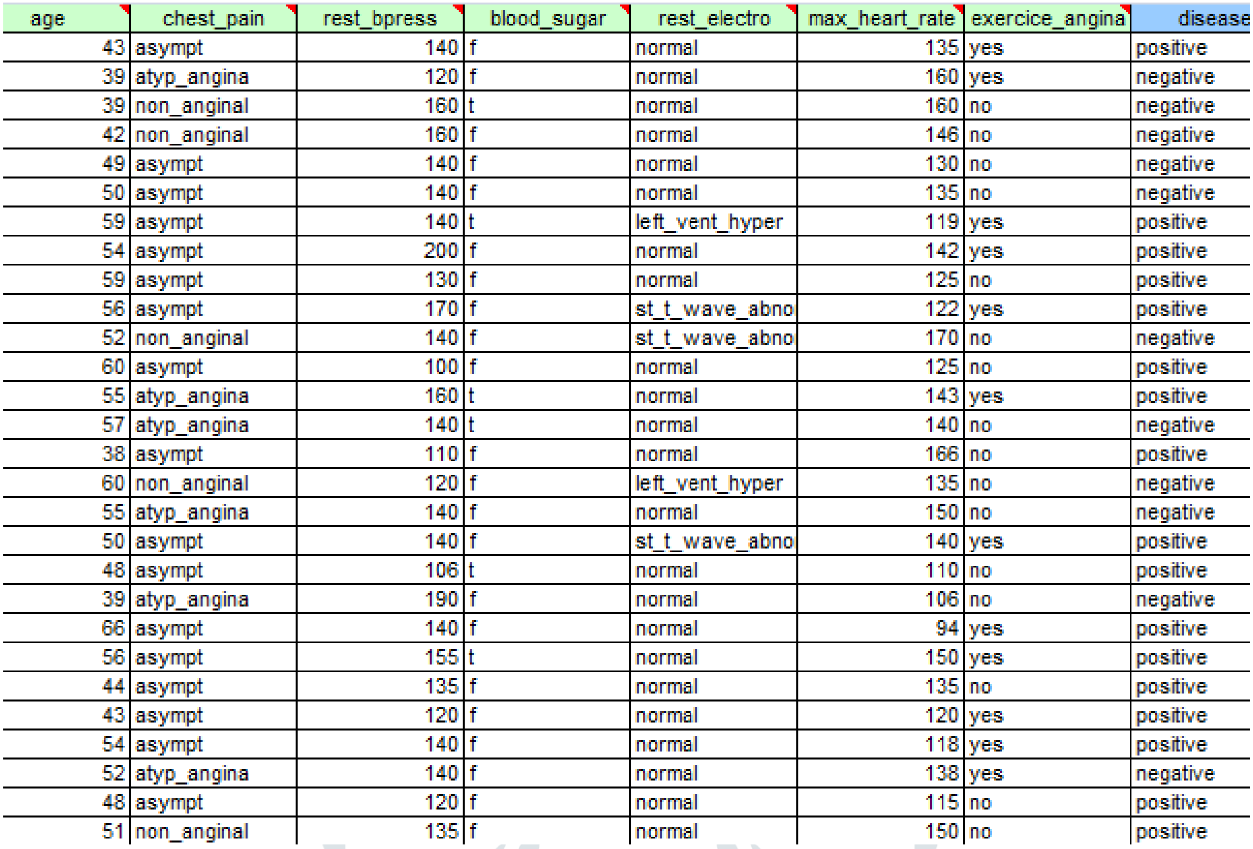
2) Attributes types

- Real: 1, 4,5,8,10,12
- Ordered: 11, Binary: 2, 6, 9
- Nominal: 7,3,13

3) Variable to be predicted

Absence (1) or presence (2) of heart

Table 1 Data Classification



age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_rate	exercice_angina	disease
43	asympt	140	f	normal	135	yes	positive
39	atyp_angina	120	f	normal	160	yes	negative
39	non_anginal	160	t	normal	160	no	negative
42	non_anginal	160	f	normal	146	no	negative
49	asympt	140	f	normal	130	no	negative
50	asympt	140	f	normal	135	no	negative
59	asympt	140	t	left_vent_hyper	119	yes	positive
54	asympt	200	f	normal	142	yes	positive
59	asympt	130	f	normal	125	no	positive
56	asympt	170	f	st_t_wave_abno	122	yes	positive
52	non_anginal	140	f	st_t_wave_abno	170	no	negative
60	asympt	100	f	normal	125	no	positive
55	atyp_angina	160	t	normal	143	yes	positive
57	atyp_angina	140	t	normal	140	no	negative
38	asympt	110	f	normal	166	no	positive
60	non_anginal	120	f	left_vent_hyper	135	no	negative
55	atyp_angina	140	f	normal	150	no	negative
50	asympt	140	f	st_t_wave_abno	140	yes	positive
48	asympt	106	t	normal	110	no	positive
39	atyp_angina	190	f	normal	106	no	negative
66	asympt	140	f	normal	94	yes	positive
56	asympt	155	t	normal	150	yes	positive
44	asympt	135	f	normal	135	no	positive
43	asympt	120	f	normal	120	yes	positive
54	asympt	140	f	normal	118	yes	positive
52	atyp_angina	140	f	normal	138	yes	negative
48	asympt	120	f	normal	115	no	positive
51	non_anginal	135	f	normal	150	no	positive

Metrics: Evaluated Performance of PSO Algorithm utilizes following Metrics: complexity, accuracy, sensitivity, specificity given in Karalolis et al.(2010) these metrics are common in medical applications are discussed below[19]
True Positive (TP): It denotes the number of Heart Disease Patients classified correctly by the Hybrid Neural Network
True Negative (TN): It denotes the number of patients not having heart disease correctly classified by the system.
False Positive (FP): It denotes the number of healthy patients wrongly classified as a heart disease patient by the system
False Negative (FN): It denotes the number of healthy patients classified as a heart disease patient by the system
The performance was evaluated in terms of accuracy, sensitivity, specificity

Table-2: Proposed algorithm performance results with different methods.

Method	Accuracy	Sensitivity	Specificity
K-NN without optimization	64.16	54.92	35.93
K-NN + ACO (optimization)	76.53	51.17	46.73
E-KNN + ACO	77.84	54.65	48.75

Table-2 and Fig.2 describes the performance analysis of proposed E-KNN+ACO.

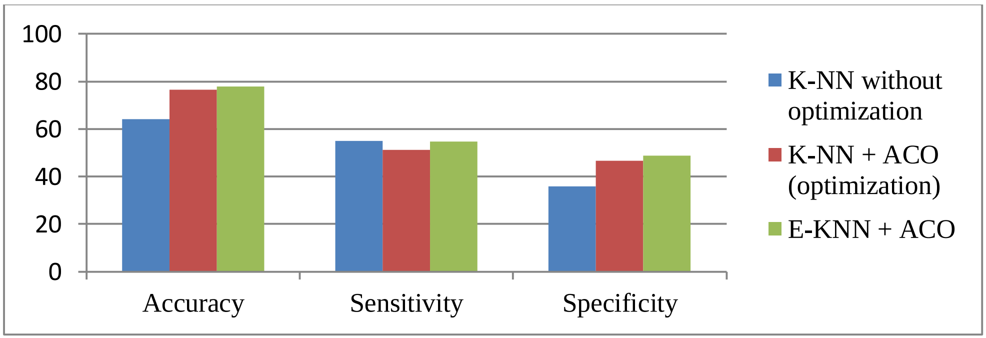


Fig.2

IV. CONCLUSION

This paper addressed the prediction of heart disease based on PSO and KNN .Our approach uses EKNN as a classifier to reduce the misclassification rate. This paper also investigates PSO based feature selection measure to select a small number of features and to improve the classification performance. The results suggest that proposed approach can significantly improve the learning accuracy. From simulation results, it is concluded that PSO based feature selection is important for classification of heart disease. This model helps the physicians in an efficient prediction of diseases with predominant features. In future, we want to integrate ensemble classifiers with PSO to develop a decision support system for early diagnosis of heart disease and also would like to compare GA and PSO for heart disease set.

REFERENCES

[1].Chen, Yumin, Duoqian Miao, and Ruizhi Wang. "A rough set approach to feature selection based on ant colony optimization." Pattern Recognition Letters 31, no. 3 (2010): 226-233.

[2] Minal Zope, Sagar Birje, Lijo John, Amit Vasudevan , Nishant Salunkhe , a system for heart disease prediction using data mining techniques , international journal of innovations in engineering research and technology [ijiert], volume 3, issue4, apr.-2016.

[3] Aqueel Ahmed, Shaikh Abdul Hannan. Data Mining Techniques to Find Out Heart Diseases: An Overview. International Journal of Innovative Technology and Exploring Engineering, Volume-1, Issue-4, September 2012.

[4] Kantesh Kumar Oad, Xu DeZhi & Pinial Khan Butt, “A Fuzzy Rule based Approach to Predict Risk Level of Heart Disease”,Global Journal of Computer Science and Technology: C Software & Data Engineering Volume 14 Issue 3 Version 1.0 Year 2014.

[5] Parvathi I, Siddharth Rautaray, Survey on Data Mining Techniques for The Diagnosis of Diseases In Medical Domain, International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 838-846.

[6] Dr. Durairaj.M, Sivagowry.S,A Pragmatic Approach of Preprocessing the Data Set for Heart Disease Prediction, International Journal of Innovative Research in Computer and Communication Engineering , Vol. 2, Issue 11, November 2014

[7]M. H. Mehta,” Hybrid Genetic Algorithm with PSO Effect for Combinatorial Optimisation Problems”, International Journal of Advanced Computer Research (IJACR), Volume-2 Number-4 Issue-6 December2012.

[8]M. Anbarasi, E. Anupriya and N.Iyengar, “Enhanced prediction of heart disease with feature subset selection using Genetic algorithm”, International Journal of Engineering Science and Technology vol.2, pp.5370- 5376, 2010.

[9] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer 2006,Vol:345, page no. 721- 727.

[10].Palaniappan, S.; Awang, R., “Intelligent heart disease prediction system using data mining techniques”, IEEE 2008.

[11] Peterson, Leif. "K-Nearest Neighbor". N.p., 2017. Print

[12] G. Karthiga, C. Preethi, and R. D. H. Devi, “Heart Disease Analysis System Using Data,” vol. 3, no. 3, pp. 3101–3105, 2014.

[13] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques, 3rd ed. Elsevier Inc., 2011.

[14] S. Tabakhi, P. Moradi, and F. Akhlaghian, “An unsupervised feature selection algorithm based on ant colony optimization”, Eng. Appl. Artif. Intell., Vol.32, pp.112–123, 2014.

[15] M. Dorigo and M. Birattari, “Ant Colony Optimization,” in In Encyclopedia of machine learning (Springer), C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, pp.36–39, 2010