IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

# Genetic Algorithm Based Approach For Predict Disease and Avoid Congestion in Data Mining

**J. Adamkani[1], M. Wasim Raja[2*]**

[1]P.G. Dept of Computer Science, The New College, Chennai,India
[2]P.G. Dept of Computer Science, The New College, Chennai,India

** Corresponding Author: kltwasim@rediffmail.com*

**Available online at: www.ijcseonline.org**

Accepted: 19/Jul/2018, Published: 31/Jul/2018

***Abstract*:** The data mining techniques is a major significant position in the field of healthcare and medical industry to analyze the medical data and finding the patterns from those data. The primary goal of the research analysis work is to predict the patient diseases from the medical data sets. Medical practitioners is getting difficult to predict the disease, actually it is one of the complex task which require their experience and knowledge. The main objective of data mining techniques to predict the possible disease from patient dataset and based on patient serious condition priority wise to reduce the congestion in the network. In this paper proposed the genetic approach is efficient for associative classification algorithm to predict the disease. The motivation is by using genetic algorithm in the discovery of high level prediction rules which can be highly comprehensible having high predictive accuracy and high interestingness values.

***Keywords*:** Data Mining, Association Rule, Keyword Based Clustering, Genetic algorithm, Classification.

## I. Introduction

The data mining techniques is a major significant position in the field of healthcare and medical industry to analyze the medical data and finding the patterns from those data. The primary goal of the research analysis work is to predict the patient diseases from the medical data sets. Most of the researchers are interestly doing their research work in this domain. By using data mining techniques, the existing available papers are mainly focusing on predicate diseases from health data sets. In recent years, the major reason is huge amount of data availability in information industry that attracted the great attention of data mining and turning is needed that the data is useful for information and knowledge [1].

The major challenging task in healthcare organizations like hospitals, medical centre etc is the provision of quality services at low costs. Quality service that helpful which implies perfectly diagnosing the patients and administering treatments. Clinical decision is poor which is leading to disastrous consequences that can be unacceptable [2]. Clinical tests taken in hospital must be in low cost. Health care organizations must have ability to analyze data. Millions of patient records are computerized and stored in data mining techniques that helps to answer the critical and important questions which is related to health care. Prediction involves in data set by using some variables to predict the variables or future value of interest [2]. In biomedical field data mining and its techniques plays an essential role for prediction of various diseases. The physicians can not able to diagnose the disease correctly because on same category the patients suffering more than one type of disease. The category of disease prediction which leads to missing concentration or unhealthy practices. The healthcare industry provides large amounts of data about healthcare and that need to be mined to ascertain hidden information for valuable decision making. Discover of hidden patterns and relationships often go unused. The patient records are classified and predicted based on the disease if the patient having the symptoms of heart disease and using disease risk factors. It is indispensable to find the best fit algorithm that has greater accuracy, less cost, speedy and memory utilization on classification in the case of heart disease prediction category [3].

Association rules are mined on a medical data set to improve heart disease diagnosis. Each rule represents a simple predictive pattern that describes a subset of the data set projected on a subset of attributes. From a medical perspective, association rules relate combinations of binary target attribute (absence/existence of artery disease) and subsets of independent attributes (risk factors and heart muscle health measurements). Association rules have important advantages over traditional algorithms [4].

 

Yet complicated task are need to be executed accurately and efficiently in Medical diagnosis which is regarded as an important. Extremely advantageous in this system is automation. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Appropriate computer based information and/or decision which supports the systems can aid in achieving at a reduced cost clinical tests. The need of comparative study between various techniques which is available for implementing the automated system is efficient and accurate. This paper aims to analyse the disease which is diagnose in recent years different predictive descriptive data mining techniques are proposed.

In this paper used Genetic algorithm [4], to reduce the size of the actual data to get the optimal subset of attributes sufficient for heart disease prediction. Classification is a supervised learning method to extract models describing important data classes or to predict future trends. Three classifiers such as Decision Tree, Naïve Bayes and Classification via clustering have been used to diagnose the presence of heart disease in patients. Classification via clustering: Clustering is the process of grouping similar elements. This technique may be used as a pre processing step before feeding the data to the classifying model[5]. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering.

Data mining is a essential step in discovery of knowledge from large data sets. In recent years, Data mining has found its significant footing in every field including health care. When compare to data analysis, the Mining process which includes classification, clustering, association rule mining and prediction. It also extents other disciplines like Data Warehousing, Statistics, Machine learning and Artificial Intelligence and so on.

## II. Literature Review

Medical diagnosis is also subjective and it is depends not only on the availability of data but physician experience also and even physician on the psycho-physiological condition. A Demonstration can be varied significantly based on the number of studies that one patient diagnosis if the patient is tested by different physicians or even by the same physician at various times.

Himigiri. Danapana et al[6] discussed about the research which intends to provide a survey about current techniques based on knowledge discovery in databases using data mining techniques that are in used in particularly Heart Disease Prediction in medical research today's. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of decision tree.

Fariba Shadabi et al[7] Artificially Intelligent (AI) tools are used to deal the uncertain and incomplete data sets. For prediction purposes neural network classifiers has been used successfully in many complex situations. Research demonstrates that AI-based data mining tools is also used successfully in many medical environments. They conclude advances research for understanding the application of Artificial Intelligence and Data Mining tools for clinical data by demonstrating the potential techniques for complex clinical situations.

Two kinds of data mining algorithms named evolutionary termed GA-KM and MPSO-KM cluster the cardiac disease data set and predict model accuracy [8]. This is a hybrid method that combines momentum-type particle swarm optimization (MPSO) and K- means technique. The comparison is made with C5, Naïve Bayes, K-means, Ga-KM and MPSO-KM to evaluate the techniques accuracy. The experimental results showed that accuracy improved when using GA-KM and MPSO-KM [8]. The researchers created class association rules using feature subset selection to predict a model for heart disease. Association rule determines relations amongst attributes values and classification predicts the class in the patient dataset [9]. The measure of Feature selection such as genetic search, to determine the attributes which contribute towards the heart diseases prediction. The researchers [10] implemented a hybrid system that uses global optimization benefit of genetic algorithm for initialization of neural network weights. The prediction of the heart disease is based on risk factors such as age, family history, diabetes, hypertension, high cholesterol, smoking, alcohol intake and obesity [10].Graph based approach for heart disease prediction was proposed by [11].Their method is based on maximum clique and weighted association rule mining. Associative classification for heart disease prediction was proposed by [12].They used Gini index based classification to predict the heart disease. The researchers [13] used the data mining algorithms decision trees, naïve bayes, neural networks, association classification and genetic algorithm for predicting and analyzing heart disease from the dataset. An experiment performed by [14] the researchers on a dataset produced a model using neural networks and hybrid intelligent algorithm, and the results shows that the hybrid intelligent technique improved accuracy of the prediction.

The artificial neural network technique is used in data mining methods for effective heart attack prediction system. First the dataset is used for heart diseases

 

prediction to pre-processed and clustered by means of K-means clustering algorithm [15]. Then neural network is trained by the selected patterns significantly. For training, Multi-layer Perceptron Neural Network with Back propagation is used. The results indicate that the algorithm is used for capable of predicting the heart diseases more efficiently. The prediction of heart diseases significantly uses 15 attributes, with basic data mining technique like ANN, Clustering and Association Rules, soft computing approaches etc. The result shows that Decision Tree performance is more and sometimes Bayesian classification is having similar accuracy when compared to decision tree but other predictive methods like K-Nearest Neighbor, Neural Networks, Classification based on clustering will not perform well [16]. By using the Weighted Associative Classifier (WAC), a slight change has been made, instead of considering 5 class labels, only 2 class labels are used. First "Heart Disease" and another one "No Heart Disease". The maximum accuracy (81.51%) has been achieved. When genetic algorithm is applied, the accuracy of the Decision Tree and Bayesian Classification is improved by reducing the actual data size. The dataset contains 909 patient records are collected data are based on heart diseases and 13 attributes has been used for consistency [17]. The patient records have been splitted equally as 455 records for training dataset and 454 records for testing dataset. The attributes is reduced to 6, after applying genetic algorithm and when compared with other algorithms, the decision tree performs more efficiently with 99.2% accuracy.

## III. RESEARCH METHOLOGLOGY

In this paper proposes to detect the accurate disease based on the user symptoms from the hospital information database by using three algorithms are:

1. Association rule mining Algorithm which is used to extract the data from the hospital information database.
2. Keyword based clustering algorithm is used to find the accurate disease which is affecting the patient.
3. Genetic algorithm is used to prioritize the patient in order to avoid the congestion.

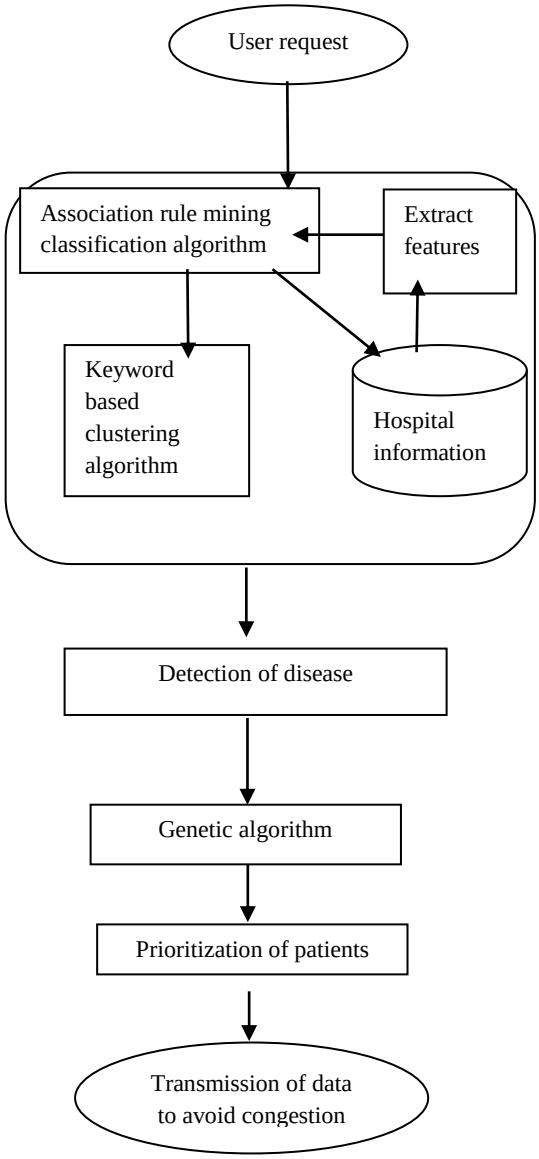


**Figure1. System Architecture**

### 3.1 Association rule algorithm

Association rule is the mining technique which is to find the association rules met the user-specified requirement is minimum support and minimum confidence from the transaction database D. The whole mining process can be carries out into the following two steps: first, to find all frequent item sets, that is, finding all the item sets is supported greater than the given support threshold; second, based on the obtained frequent item sets, generate a corresponding strong association rule, that is, generate the association rules which support and confidence respectively greater than or equal to the given support threshold and confidence threshold.

 

Let I = ( i1,i2....im) be a set of literals, called items. Let D be a database of transaction, where each transaction T is a set of items such that T #I. For a given item set X #I and a given transaction T, we say that T contains X if and only if X #I. The support count of an item set X is defined to be supx = the number of transactions in D that contain X. we say that an item set X is large, with respect to support threshold of s%, if supx P /D/ s%, where /D/ is the number of transactions in the database D. An association rule is an implication of the form ''X ) Y", where X #I, Y #I and X \ Y = ;. The AR ''X ) Y" is said to hold in database D with confidence c% if no less than c% of the transactions in D that contain X also contain Y. The rule X ) Y has support s% in D if supX[Y] =/D/_s%

**3.1.1 Apriori algorithm**

The Apriori algorithm works iteratively. First step in this algorithm to find the set of large 1-item sets, and then set of 2-itemsets, and so on. Based on maximum item set length the number of scan over the transaction database. Apriori is based on the following fact: The simple but powerful observation leads to the generation of a smaller candidate set using the set of large item sets found in the previous iteration. The Apriori algorithm presented is given as follows:

```
Apriori()
L1 ={large 1-itemsets}
k = 2
while Lk_1–/ do
begin
Ck ¼ apriori genðLk_1Þ
for all transactions t in D do
begin
Ct ¼ subsetðCk; tÞ
      for all candiate c 2 Ct do
c.count = c.count+1
end
Lk ¼ fc 2 Ckjc:count P minsupg
k = k + 1
end
```

First Apriori algorithm scans the transaction databases D in order to count the support of each item i in I, and determines the set of large 1-itemsets. Then, the iteration is performed, each of the computation of the set of 2-itemsets, 3-itemsets, and so on. There are two steps for kth iteration:

Generate the candidate set Ck from the set of large (k-1)-itemsets,

Scan the database in order to compute the support of each candidate itemset in Ck.

**3.2 Keyword based clustering algorithm**

Keyword-based document clustering creates a cluster by the keywords of each document. Suppose that C is a set of clusters that is finally created by the clustering algorithm. If n is the number of clusters in C, then C is a set of clusters C1,C2,C3...Cn.

$$C=\{C1,C2,C3....Cn\}$$

Each cluster is initialized by document d that is not assigned to the existing clusters, and d is a seed document of . When a new cluster is created, expansion and reduction steps are repeated until it reaches a stable state from the start state.

**3.2.1 Cluster Initialization**

The clustering algorithm first step is a creation and initialization of a new cluster. A document D is selected that does not belong to any other cluster, and it is assigned to a new cluster $C_i$ that is an initial state of cluster.

$$Ci= \{0\}$$

At this time, a document D that is the first document in the new cluster is called a seed document.

**3.2.2 Expansion of cluster**

In the initialisation step of the cluster, a new cluster $C_i$ , an initial state of cluster C, is established as the seed document, and the keyword set i is initialised by the key word Kc set of the seed document. In the expanding step of the cluster, by adding more related documents to the cluster results the cluster get expanded, that include the keywords of the seed document as the related documents of the seed document. The cluster expansion is performed by iteration which is keyword expansion and cluster expansion. More documents are added to a cluster by the similarity evaluation between the keyword set and the document. If a new document is added to a cluster, then the keywords also added in the document to the keyword set in the cluster.

**3.2.3 Cluster reduction**

This step is to produce a complete cluster by removing the documents that are not related to the cluster. For the cluster Ci documents of a low similarity to the cluster are removed, that are not related to a cluster Ci through the similarity computation with the cluster. Ultimately, the cluster C is completed which consists of the related documents after filtering from the non-related documents. If a cluster C is completed the next cluster is created through the same process. Clustering is terminated if all the documents are clustered or no more clusters are created.

**3.3 Entrophy based Genetic algorithm**

Genetic algorithms are computing methodologies constructed in analogy with the process of evolution. It closely resembles the natural process of regeneration, reproduction, inheritance evolution. Genetic algorithms are typically used for problems that cannot be solved efficiently with traditional techniques. Genetic algorithms

 

is useful for searching very general spaces and optimization problems. Each solution generated in Genetic algorithms is called a chromosome (individual).Each chromosome is made up of genes, which are the individual elements (alleles) that represents the problem. The collection of chromosomes is called a population. The internal representation of the chromosomes is known as its genotype.

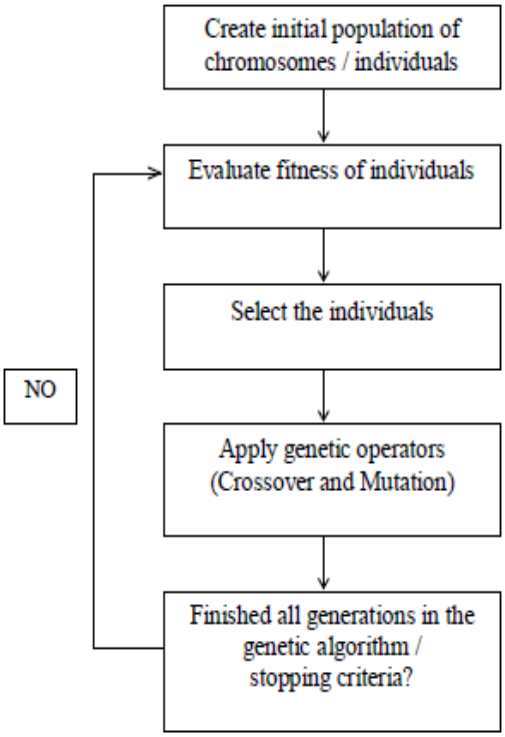


**Figure 2: Flow chart of genetic algorithm**

The functions of genetic operators are as follows:
1) Selection: selection deals with the probabilistic survival of the fittest in that, more fit chromosomes are chosen to survive.

2) Crossover: This operation is performed by selecting a random gene along the length of the chromosomes and swapping all the genes after that point. Various types of crossover operators are a) single point b) two point c) uniform d) half uniform e) reduced surrogate crossover f) shuffle crossover g) segmented crossover

3) Mutation: mutation alters the new solutions so as to add stochasticity in the search for better solution. The most common method way of implementing mutations is to select a bit at random and flip (change) its value. There are 2 types of mutations use in genetic network programming 1) mutating the judgment node 2) mutating the value of the judgment node. In associative classification attributes and their values are taken as judgment nodes and class values as processing nodes.

### 3.3.1 Entropy measures

Entropy is a commonly used to measure the information theory. Originally is used to characterize the impurity of an arbitrary collection of examples. In our implementation, the entropy is used to measure the homogeneity of the rule matches. Given a collection *S*, containing the examples that a certain rule *R* matches, let $P_i$ be the proportion of examples in *S* belonging to class *i*, then the entropy *Entropy(R)* related to this rule is defined as:

$$Entropy(R) = -\sum_{i=1}^{n} (p_i \log_2 (p_i))$$

where *n* is the number of target classifications. While an individual consists of a number of rules, the entropy measure of an individual is calculated by averaging the entropy of each rule:

$$Entropy(individual) = \frac{\sum_{i=1}^{N_R} Entropy(R_i)}{N_R}$$

where $N_R$ is number of rules in the individual

## IV. PERFORMANCE ANALYSIS

The performance of the algorithm is evaluated using the measures like accuracy, Time computation, efficiency defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

**Table 1 The comparison results on the prediction accuracy and standard deviation**

| Parameters | Decision trees | Decision tree With boosting | Neural networks | Navie bayes | Entrophy based GA |
|---|---|---|---|---|---|
| Accuracy | 67.56 | 60.98 | 56.09 | 53.66 | 80.04 |
| Time computation | 0.08 | 0.87 | 0.01 | 0.15 | 0.03 |
| Efficiency | 44.93 | 45.44 | 43.34 | 39.60 | 49.52 |

 

Table 2. Comparison of parameters between Non GA and GA

| | Our GA approach | Decision trees | Decision trees with boosting | Neural networks | Naive Bayes |
|---|---|---|---|---|---|
| Run 1 | 90.77 | 87.69 | 89.23 | 89.23 | 66.15 |
| Run 2 | 89.23 | 84.62 | 86.15 | 86.15 | 78.46 |
| Run 3 | 89.23 | 89.23 | 90.77 | 90.77 | 84.62 |
| Run 4 | 92.31 | 90.77 | 90.77 | 89.23 | 81.54 |
| Run 5 | 86.15 | 81.54 | 81.54 | 84.62 | 75.38 |
| Run 6 | 89.23 | 87.69 | 87.69 | 87.69 | 80.00 |
| Run 7 | 84.62 | 81.54 | 84.62 | 84.62 | 73.85 |
| Run 8 | 87.69 | 86.15 | 87.69 | 86.15 | 83.08 |
| Run 9 | 90.77 | 86.15 | 89.23 | 87.69 | 76.92 |
| Run 10 | 86.76 | 88.24 | 91.18 | 86.76 | 75.00 |
| Average | 88.68 | 86.36 | 87.89 | 87.29 | 77.50 |
| Standard deviation | 2.37 | 3.06 | 3.08 | 2.03 | 5.36 |

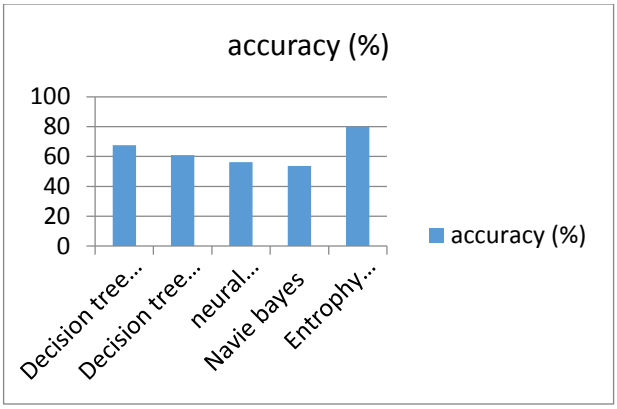


Figure 3. Comparison of accuracy in (%)

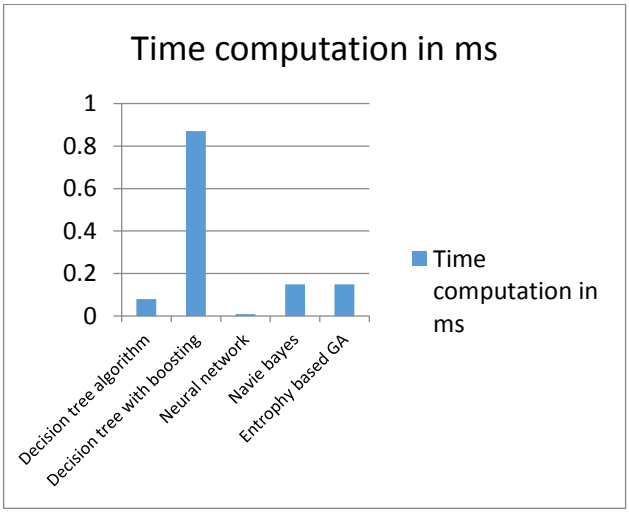


Figure 4. Comparison of computational time

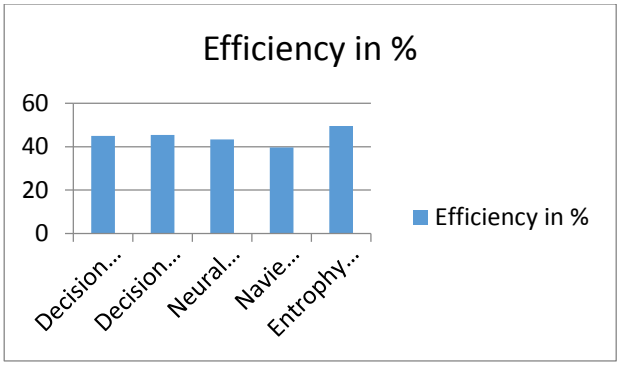


Figure 5. Comparison of efficiency in (%)

## V. CONCLUSION

Data mining is the process of analysing a data from different prospectivee and provides useful information is based on that the outcomes of predicting the diseases for a patient from the huge volume of data presents in the hospital information database. In this paper using the association rule mining algorithm for extract the matched features from the hospital information database and keyword based clustering algorithm is used to find the accurate disease which is affected by the patient. The proposed efficient associative classification algorithm using entrophy genetic approach for disease prediction resulted in having high predictive accuracy and of high efficiency values.

## REFERENCE

[1] S. Vijayarani* and S. Sudha "An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples" Indian Journal of Science and Technology, Vol 8(17), August 2015.

 

[2] Aswathy Wilson, Gloria Wilson, Likhiya Joy K "Heart disease prediction using data mining techniques"
[3] G. Purusothaman* and P. Krishnakumari "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease" Indian Journal of Science and Technology, Vol 8(12), DOI: 10.17485/ijst/2015/v8i12/58385, June 2015.
[4] Jyoti Soni Ujma Ansari Dipesh Sharma "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" international Journal of Computer Applications, Volume 17– No.8, March 2011
[5] Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", International Conference on Computer Research and Development, ISBN: 978-1-61284-840-2,2011
[6]Himigiri. Danapana, M. Sumender Roy,‖ Effective Data Mining Association Rules for Heart Disease Prediction System‖ IJCST Vol. 2, Issue 4, Oct . - Dec. 2011.
[7]Fariba Shadabi and Dharmendra Sharma,‖ Artificial Intelligence and Data Mining Techniques in Medicine – Success Stories‖ International Conference on BioMedical Engineering and Informatics- 2008.
[8] J. Liu, Y.-T. HSU, and C.-L. Hung, "Development of Evolutionary Data Mining Algorithms and their Applications to Cardiac Disease Diagnosis," in WCCI 2012 IEEE World Congress on Computational Intelligence, 2012, pp. 10–15.
[9] P. Chandra, M. . Jabbar, and B. . Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection," in 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 628– 634.
[10] S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), 2013, no. Ict, pp. 1227– 1231.
[11]Zhao, Q., Rezaei, M., Chen, H., Franti, and P.: Keyword clustering for automatic categorization. Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, (2012).
[12]Michael Pucher, F. T. W.: Performance Evaluation of WordNet-based Semantic Relatedness Measures for Word Prediction in Conversational Speech. (2004).
[13] K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.
[14] R. Chitra and V. Seenivasagam, "REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES," Journal on Soft Computing (ICTACT), vol. 3, no. 4, pp. 605–609, 2013.
[15]. Shanta kumar, B.Patil,Y.S.Kumaraswamy, "Predictive data mining for medical diagnosis of heart disease prediction" IJCSE Vol .17, 2011
[16]. M. Anbarasi et. al. "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376 ,2010.
[17]. Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", IEEE, 2011.
[18] MA.Jabbar, Priti Chandra, B.L.Deekshatulu..:Cluster based association rule mining for heart attack prediction,JATIT,vol 32,no2,(Oct 2011)
[19] Ping Ning tan, Steinbach, vipin Kumar. : Introduction to Data Mining, Pearson Education, (2006).
[20] 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)
[21] P.S.Mishra "Optimization of the Radial Basis Function Neural Networks Using Genetic Algorithm for Stock Index Prediction", International Journal of Computer Science and Engineering Vol. 6(6), 2347-2693, 2018.
[22] K. Sivaranjani, A. Nisha Jebaseeli "Survey on Disease Diagnostic using Data Mining Techniques", International Journal of Computer Science and Engineering Vol. 6(2), 2347-2693 ,2018.

**Author Profile**

Dr.J. Adamkani received his M.Sc., degree in The New College, Chennai, India in 2003. He also received his Ph.D degree in University of Madras, India in 2017. Now he is working as a Assistant Professor with Department of Computer Science, The New College, Chennai, India. His research area is Data Mining and Network Security.

Dr.M. Wasim Raja received his M.Sc., degree in Jamal Mohamed College, Trichy, India in 2002. He also received his Ph.D degree in University of Madras, India in 2017. Now he is working as a Assistant Professor with Department of Computer Science, The New College, Chennai, India. His research area is E-Learning and Software Engineering.

 