Efficient Analytical Approaches to Predict the Human Diseases Using FP-Tree with Bagging Algorithm

Prakash S.[1*], Sangeetha K.[2] and Ramkumar N.[3]
1. IT, Sri Shakthi Institute of Engineering and Technology, Coimbatore, INDIA
2. CSE, SNS College of Technology, Coimbatore, INDIA
*prakashsce18@gmail.com

Abstract
The disease is an unhealthy condition of a human. It disturbs the all the human parts of the body, mind, metabolism and day to day activities. The diseases are classified into many types, namely chronic disease, incurable disease and terminal disease. One of the most hazardous disease which affects the human being is heart and the liver. In order to identify the specific disease the data's are collected from the hospitals. The data set provides all the relevant information about the patient. Particularly, based on the symptoms of the patient the type of the disease predicted. In the health care, Data mining technology is widely used for identifying and predicting the disease. One of the most popular technique used is Associative rule mining (ARM) is tested on the collected data set of the patient to identify the correlation among the different symptoms characteristics. In order to identify the correlation among the characteristics, the frequent item set is identified with the help of minimum support. The strength of the association is calculated based on the minimum confidence value. In order to classify the different disease, the classification algorithm is used to classify the data into different groups. There are number of classification techniques such as Nearest Neighbour, Support Vector Machine (SVM) Naive Bayes, and J48. Out of all these techniques, the bagging method and FP Tree is applied to the data set and the accuracy and efficiency is compared with other techniques.

Keywords: Association Rule Mining, Naïve Bayes, J48, Support vector machine, FP Tree and Bagging

Introduction

In the recent years, many people are affected from heart attacks and liver failures. Many researchers have been carried out in order to prevent the diseases in the early stage. The heart attack is mainly due to stress, tension, blood pressure, smoking, continuous alcohol consumption and also due to hereditary. Heart attack occurs due to the block in the blood vessels. The symptoms of the heart attack are chest pain, sweat, arm pain, vomiting and motion.

The liver failure occurs due to hepatic insufficiency. Hepatitis insufficiency is the liver will get damage and it will function abnormally. There are two different types of liver failure are categorized as acute and chronic and recently third type called acute on chronic liver failure has been recognized. The symptoms of the liver disease are vomiting, pain in the abdomen, jaundice, weakness and loss of weight. The liver failure can occur all of a sudden or it will damage gradually.

Association Rule Mining: The most popular Data mining techniques is Association Rule Mining. It is one of the techniques for analyzing the relationships between variables from large data sets. It is primarily used for find the correlations between different attributes of a given data set. In order to find the correlations the FP Tree technique is used for finding the correlation. The Authors has chosen Association rule mining because it generates the frequent item set, the frequent item set is used to find the correlation between items. When the A and B are the items in the transaction, the support value is calculated as Number of transactions in a data set contains the items A and B / Total number of transactions. The Confidence value is calculated as Number of transactions in a data set contains the items A and B / Number of transactions contain A

The minimum support value is a threshold to find the frequent item set, how frequent the different items are occurred in the set of transactions, depends on the prediction level, the support value may vary.

FP Tree: FP Tree was proposed by Han. It is a structure that stores information about mostly used patterns in a collection of data. It is defined as it has root named null which has sub trees as children, mostly used data are represented in the form of a table. Every node has prefix sub tree and it has three fields, they are item, count and node. FP tree takes lesser time to identify the frequent itemset when compared to the traditional Apriori algorithm, because FP Tree scans the dataset only once while finding the frequent itemset whereas other algorithm scans the data set upto 'N' times to generate the frequent itemset

Literature Survey

The heart disease is one of the major impact that leads to death. The doctors are finding difficult to predict the heart problem as it requires more experience. There are several techniques and algorithms have been introduced in the latest information technology sector. In order to make decisions data mining provides many techniques such as J48, Naive Bayes, REPTREE, CART and Bayes Network are used for finding the heart disease. [1]

The authors proposed that the data mining concepts such as Association rule mining and classification re use for processing the huge number of data sets. The results of the

data set is direct and it is easy for the user to understand. These techniques are widely useful in medical diagnosis. It is also used for analyzing the data so as to predict the type of disease the patient is going to be affected. Based on the analysis, the doctors will take wide decisions. Many classification methods such as CPAR, MMAC, CAR, and CBE CMAR are used for segregating the characteristics from the data set. [2]

The heart disease diagnosing is one of the major concern in the data analysis field. Many research works is going on in the medical expert system. The technique called Efficient Nearest Neighbor (ENN). It is one of the most popular algorithms for pattern recognition. In ENN approach, the data sets are divided based on the characteristics of the neighbor. [3, 4]

The techniques such as Associative Rule Mining along with the classification gives the better result. In terms of prediction, classifiers play a major technique for analyzing the data for different applications. In the present and the past world, heart disease is considered as a most hazardous disease and it will lead to death. The author uses ARM along with the genetic algorithm are used for predicting the heart disease. [5]

There are many data mining algorithms to predict the heart disease. Most common data sets in all the research papers consist of attributes of person age, sex, blood test. The attributes such as the diabetes level, history of a patient, alcohol consumption, smoking, tobacco consumption, obesity were not used for analyzing purposes. The author uses Naïve Bayes algorithm for analyzing the given data set. It is highly used for the physicians for predicting the heart problem. [6]

The support vector machine the data can be analyzed the Hepatitis level of the patient from the data set. The drawback of SVM are the collection of new data cannot be assured, in order to find the hidden relations, from the unknown data, the wrapper method is introduced. The wrapper method is used to clear all irrelevant records and provides the better result. Rapid Minor is a tool used for collecting all the patient test information from the clinical data in the hospital. [7]

The classification technique is widely used in the field of medical diagnosis. This technique helps to find out the relations in order to model the prediction network. With the help of Chi-Square, the characteristics will improve the classification technique. The machine learning algorithms are used analyzing the Hepatitis patients. [8]

The ARM is used to find the relationships between the characteristics and it is a very effective technique compared to other techniques. The heart attack is one of the hazardous disease and it leads to death. In the recent technology era, it is difficult to estimate the probability of patients getting heart attack. The data mining concept is widely used to identify from the given data set. The authors used decision tree in order to identify the heart attack. [9]

The data analysis is widely used in medical diagnosis. Liver cancer is one of the major death causing disease. The diagnosing the liver cancer at the early stage is very difficult. Various techniques such as Decision tree, C4.5, Bayesian network, Support vector machine, KNN and neural networks are used to predict the liver cancer. [10]

The count of the death increases due to the cardio vascular disease. In order to analyze the number of people affected by heart disease is by using different techniques like ID3 and J48 to predict the with the help of the dataset. Authors used 10-fold cross validation model is used to calculate the estimated data for constructing the predicted model. [11]

The different algorithms such as C4.5, KNN, Naïve Bayes, SVM and SMO algorithms are applied to the different sets. The author proposed out of all the algorithms SMO is the one gives better accurate rate. [12]

In [13], the author said that association rule mining used for searching a interesting relationships among items is given dataset.

Methodology

FP Growth algorithm

Input:
D – Hepatitis dataset
MS – minimum support value to find frequent item set
Output:
Frequent patterns are mined
Method

1. Scan D, find the repeated 1 item set and support count F. Sort F in support count descending order and call it as L1
2. Create FP tree root with null and insert every transaction into the tree
3. Find the possible path to reach a particular item in a tree
4. Count how many times other item count is increased because of a given item
5. Find the frequent item set from the conditional pattern tree, which satisfies the minimum support count
6. Repeat step 3 to 5 until all the items are mined

Bagging Method: Bagging method is used for increasing the classification accuracy. It uses the combination of various classification models and combines a series of k learned models.

Bagging is a classification method which increases the accuracy level in identifying the class label attribute. Bagging method consists of a medical dataset and it is categorized into three different parts M1, M2, M3,….Mk. The medical dataset consists of the patient information and the test result. The output of every classification method is combined and votes of all the algorithm are combined. The new data samples are combined along with the votes to

provide better outcome. Bagging is one of the efficient algorithm than other techniques

For predicting the appropriate disease and improve the accuracy, the proposed method is applied. From the dataset D, the FP pattern algorithm is applied to find the association from the set of attributes. Then for classifying the attributes bagging method is used.

The bagging method is applied into the dataset which is a combination of different classification algorithms. The class label attribute values from each method are measured. Based on the combined votes the accurate class label is generated.

Experimental Results

The dataset is collected from the UCI data repository. Hepatitis dataset consist of 19 characteristics and 1 class label attribute. For experimentation purposes 155 instances are considered. The proposed algorithm is applied in order to identify the classification accuracy in predicting whether the person will still be dead or alive. The Association rule mining is applied to identify the correlation among all the 19 attributes. After identifying the correlations the bagging classification is applied into it.

Finally, the decision tree is constructed, which is shown in Fig 2.

The Authors considered 20 attributes to conduct the experiment they are age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen, palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin protime, HISTOLOGY, DIE. From the frequent item set which has the minimum confidence value is taken into an account to write the Association rule.

The possibility of occurrence will be given in fig.3

There are many existing algorithms are available for classification, bagging produces accurate classification. The table II and figure 4 shows the accuracy level of various algorithms.

From the collected dataset, the 11 attributes are of type Boolean attributes, the repeated usages of sets are identified with the help of FP Tree algorithm. Correlation among various attributes is identified. Association rule is formed from the frequent itemset. From the data set, it is inferred as If(Fatigue=No)&&(Spider – No)&&(Bilirubin >1) && (Age>40) &&(Sex=male) =>(condition = DIE)

The time taken to classify is less in the association rule mining compared to the traditional Apriori method, because the FP Tree algorithm minimizes the number of scan in the dataset. The accuracy rate in predicting the class label attribute is 76%. In order to improve the accuracy, the bagging algorithm is applied to the collected dataset. In the bagging algorithm the decision tree induction and back propagation algorithm are applied in parallel. All the attributes in the dataset are contributing towards identifying the appropriate class label attribute. The output is collected from the different algorithms and it is combined as a vote and the decision is made based on the bagging methodology. It is observed that 87% accuracy is produced by bagging method.

Conclusion

The data mining algorithms are widely used in the medical diagnosis field for the predicting the disease. These techniques are very much useful for the doctors to identify the disease if they are not able to predict based on the experiences. The algorithms such as Naïve Bayes, J48, Nearest Neighbor, Support vector machine are tested from the collected data set and analysis has been performed in order to identify the accuracy of among all the four techniques. The Naïve Bayes has 73% accuracy level, J48 has 71% accuracy level, Nearest Neighbor has 62% accuracy level and Support vector machine has 78% accuracy level. With the comparison of the existing methods, the new technique called FP tree and Bagging has been applied to the data set and it is analyzed. It is found that the FP Tree gives 81% accuracy level and Bagging method gives 87% accuracy. On comparing the above traditional data mining technique, the new proposed system produces 10% accuracy level. With help of the proposed method it is also analyzed that person will die or alive based on the different parameters.

References

1. Hlaudi Daniel Masethe, Mosima Anna Masethe " Prediction of Heart Disease using Classification Algorithms" Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA

2. D. Sasirekha and A. Punitha "A Comprehensive Analysis on Associative Classification in Medical Datasets" Indian Journal of Science and Technology, Vol 8(33), DOI: 10.17485/ijst/2015/v8i33/80081, December 2015

3. K.Jayavani and Kadhar Nawaz "Optimal Data Prediction and Classification Applicable for Intelligent Heart Disease Diagnosis System" International Journal of Computational Intelligence and Informatics, Vol. 5: No. 2, September 2015

4. M.Akhil jabbar B.L Deekshatulu Priti Chandra ""Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm" International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013 page 85-94

5. M.Akhil jabbar , Dr.Priti Chandra , Dr.B.L Deekshatulu ""Heart Disease Prediction System using Associative Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012

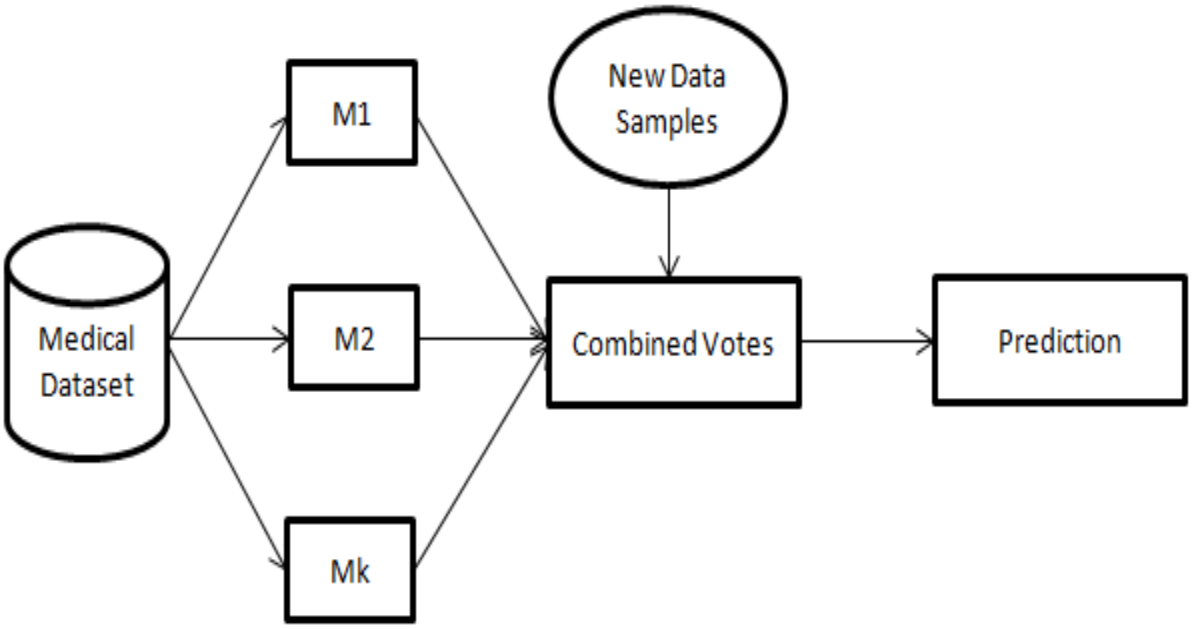

Fig. 1: Bagging method

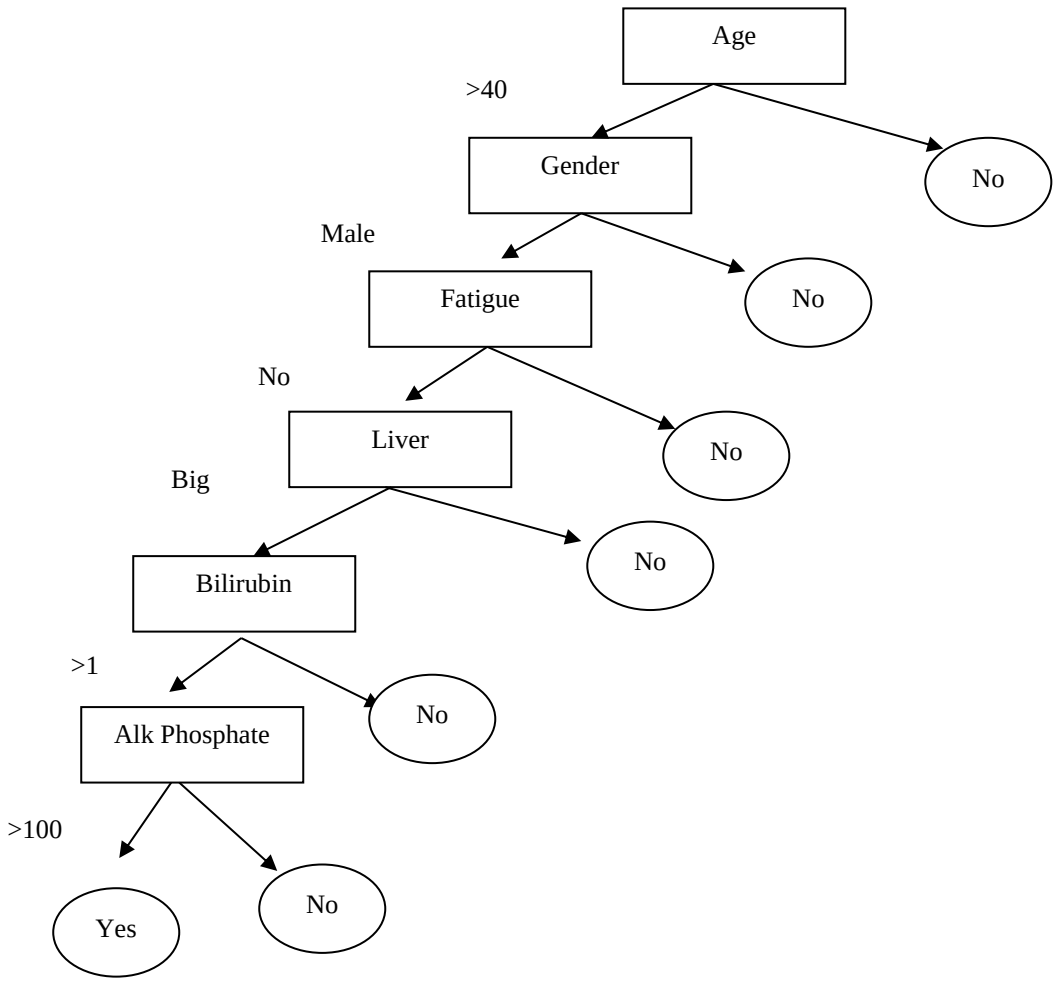

Fig. 2: Decision tree construction for medical dataset to predict the disease

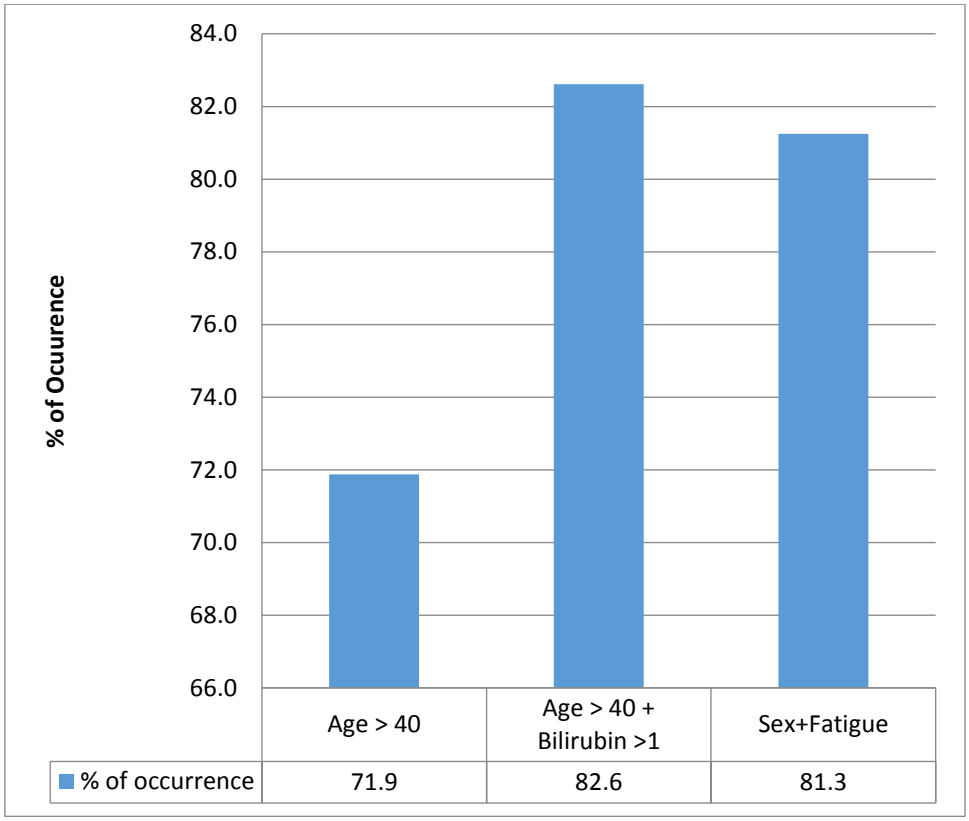

Figure 3: Possibility of disease occurrence

Table I
Example training data

S.N.	AGE	SEX	STEROID	ANTIVIRALS	FATIGUE	MALAISE	ANOREXIA	LIVER BIG	LIVER FIRM	SPLEEN PALPABLE	SPIDERS	ASCITES	VARICES	BILIRUBIN	ALK PHOSPHATE	SGOT	ALBUMIN	PROTIME	HISTOLOGY	DIE
1	30	2	1	2	2	2	2	1	2	2	2	2	2	1.0	85	18	4	?	1	2
2	41	1	1	1	1	1	2	2	1	2	2	2	2	2.3	280	98	3.8	40	1	1
3	35	1	2	2	1	2	2	2	2	2	2	2	2	0.9	58	92	4.3	73	1	2
4	40	1	2	2	1	2	2	2	2	2	1	2	2	0.6	67	28	4.2	?	1	1
5	57	1	2	2	1	1	1	2	2	2	1	1	2	4.1	?	48	2.6	73	1	1
6	47	1	2	2	1	1	2	2	1	2	2	1	1	1.7	86	20	2.1	46	2	1

Table II
Accuracy % comparison

Name of the Algorithm	Navie Bays	J48	Nearest Neighbour	Support Vector Machine	Proposed FP Tree	Proposed Bagging
Accuracy in %	73	71	62	78	81	87

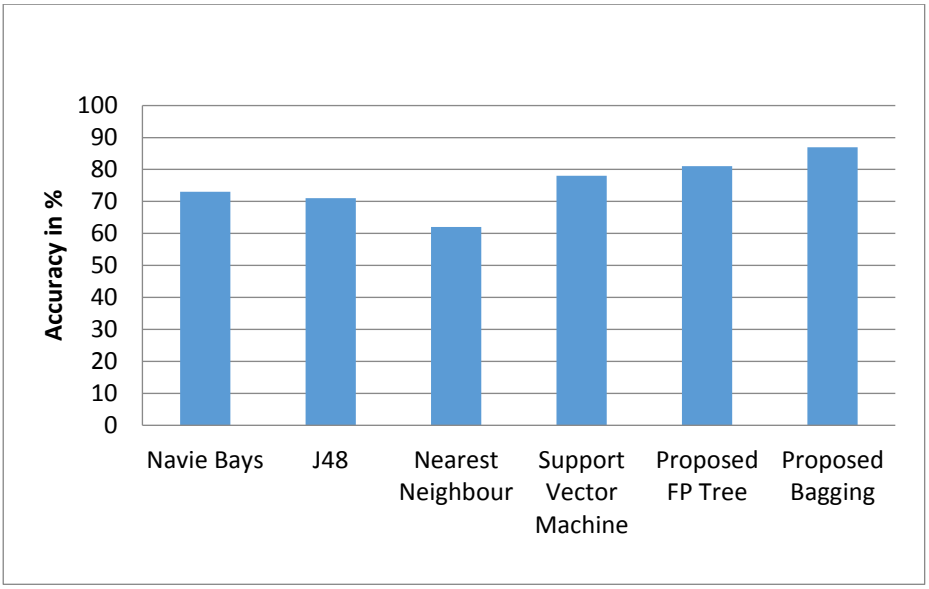

Figure 4: Accuracy Comparison

6. Shinde S. B., Amrit Priyadarshi "Diagnosis of Heart Disease Using Data Mining Technique" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

7. C. Barath Kumar1, M. Varun Kumar2, T. Gayathri3, S. Rajesh Kumar "Data Analysis and Prediction of Hepatitis Using Support Vector Machine (SVM)" C. Barath Kumar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2235-2237

8. Varun kumar, Vijay Sharathi and Gayathri devi "Hepatitis Prediction Model based on Data Mining Algorithm and Optimal Feature Selection to Improve Predictive Accuracy" International Journal of Computer Applications (0975 – 8887) Volume 51– No.19, August 2012

9. Zarna Parekh, Avaniba Parma "An Approach for Early Diagnosis of CardioVascular Disease Using Modified Decision Tree" COMPUSOFT, An international journal of advanced computer technology, 4 (5), May-2015 (Volume-IV, Issue-V)

10. Reetu1, Narender Kumar "Medical Diagnosis for Liver Cancer using Classification Techniques" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

11. Jyoti Rohilla , Preeti Gulia "Analysis of Data Mining Techniques for Diagnosing Heart Disease" International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 7, July 2015

12. Chetana Yadav, and Shrikant Lade "A Survey on Data Mining Techniques for the Diagnosis of Coronary Artery Disease" International Journal of Advanced Research in Computer Science and Software Engineering" Volume 3, Issue 10, October 2013

13. Sree Subhasini and Bhakyalakshmi "Parallel mining of frequent itemsets using map reduce and fidoop" International Journal of Innovations in Scientific and Engineering Research , Volume 3 ,Issue 11, pp. 94-97, Nov 2016