International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July2016

# Expert Medical Systems

**Mr Deepak P, Prof. Hemanth S R**

*P G Scholar, Department of Computer Science and Engineering, Maharaja Institute of Technology (MIT), Mysore, Karnataka, India*

*Professor, Department of Computer Science and Engineering, Maharaja Institute of Technology (MIT), Mysore, Karnataka, India*

***Abstract***: The Objective of the project is to design an expert system that predicts medical diseases like heart and diabetes diseases with reduced number of attributes. Medical data nowadays is available in abundance but without proper usage. The goal is to turn data that are facts, numbers, or text which can be processed by a computer into information and knowledge. Using data mining techniques on medical data several critical issues can be understood better and dealt with starting from studying risk factors of several diseases to identification of the diseases occurring frequently or taking care of hospital information. The main aim is to analyze the uniqueness of medical data mining using an Expert Medical System. They have potential in clinical laboratory reporting by supporting automation of the process, improving accuracy and consistency, and enhancing the quality of work. Classification of knowledge objects is a knowledge mining and knowledge management process used in grouping similar knowledge objects together. There are plenty of classification algorithms available in literature but decision tree is the most often used because of its ease of implementation and simpler to understand, when compared to other classification algorithms. There are many classifiers but we have used C4.5 for more accuracy and less run time. The decision tree algorithm has been applied on the knowledge of heart and diabetes disease to foretell whether disease is present or not.

***Keywords***:Machine Learning, Classification, C4.5 Algorithm**.**

## I. INTRODUCTION

Expert Medical Systems are the programs designed to solve problems at a level comparable to that of a human expert in a given domain or a computer system that operates by applying an inference mechanism to a body of specialist expertise represented in the form of 'knowledge' which is an automated process. An application of Artificial Intelligence is Expert System which can be used to make the machine to learn the frameworks to create the information which could be communicated as basic controls or as a choice Tree.

It includes a knowledge acquisition component which processes the data and information into rules and use knowledge about the diseases and facts about the patients to suggest diagnosis. It consists of five major components: *Knowledge base* which contains the specific knowledge related to the area of application of the system and *Database* acts as a working memory and contains current facts or past data. A *Rule Base* that supports the work with the knowledge to obtain a diagnosis. An *Explanation Component* making the user knows how the system arrived to suggested diagnosis and the *User Interface* for adequate communication between consulting patients and the system.

C4.5 decision tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. Medicine has formed a rich test-bed for machine learning experiments in the past, allowing scientists to develop complex and powerful learning systems. While there

has been much practical use of expert systems in routine clinical settings, at present machine learning systems still seem to be used in a more experimental way. Machine learning systems can be used to develop the knowledge bases used by expert systems. C4.5 Algorithm provides more accuracy and takes less run time compared to others and the decision tree can be used to tell whether the disease is present or not and the simulation result enables us to predict the patterns and relationships from the model for performance.

## II. RELATED WORK

In Rule-based expert system, system takes some facts from the real world and makes decision based on those facts. But the efficiency of the system is decrease when size of problem domain is large and for this purpose, to improve the efficiency and accuracy we use classification technique in expert system. Data classification process contains two steps. 1. **Learning step** which is constructed using training dataset. 2. **Classification step**: Constructed model is used to classify the testing dataset. In expert system, the knowledge is represented in terms of if-then rules. Knowledge base is the main component and is a collection of rules which are produce by the knowledge engineer through knowledge acquisition. Knowledge is acquired from the knowledge expert. User interface provide the input for the system. After getting the input from the user the Inference engine is come into the work. Inference engine takes input from the user interface and then perform the matching with the predefined rules which is store in the knowledge base. When it finds the perfect match then it takes decision as described by the system. For this process of rule matching temporary working memory is used. At last the decision made by the system is transfer to the user or client using user interface. **Decision tree algorithm**: Decision tree is a structure which is a flowchart which perform specific test on an attribute which represented as an internal node of tree and leaf node represents the class label. Tree contains only one root node which indicates the starting of tree. The accuracy of the decision tree algorithm is good. C4.5 is the most efficient and useful algorithm used for decision tree-based approach. In this algorithm initially dataset is sorted according to attribute value. This procedure is continued until all the attributes are classified. When node of a tree contains same attribute records then that node is consider as a leaf node as represent the class label. Decision tree is performing well when size of a dataset is large. [1]

Comparison of Data mining techniques in classification is to find the best technique for creating risk prediction model of heart disease at minimum effort. There are two types of model used in analysis of data. First one is applying single model to various heart data and another one is applying combined model to the data and the combined model is known as hybrid model. This paper provides a quick and easy understanding of various prediction models in data mining and helps to find best mode. The objective of classification is to predict accurately the target class for all case in the data. Association Rule for classification of Heart-attack patients was proposed. First stage involved the data warehouse pre processed to make the mining work more efficient. The extraction of significant patterns from the heart disease data warehouse was presented. The Association Rule used to pre process in order to handle missing values and applied equal interval with approximate values based on medical data. Research shows that comparing the effectiveness of the three popular classification algorithms namely C4.5, ID3 and CART to classify disease dataset. It is more useful in medical research to construct algorithms for disease classification and prediction. The observation shows that C4.5 and

CART performs better in terms of accuracy. There is a requirement to have reliable model for predicting the existence or absence of heart disease with known and unknown risk factors. [2]

The architecture of our proposed method is designed based on how a medical doctor concluded related to indication that someone has the potential against DM, which is the model has been adjusted with the data and the approach to early prediction of the Data Mining with a combination of two techniques of computational intelligence in the form of fuzzy logic and artificial neural network and case-based reasoning for knowledge engineering techniques. To improve the accuracy of prediction results on each approach, they implemented a rule-based algorithm to assist in classifying and predicting specific data. They conducted a study to perform a comparison between the performance of the three methods, namely artificial neural network prediction, modeling decision tree and logistic regression, to predict DM using the 12 risk factors and the result suggested that the decision tree modeling produces the highest level of classification accuracy. [3]

Fuzzy expert system is a group of membership functions and rules. Fuzzy expert systems are tilting toward numerical processing. This paper recapitulates the essential distinction between the Mamdani-type and Sugeno-type fuzzy expert systems by using the input parameters such as age, obesity, RBS(Random Blood Sugar), family history and diet. The MATLAB fuzzy logic toolbox is used for the imitation of both the models. Based on the membership function of input and output values the diabetes disease is concluded. From the results it is concluded that the accuracy in case of sugeno-type fuzzy expert system is quite more than Mamdani-type fuzzy expert system. [4]

Expert systems have provided solutions to different problems from strategic planning of marketing to consulting in process reengineering. In general, the majority of studies published are based on advanced techniques of artificial intelligence using specific languages or tools that require certain knowledge of reasoning processes to model information. With the arrival of web-based expert systems that can connect to the Internet has made it easy to access information from any place at any time, creating new requirements for web systems. In this research, a tool is proposed for development of web based expert systems and utilizes Semantic Web technology which permits the knowledge engineer and domain expert to define the knowledge without having to know anything about programming languages and AI.The proposed tool enables the knowledge engineer to insert and update the domain knowledge such as Facts and Rules. The tool can induce new rules based on the semantic concepts and relations. Using Semantic Web technology supports the tool to utilize the ontology as knowledge formalization. Most representation mechanisms must provide support for three aspects of knowledge conceptual representation, relational representation, and uncertainty representation. Four schemes are commonly used for knowledge representation such as an Inference engine stands between the user and knowledge base. The inference engine, including Inference and Control, performs two major tasks. First, it examines existing facts and rules and adds new facts when possible. Second, it decides the order in which inferences are made. In doing so, the inference engine conducts the consultation with the user. It uses the information in the working memory along with the rules in the knowledge base to derive the conclusion. [5]

Diagnosis is a complicated and an important task that needs to be executed accurately and efficiently. The diagnosis is made based on doctor's experience & knowledge which leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an Automatic medical diagnosis system would be beneficial. In the dataset provided where the number of attributes which were used for heart disease diagnosis were reduced from 13 attributes to 6 using Genetic Algorithm and Feature Subset Selection. This paper uses the Classification Modeling Techniques such as Decision Trees, Naive Bayes and Neural Network, along with weighted association Apriori algorithm and MAFIA algorithm in Heart Disease Prediction.Mining frequent itemsets is an active area in data mining that aims at searching interesting relationships between items in databases. The proposed approach utilizes an efficient algorithm called MAFIA (Maximal Frequent Itemset Algorithm) which combines diverse old and new algorithmic ideas to form a practical algorithm.The proposed algorithm is employed for the extraction of association rules from the clustered dataset performing efficiently when the database consists of huge dataset. After mining the frequent patterns using MAFIA algorithm, the significance weightage of each pattern is calculated. It is calculated based on the weightage of each attribute present in the pattern and the frequency of each pattern. Among the various modeling techniques, Decision Tree has outperformed with 99.62% accuracy by using attributes and the accuracy of the Decision Tree. [6]

The purpose of Decision Support Systems is to make rules and plan to manage basic symptom knowledge to face challenge using data processing techniques and mathematical modeling tools. The concept of a fuzzy set has better knowledge representation to improve the decision making process. The knowledge-base of systems is composed of structured and concise representation of the knowledge of domain experts. The structure knowledge belongs to facts, rules and events of different type of diabetes, which were commonly agreed upon by experts in the field of medicine. Fuzzy inference is the process of mapping from a given input to an output using the theory of fuzzy sets. The core of decision making output is a process by the inference engine using the rules contained in rule base. The fuzzy inference mechanism used here is a MamdaniInference . The fuzzy inference engine utilize the rules in the knowledge-base and derives conclusion base on the rules. Inference engine uses a forward chaining mechanism to search the knowledge for the symptoms of various type of diabetes. The selection of a particular methodology also depends upon some external parameters such as the cost of system, efficiency required, and amount of data available and the sensitivity of the system. Clinical decision systems enhance the quality of healthcare services and by control the cost-effectiveness of medical examinations and treatment. [7]

Data mining research on heart disease diagnosis and prediction uses the data mining techniques to enhance heart disease diagnosis and prediction including decision trees, Naive Bayes classifiers, K-nearest neighbour classification (KNN), support vector machine (SVM), and artificial neural networks techniques. Results show that SVM and neural networks perform positively high to predict the presence of coronary heart diseases (CHD). Decision trees after features reduction is the best recommended classifier to diagnose cardiovascular disease (CVD). It is found that decision trees and Naive bayes classifiers are recommended for CVD diagnosis with an accuracy

reaching more than 95%. Classifiers such as C5, SVM, and neural networks are the best recommended for CHD prediction. It is observed that the prediction results of various data mining classification techniques are strongly encouraging and would assist the physicians to do early diagnosis and make more accurate decisions. But the performance of data mining techniques to detect coronary arteries diseases (CAD) is between 60 - 75% and further improvement should be carried for more accuracy. [8]

Rule based reasoning techniques are used to identify and then takes decision for the disease. The objectives of the system are initial classification of the 11 parameters inputted by the user. Rules were in the form of "if- then" and fuzzification will be done using associated membership functions and perform aggregation if needed. Match the classified inputted parameter with rules and identify the maximum degree of occurrence of result and then defuzzify the result. It provided the severity of the heart disease to the user on the basis of the result. Designing the system with fuzzy base in comparison with designed improves the results and more efficient. [9]

Data mining (DM) is the core stage of knowledge discovery in databases (KDD) which applies machine learning and statistical methods in order to discover areas of previously unknown knowledge. KDD involves the following steps: data selection, data pre-processing, transformation and interpretation of results. In the past decades, data mining have played an important role in heart disease research for to find the hidden medical information from the different expression between the healthy and the heart disease individuals in the clinical data are a powerful approach in the study of heart disease classification. Statistics and machine learning are two main approaches which have been applied to predict the status of heart disease based on the clinical data. One of the most common classification models is the decision tree, which is a tree-like structure where each internal node denotes a test on a predictive attribute and each branch denotes an attribute value. A leaf node is used to predict class distributions.The decision tree type used is the gain ratio decision tree. The gain ratio decision tree is based on the entropy (information gain) approach, which selects the splitting attribute that minimizes the value of entropy, thus maximizing the information gain. Information gain is the difference between the original information content and the amount of information needed. The features are ranked by the information gains and the top ranked features are chosen as the potential attributes used in the classifier. To identify the splitting attribute of the decision tree, one must calculate the information gain for each attribute and then select the attribute that maximizes the information gain. The result obtained that which shows that fasting blood sugar is the most important attribute which gives better classification against the other attribute and they develop a prediction model using J48 decision tree for classifying heart disease based on the clinical features against unpruned, pruned and pruned with reduced error pruning approach. The accuracy of Pruned J48 Decision Tree with Reduced Error Pruning Approach is more better then the simple Pruned and Unpruned approach. [10]

Objective of the work is to predict more accurately the presence of heart disease with reduced number of attributes. Three classifiers like Naive Bayes, J48 Decision Tree and Bagging algorithm are used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. Bagging plays an important role in the field of medical diagnosis and used to improve model

stability and accuracy. Boosting can be used with any type of model and can reduce variance and bias in predictions. Bagging means Bootstrap aggregation an ensemble method to classify the data with good accuracy. J48 Decision tree with reduced error. It requires no domain knowledge or parameter setting and can handle high dimensional data. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes. Weka is a collection of machine learning algorithms for data mining tasks.The classify panel enables the user to apply classification and regression algorithms to the resulting dataset to estimate the accuracy of the resulting predictive model and to visualize erroneous predictions. The predictive models provide a way to predict whether a patient having heart disease or not. Execution of the learning techniques is highly dependent on the nature of the training data. To evaluate the robustness of classifier, the usual methodology is to perform cross validation on the classifier and investigation on the data experiments are conducted to find the best classifier for predicting the diagnosis of heart disease patients and the results show that bagging algorithm accuracy of 85.03% and the total time taken to build the model is at 0.05 seconds in the diagnosis of heart disease patients.[11]

Data mining, an application is medical areas to form decision support systems for diagnosis by inventing meaningful information from given medical data. In this study a decision support system for diagnosis of illness that make use of data mining and three different artificial intelligence classifier algorithms namely Multilayer Perceptron, Naive Bayes Classifier and J.48. Data mining represents a significant advance in the type of analytical tools currently available and used as a valid, sensitive and reliable method to discover patterns and relationships. Benefits of data mining in medical analysis are to increase diagnostic accuracy, to reduce costs and to reduce human resources. The data set used is Diabetes set. WEKA **(Waikato Environment for Knowledge Analysis)** software package is used in the study. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. The decision trees generated by C4.5 can be used for classification and for this reason; it is referred as a statistical classifier. The Naive Bayes Classifier technique is mainly suited when the dimensionality of the inputs is high and the results prediction accuracy ratios of 75.13%, 73.82% and 76.30% for MLP, J48, and Naive Bayes respectively. [12]

The primary goal was to design and develop a model and design efficient approach for detection of heart disease, which can be utilized for real world applications as a computer aided diagnostic tool. Identification of information and valuing it for decision making from a large collection of data has been increased recently which is an interactive and iterative process encompassing several subtasks and decisions and is known as Knowledge Discovery from Data. The data sets were trained to carry out with the aid of back propagation techniques. Whenever unknown data was fed by the doctor, the system identified the unknown data from comparisons with the trained data and generates a list of probable diseases that the patient may vulnerable using machine learning algorithms. Artificial Neural Network technique is used to get the fitness value of any chromosomes; weights are extracted from that chromosome which is a reverse process of chromosome formulation. With the available inputs and above defined weights, architecture of neural network formed and output

generated. From the output, error value is defined. Reciprocal of this error is considered as fitness value of that particular chromosome. Most of the developed solution utilized the feed forward architecture and back propagation as a learning algorithm. Because of trapping tendency in local minima, problem may appear at the time of up gradation and in result no consistency. To overcome this problem a new way, Genetic algorithm has been applied for training purpose. The untold and unseen side of trained data selection is to be discovered which gives a new label to understand the selection method. [13]

The data classification is based on supervised machine learning algorithms which result in accuracy and time taken to build the algorithm and tool named Tanagra is used to compare the performance accuracy of data mining algorithms for diagnosis of heart disease dataset. It contains tools for data classification, statistics, clustering, supervised learning and visualization and uses two learning performance evaluators such as first it splits a dataset into training and test data, while the second performs cross-validation using folds. Evaluation is described by accuracy, error, precision. They have used a Bayes classifier, a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions and it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature and uses the method of maximum likelihood. Naive Bayes algorithm is the best compact time for processing dataset and shows better performance in accuracy of 52.33% and prediction in 609ms.[14]

## III. PROPOSED METHODOLOGY

**Existing System**

Existing system was designed to predictdiseases like heart and diabetes was time consuming, accuracy was very less and used redundant parameters which are not required in classification. To build a decision tree which gone take very less time to predict the accuracy of existing system to predict heart and diabetes is low which was 50-80ms and the accuracy was 73-80%.

**Proposed system**

Implementation of the proposed system is done usingMVCArchitecture (Model-View-Controller).Model is concerned with the database, view is concerned with the forms and Controllers are used to communicate between model and view. C4.5 algorithm is used to improve the existing system which uses more optimal attributes that are more important in classification of proposed system which gives c4.5 better accuracy and reduced run time.Proposed system has the following advantages over the existing system:

- Provide answers for decisions, processes and tasks that are repetitive.
- Hold huge amounts of information.
- Minimize employee training costs.
- Centralize the decision making process.
- Make things more efficient by reducing the time needed to solve problems.
- Combine various human expert intelligences.
- Reduce the number of human errors.

**Algorithm: C4.5**

1. Check for base cases
2. For each attribute a
   - Find the normalized inform
3. Let a_best be the attribute with the hig
4. Create a choice node that split on a_be
5. Recurse on the sub lists obtained by sp
node.

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3algorithm.C4.5 algorithm and procedural steps are explained below.

**Procedural steps:**

**a)** In general, if we are given a probability distribution P = (p1, p2, .., pn) then the *Information conveyed by this distribution*, also called *the Entropy of P*, is:

I(P) = -(p1*log(p1) + p2*log(p2) + .. + pn*log(pn))----------1

For example, if P is (0.5, 0.5) then I(P) is 1, if P is (0.67, 0.33) then I(P) is 0.92.

b) Calculate information Entropy for each attribute. This is used to calculate the gain.

Consider the quantity Gain(X,T) defined as

Gain(X,T) = Info(T) - Info(X,T)----------2

Equation 2 represents difference between information needed to identify an elementof T and information needed to identify an element T after the value of attribute Xhas been obtained, that is, this is the gain in information due to attribute X.

**c)** For all the attribute calculate the quantity Gain(X,T) for each Attribute. Theattribute with maximum gain is used to split the decision tree. Repeat from (a).

## IV. CONCLUSION

In medical diagnosis various data mining techniques are available. In this study, for classification of medical data we employed c4.5 algorithm because it is more accurate and takes less time and it also produce human readable classification rules which are easy to interpret. The result shows that our c4.5 algorithm accuracy is 86% and time taken to process the design is 116sec in predicting heart infection patients and 84% accurate and time taken to process the design is 64sec in analysis of diabetes infected patients. The Results are shown in the form of Decision trees.

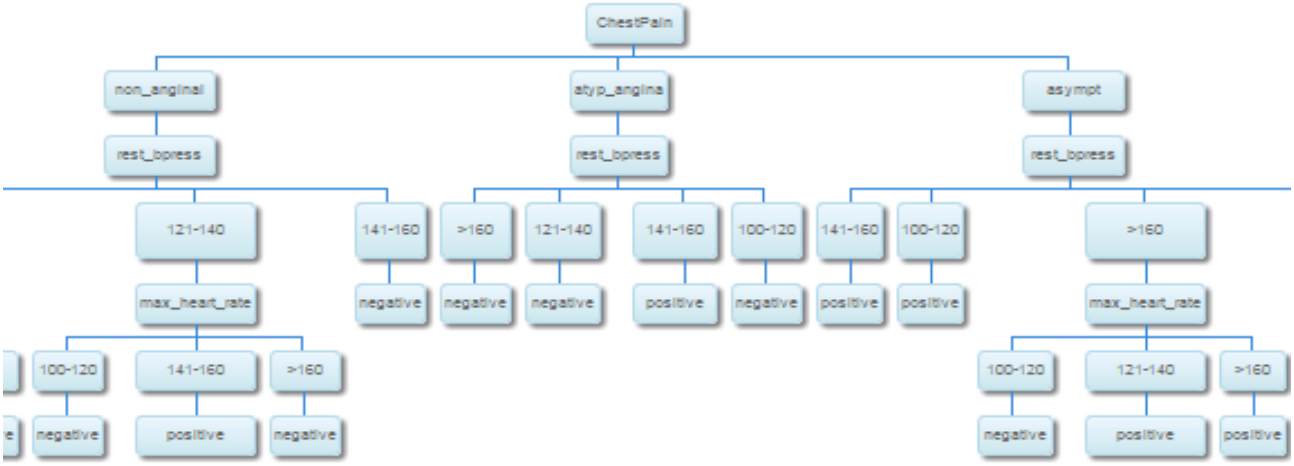


Fig 1 Decision tree obtained for Heart Disease

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July2016

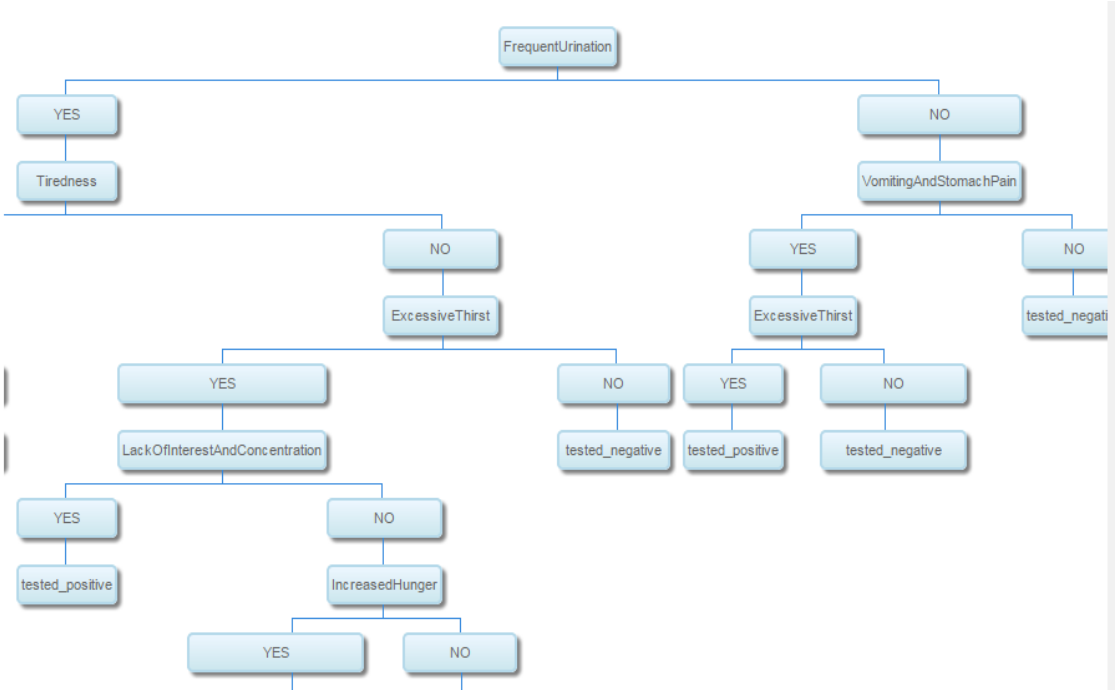


Fig 2 Decision tree obtained for Diabetes Diseases

REFERENCES

[1].Jatin Patel, Nikita D Patel, Nikita S Patel "A Research on Expert System using Decision Tree and Naive Bays Classifier", IJCSMC, Vol. 4, Issue.5, pg.341 – 348, May 2015.

[2]. P. Krishnakumari and G.Purusothaman "A Survey of Data Mining Techniques on Risk Prediction of Heart Disease", *Vol8(12),June 2015*.

[3]. Rian Budi Lukmantoa, Irwansyah. E "The Early Detection of Diabetes Mellitus Using Fuzzy Hierarchical Model", 2015

[4]. VishaliBhandari, Rajeev Kumar" Comparative Analysis of Fuzzy Expert Systems for Diabetic Diagnosis" in IJCA (0975 – 8887) Volume 132 – No.6, December 2015

[5].MostafaNofal, Khaled M. Fouad "Developing Web-Based Semantic Expert Systems", International Journal of Computer Science Issues, Vol. 11, Issue 1, No 1, January 2014

[6]. AdityaMethaila, Prince Kansal, HimanshuArya, PankajKumar"Early Heart Disease Prediction Using Data Mining Techniques", 2014

[7]. Nagendra Singh Rana, D.B.V. Singh, Piyush Mishra , Shailendrasengar" Clinical Decision Support System for Diabetes Disease Diagnosis", 2014

[8]. Salha M. Alzahani, AfnanAlthopity, AshwagAlghamdi, BoushraAlshehri and SuheerAljuaid "An Overview of data mining techniques applied for Heart disease diagnosis and prediction", 2014

[9]. M. EswaraRao and Dr. S. GovindaRao"Expert System for Heart Problems", 2014

[10]. "A Heart Disease Prediction Model using Decision Tree", 2013 by

K.L. Jaiswal, PrabhatPandey, Atul Kumar Pandey ,Ashish Kumar Sen.

[11]. "Data Mining Approach to Detect Heart Diseases", 2013 by VikasChaurasia, Saurabh Pal.

[12]. Murat Koklu and YavuzUnal" Analysis of a Population of Diabetic Patients Databases with Classifiers", 2013

[13]. K S Kavitha , K V Ramakrishnan, M K Singh "Modeling and design of evolutionary neural network for heart disease detection", 2010

[14].AshaRajkumar, Mrs. G.SophiaReena "Diagnosis of heart diseases using Data mining Algorithm", 2010