ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Prediction of Heart Diseases using Data Mining Techniques

K.Manimekalai

Assistant Professor, Department of BCA, Sri GVG Visalakshi College for Women, Udumalpet, India

ABSTRACT: Heart Disease is a major disease all over the world. In medical science, prediction of heart disease is very important. Data Mining Techniques are used to find the heart diseases of patients. Based on the risk factors the heart diseases can predict easily. Risk factors are also helpful to find out the heart disease by the medical experts as well as the patients. The objective of this paper is to study the various data mining Techniques available to predict the heart disease and find the best method of prediction.

KEYWORDS: Data Mining, Heart Disease, SVM classifier with Genetic algorithm, Naïve Bayesian, C 5.0, Neural Network, KNN, J4.8 Decision Tree, fuzzy mechanism.

I. INTRODUCTION

In the modern society most of the persons affecting heart diseases. In this fast growing world people want to live a luxurious life. In order to earn a lot of money and live a comfortable life they work like a machine. They never concentrate the organs of their body. Day by day the risk factors are increased for heart disease. Depends upon their working condition they change their food habits. It leads to blood pressure, sugar in young age.

Most of the time people never consult a medical practitioner. They are taking their own medicine. This type of method leads a side effects. At present, millions of people suffer from heart disease annually. Heart disease is a major cause of morbidity and mortality. According to the World Health Organization, 12 million deaths are caused by heart diseases in the world annually, 50 percent of which can be prevented by controlling risk factors. Heart diseases are expected to be the main reason for 35 to 60 percent of total deaths expected worldwide by 2025.

Data mining is the process of extracting hidden knowledge from data. It can reveal the patterns and relationships among large amount of data in a single or several datasets. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. In other words Data mining is one of the steps of knowledge discovery for extracting implicit patterns from vast, incomplete and noisy data. It is a field with the confluences of various disciplines that has brought statistical analysis, machine learning techniques, artificial intelligence and database management systems together to address the issues.

II. HEART DISEASE

Life is itself on efficient working of heart. The heart is important part of human body. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it.

A number of factors have been shown that increases the risk of Heart disease:

- Family history
- Smoking
- Poor diet
- High blood pressure
- High blood cholesterol
- Obesity
- Physical inactivity
- Hyper tension

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient's current test results & doctor's experience. Thus the diagnosis is a complex task that requires much experience & high skill.

Types of Heart diseases

Heart disease is a broad term that includes all types of diseases affecting different components of the heart. Heart means 'cardio.' Therefore, all heart diseases belong to the category of cardiovascular diseases. Some types of Heart diseases are

Coronary heart disease

It also known as coronary artery disease (CAD), it is the most common type of heart disease across the world. It is a condition in which plaque deposits block the coronary blood vessels leading to a reduced supply of blood and oxygen to the heart.

Congestive heart failure

It is a condition where the heart cannot pump enough blood to the rest of the body. It is commonly known as heart failure.

Cardiomyopathy

It is the weakening of the heart muscle or a change in the structure of the muscle due to inadequate heart pumping. Some of the common causes of cardiomyopathy are hypertension, alcohol consumption, viral infections, and genetic defects.

Congenital heart disease

It also known as congenital heart defect, it refers to the formation of an abnormal heart due to a defect in the structure of the heart or its functioning. It is also a type of congenital disease.

Arrhythmias

It is associated with a disorder in the rhythmic movement of the heartbeat. The heartbeat can be slow, fast, or irregular. These abnormal heartbeats are caused by a short circuit in the heart's electrical system.

Myocarditis

It is an inflammation of the heart muscle usually caused by viral, fungal, and bacterial infections affecting the heart. It is an uncommon disease with few symptoms like joins pain, leg swelling or fever that cannot be directly related to the heart.

III. LITERATURE SURVEY

Numerous studies have been done that have focus on diagnosis of heart disease. They have used different attributes and applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

Rajwant Kau et al[1] provided a study of different data mining techniques that many risk factors that cause heart disease and it is very difficult to understand and categorized. Most of time Heart Diseases are detected when a patient reaches at last stage of disease. The Risk Factors help to analyse the disease in advance. They collected 50 patients database and used SVM Classifier with Genetic Algorithms.

Moloud Adbar et al[2] applied and compared data mining techniques to predict the risk of heart diseases. They applied five algorithms including c5.0, Neural Network, SVM, KNN and Logistic Regression.

B.V. Baiju and R.J.Remy Janet[3] used continuous data instead of categorical data. They prefer Naïve Bayes or Bayes Rule for the following reasons.

- When the data is high
- When the attributes are independent of each other
- When they want more efficient output, as compared to other methods output.

The technique Naïve Bayes is used to find the result.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Another study conducted by Yangquan Lyu et al [4] has been based on the evaluation model of coronary artery disease by using data mining algorithm. In this research a new dynamic model, which makes it possible to assess lifetime, suggests linear time-invarient approach to assess CHD.

In the paper presented by K.Rajeswari et al[5] examined the heart disease using Neural Network. They studied the influence feature selection for neural network algorithm in identifying patients with Ischemic heart disease.

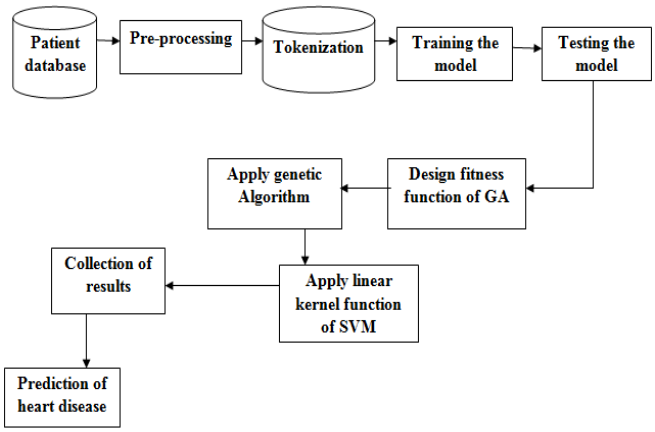
The work done by S.Suganya et al[6], for finding minimum distance CART classifier was incorporated to classify the data among various groups.

K.Manimekalai [31] explained that some attributes are mainly responsible for heart disease. In heart disease prediction system, the input parameters play a major role for efficient prediction. Generally many attributes are involved for efficient prediction. In order to predict the disease based upon patient's raw symptom description, different input parameters can be collected. To help the physicians and healthcare professionals in the prediction of heart disease the number of attributes should standard.

IV. EXISTING TECHNIQUES

i) SVM weight optimization by Genetic Algorithm

Support Vector Machine weight optimization by Genetic Algorithm. This system uses Linear Kernel function for learning and training the SVM Classifier. But the results produced by this were not much better. To get more accurate results the genetic algorithm is used with SVM as optimizer. For this, the fitness function of Genetic Algorithm is used as input for SVM Classifier. The Linear Kernel function is used to Train the dataset using the weights optimized by Genetic Algorithm.



ii) C5.0 Algorithm

C5.0 algorithm developed from C4.5 algorithm is one of the most important and widely used algorithms in data mining. C4.5 itself is the extended form of ID3 algorithm. C5.0 has the ability to be applied for classifying as a decision tree or a set of rules. Because of the understandability of their rules set, they are preferred in many applications. The strength of the algorithm is in handling missing values or its large number of entries, as well as the fact that less time is necessary to learn it [14-17]. If S is training set and X contains n attributes so that the set S is divided into N sub categories: The algorithm to test the features makes use of element is called the gain ratio [18]. The number of samples in the S is displayed in (S1, S2, S3,....Sn). For calculating the number of samples that belong to Ci (the value Parameter i is [i = 1,2,3,4, ..., N]) is used in the following formula: freq. Also for calculate an instance belonging the Ci is used to the formula: freq (Ci, S) / S

iii) Neural Network Algorithm

Artificial Neural Network is a data processing algorithm, originated from human brain. The system includes a large number of tiny processors to handle data processing. The processors act in the form of an interconnected network parallel to each other to solve a problem. Using programming knowledge, in this networks a data structure is designed that can act as neurons. This data structure is called the neuron [8-11].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

iv) SVM Algorithm

Support Vector Machine (SVM) is a regulatory algorithm introduced by Vapnik in 1995. The base of the algorithm is using the precision to generalize the errors. The algorithm makes "hyperplane" and divides the data into classes so that all samples belonging to one class will be categorized on one side and the rest on the other side. Linear SVM Classifier is defined for the SVM classifying task, and dividing them occurs provided that the chosen line involves the most marginalized sure [19], [20].

v) Neural Network Algorithm

Artificial Neural Network is a data processing algorithm, originated from human brain. The system includes a large number of tiny processors to handle data processing. The processors act in the form of an interconnected network parallel to each other to solve a problem. Using programming knowledge, in this networks a data structure is designed that can act as neurons. This data structure is called the neuron [21-24].

vi) Naive Bayes Algorithm

Naïve Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. A naive Bayes classifier is a term dealing with a simple probabilistic classification based on applying Bayes theorem. In simple terms, a naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It learns from the “evidence” by calculating the correlation between the target (i.e., dent) and other (i.e., independent) variables. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Here independent variables are considered for the purpose of prediction or occurrence of the event. The algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. For example, a patient may bed to have certain symptoms. Based on the observation, Baye’s theorem can be used to compute the probability that a proposed diagnosis is correct. Baye’s Theorem finds the probability of an even occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes theorem can be stated as follows, P (B given A) = Prob(A and B) / Prob (A) To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur
P (Ci/X)>P (Cj/X) for all 1< = j< = m and j!= i

vii) J48 DecisionTree Classifier

Decision tree is a kind of classifying and predicting data mining technology, belonging to inductive learning and supervised knowledge mining technology. As decision tree is advantageous in fast construction and generating easy-to-interpret If-Then decision rule, it has become the most widely applied technique among numerous classification methods [27]. Decision tree algorithm has been applied in many medical tasks, for examples, in increasing quality of dermatologic diagnosis [28], predicting essential hypertension [29], and predicting cardiovascular disease [30]. Decision tree is one of the most popular tools for classification and prediction. Production of a decision tree is an efficient method for classification of data. This tree using a top-down strategy to build a test on each node. J48 decision tree method is the implementation of c4.5 decision tree in weka data mining tool. J48 decision tree supports continuous and discrete features. It can also manage features with missing value.

viii) Fuzzy Mechanism

Crisp input values are transferred into fuzzy values in the stage of fuzzification [25]. The Fuzzy values are taken into the Generating rules for Advanced Fuzzy Resolution Mechanism.

ix) CART Classifier

The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be. The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.
The main elements of CART are:

a) Rules for splitting data at a node based on the value of one variable.

b) Stopping rules for deciding when a branch is terminal and can be split no more; and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

c) Finally, a prediction for the target variable in each terminal node.

V. PROPOSED PREDICTION SYSTEM

In this paper 9 techniques are discussed. It is shown in the table 1. S.Suganya, P.Tamil Selvy [6] used CART Classifier Technique. Another author Rajwant Kaur and Sukhpreet Kaur [1] combined the technique SVM Classifier with Genetic Algorithms. Moloud Adbar, Sharareh R.Niakan Kalhori, Tole Sutikno, Imam Much Ibnu Subroto, goli Arji [2] presented the experimental results and used various data mining techniques like C5.0, Neural Network, Support Vector Machine, KNN. B.V.Baiju and R.J.Remy Janet [3] followed Naïve Baysesian classification Technique. Yongquiag L, Jiaming H, Yiran w, Jijiang Y, Yida T, Wenyao W, Nazim[4] used Neural Network. A.V.Senthilkumar [7] used Fuzzy Mechanism technique. K.Rajeswari, V.Vaithiyanathan and T.R. Neelakanttan [5] presented the experimental results and used Neural Network technique.

Table 1 : Comparison of Techniques

Year	Author	Techniques Used
2016	S.Suganya, P.Tamil Selvy[6]	CART Classifier
2015	Rajwant Kaur and Sukhpreet Kaur[1]	SVM Classifier with Genetic Algorithms
2015	Moloud Adbar, Sharareh R.Niakan Kalhori, Tole Sutikno, Imam Much Ibnu Subroto, goli Arji[2]	C5.0, NN, SVM, KNN
2015	B.V.Baiju and R.J.Remy Janet[3]	NaïveBaysesian classification Technique
2015	Yongquiang L, Jiaming H, Yiran w, Jijiang Y, Yida T, Wenyao W, Nazim[4]	Neural Network
2013	A.V.Senthilkumar [7]	Fuzzy Mechanism
2012	K.Rajeswari, V.Vaithiyanathan, T.R. Neelakanttan[5]	Neural Network

Most of the authors explained individual techniques. The technique Neural Network is repeated twice.

VI. RESULTS AND DISCUSSION

Comparison of accuracy is shown in the table 2. When CART Classifier Technique is used 83% accuracy is achieved. The technique SVM Classifier is combined with Genetic Algorithms the accuracy of 93.02%; It is very high. The various data mining techniques like C5.0, Neural Network, Support Vector Machine, KNN are used and got the accuracy of 93.02%, 89.4%, 86.05%, 80.23% respectively. When Naïve Baysesian classification Technique and obtained 81% accuracy. When the Neural Network is used and the accuracy is 88.37%. The technique Fuzzy Mechanism is used and got 94.11% accuracy. When Neural Network technique and got 94% accuracy.

The technique Neural Network is repeatedly used by various authors. Data mining techniques applied in Medical field provides innovative results and decision support system used to improve the health of patients and for other medical services. But still it needs to improve the system to predict probable complications in advance. When the data mining technique SVM Classifier with Genetic Algorithm is used it has the highest accuracy 95%. When KNN Technique is used it has the lowest accuracy 80.23%.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

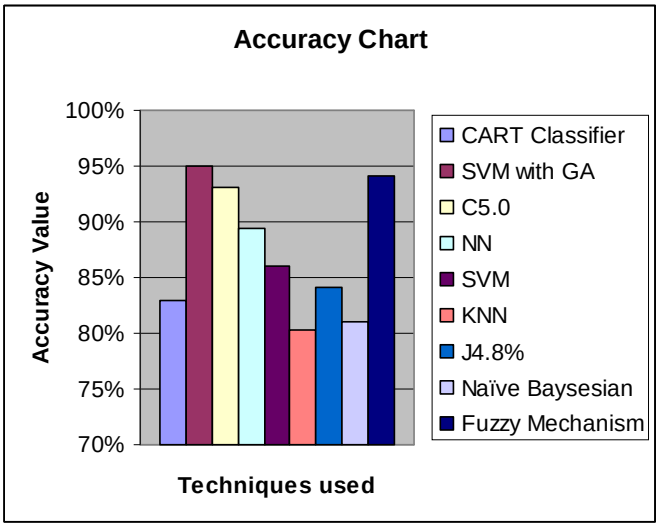
Vol. 4, Issue 2, February 2016

Table2: Comparison of Accuracy

Techniques Used	Accuracy
CART Classifier	**83%**
SVM Classifier with Genetic Algorithms	**95%**
C5.0	**93.02%**
Neural Network	**89.4%**
Support Vector Machine	**86.05%**
KNN	**80.23%**
J4.8%	**84.1%**
Naïve Baysesian classification Technique	**81%**
Fuzzy Mechanism	**94.11%**

Chart 1: Accuracy Chart for various Data Mining Algorithms

The chart is drawn for the data mining techniques and their accuracy value. SVM with Genetic algorithm has a high accuracy of 95% and the lowest accuracy is KNN.



VII. CONCLUSION

The objective of this work is to provide a study of Heart Diseases using various Data Mining Techniques. When the data mining technique is used separately the accuracy is low. To improve the accuracy value, data mining techniques should be combined together. SVM Classifier with genetic Algorithm contains 95% high accuracy than other technique. In future the techniques are hybrid, the accuracy will high. It will definitely help the patients as well as the medical practitioners to predict the heart disease.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

REFERENCES

[1] Rajwant Kaur, Sukhpreet Kaur, "Prediction of Heart disease Based on Risk Factors Using Genetic SVM Classifier", IJARCSSE Volume 5, Issue 12, December 2015.
[2] Moloud Adbar, Sharareh R.Niakan Kalhori, tole Sutikno, Imam Much Ibnu Subroto, Goli Arji, "Comparing Performance of Data Mining algorithms in Prediction Heart Diseases", IJECE, Vol 5, No 6, December 2015.
[3] B.V. Baiju and R.J.Remy Janet, "A survey on heart disease Diagnosis and Prediction using Naïve Bayes in Data Mining", IJCET, Vol 5, No.2, April 2015.
[4] Yongquiang L, Jiaming H, Yiran w, Jijiang Y, Yida T, Wenyao W, Nazim A. Dynamic evaluation model of coronary heart disease for ubiquitous healthcare. Computer Industry, 2015; 69:35-44.
[5] K.Rajeswari, V.Vaithiyanathan, T.R. Neelakanttan , "Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks. International Symposium on Robotics and Intelligent Sensors 2012, Procedia Engineering, 2012;41:1818-1823.
[6] S.Suganya, P.Tamil Selvy, " A Proficient Heart Disease Prediction Method Using Fuzzy-Cart Algorithm, "IJSEAS, Volume2, Issue1, January2016.
[7] A.V Senthil Kumar. Generating Rules for Advanced Fuzzy Resolution Mechanism to Diagnosis Heart Disease. International Journal of Computer Applications, 2013; 77(11): 6-12.
[8] Yazdani A, Ebrahimi T, Hoffmann U. Classification of EEG signals using Dempster Shafer theory and a K-nearest neighbor classifier. IEEE. In: Proc of the 4th int EMBS conf on neural engineering, 2009: 327-30.
[9] Jenn-LongLiu, Yu-Tzu, Chih-Lung Hung, "Development of evolutionary data mining algorithm and their applications to cardiac disease diagnosis", World congress on computational intelligence, 2012.
[10] Daubechies I. The wavelet transform, time-frequency localization and signal analysis. IEEE. Trans Inform Theory 1990; 36: 961-1005.
[11] Demuth H, Beale M, Hagan M. Neural network Toolbox™ user's guide. The MathWorks, Inc.; 2009.
[12] Leng, G., McGinnity, T.M., Prasad, G. Design for self-organizing fuzzy neural networks based on genetic algorithms. IEEE. Trans. Fuzzy Syst. 2006; 14 (6): 755-766.
[13] Frank H. F. Leung, H. K. Lam, S. H. Ling, Peter K. S. Tam. Tuning of the structure and parameters of a neural network using an improved genetic algorithm. IEEE. Trans. Neural Networks, 2003; 14(1):79-88.
[14] Quinlan J R. Induction of decision trees. Machine Learning, 1986; 4:81-106.
[15] Quinlan J R. C4.5: Programs for machine learning. Machine Learning, 1994; 3:235-240.
[16] Quinlan J R. Bagging, Boosting and C4.5. Proceedings of 14th National Conference on Artificial Intelligence, 1996: 725-730.
[17] Xindong W , Vipin K , J. Ross Q , Joydeep Gh, Qiang Y, Hiroshi M , Geoffrey J. M, Angus Ng, Bing L, Philip S. Yu, Zhi-Hua Z, Michael S, David JH, Dan S. Top 10 algorithms in data mining. Springer,2008;14(1):1-37.
[18] Shuonan H, Rongtao H, Xinming S, Jun W, Chengshang Y, Research on C5.0 Algorithm Improvement and the Test in Lightning Disaster Statistics", International Journal of Control and Automation,vol.7,no1,pp.181-190,2014.
[19] Sumit B, Praveen P, G.N. Pillai. SVM Based Decision Support System for Heart Disease Classification with Integer- Coded Genetic Algorithm to Select Critical Features. WCECS. Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, October 22 - 24, 2008
[20] Vapnik, V. N. The nature of statistical learning theory. New York: Springer, 1995.
[21] Daubechies I. The wavelet transform, time-frequency localization and signal analysis. IEEE. Trans Inform Theor,1990; 36: 961-1005.
[22] Demuth H, Beale M, Hagan M. Neural network Toolbox™ user's guide.TheMathWorks,Inc.;2009.
[23] Leng, G., McGinnity, T.M., Prasad, G. Design for self-organizing fuzzy neural networks based on genetic algorithms. IEEE. Trans. FuzzySyst.2006;14(6):755-766.
[24] Frank H. F. Leung, H. K. Lam, S. H. Ling, Peter K. S. Tam . Tuning of the structure and parameters of a neural network using an improved genetic algorithm. IEEE. Trans. Neural Networks, 2003;14(1):79-88.
[25] Mehdi Fasanghari, Gholam Ali Montazer,"Design and implementation of fuzzy expert system for Tehran Stock Exchange portfolio recommendation", Expert Systems with Applications 37 pp.6138–6147, 2010.
[26] Boshra Bahrami, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", JMEST, Vol. 2 Issue 2, February - 2015
[27] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. "Discovering data mining: From concept to implementation,". New Jersey: Prentice Hall, 1997.
[28] Chang, C.-L., and ChenC.-H."Applying decision tree and neural network to increase quality of dermatologic diagnosis,"Expert Systems with Applications, 4035-4041.,2009.
[29] Ture, M., Kurt, I., Kurum, A. T., and Ozdamar, K,"Comparing classification techniques for predicting essential hypertension,".Expert Systems with Applications, 2005.
[30] Eom, J.-H., Kim, S.-C., and Zhang, B.-T. "AptaCDSS-E: A classifier ensemble based clinical decision support system for cardiovascular disease level prediction," Expert Systems with Applications, 2008
[31] Manimekalai.K "A Review on Prediction of Heart Diseases by Comparing Risk Factors in Data Mining", IJCSIT, Vol.7(1),2016,396-398.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

BIOGRAPHY

Manimekalai K is working as a Head & Assistant professor of Computer Applications in Sri GVG Visalakshi College for Women (Affiliated to Bharathiyar University), Udumalpet, Tamilnadu, India. She is interested in Medical datamining. She has 11 years of experience. She received her B.Sc in Physics Degree from the University of Bharathiyar in 1998. She did her MCA in Bharathidasan University in 2003. She had completed M.Phil in Bharathidasan University in 2007. She has published a journal on "A Review on Prediction of Heart Diseases by Comparing Risk Factors in Data Mining", IJCSIT, Feb 2016.