ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 1 , January 2017

Deep Learning Approach for Prediction of Heart Disease Using Data mining Classification Algorithm Deep Belief Network

Dr. T. Karthikeyan, V.A.Kanimozhi

Associate Professor, Dept. of Computer Science, PSG College of Arts & Science, Coimbatore, India
Research Scholar, Department of Computer Science, PSG College of Arts & Science, Coimbatore, India,

ABSTRACT: Data mining is an iterative and interactive process of discovering novel, valid, useful, comprehensive and understandable patterns and models in enormous data sources. The challenge faced by healthcare industry with regard to the massive data-rich but information-poor collection is to extraction of valuable information. Medical data mining is used to extract hidden potentials for effective decision making process. This proposed Heart Disease Prediction System utilizes Deep Belief Network classification algorithm to predict the likely chances of heart related diseases of the user. Deep Belief Network is one of the proficient classification algorithm which employs Deep Learning approach in Deep Neural Network. This proposed work contains comparison of Convolutional Neural Network [CNN] and Deep Belief Network classification [DBN] algorithms. Convolutional Neural Network algorithm is one of the unsupervised algorithm. It provides 82% of accuracy in the prediction of heart diseases. But the proposed Deep Belief Network algorithm provides 90% accuracy in heart diseases prediction which enhances the prediction accuracy of heart disease prediction system. It is designed in the MATLAB 8.1 development environment.

KEYWORDS: Heart Disease Prediction System, Data Mining techniques, Medical Data Mining, Deep Neural Network, Deep Belief Network classification algorithm,

I. INTRODUCTION

A. Data Mining : Data Mining is the non-trivial process of identifying valid, novel, potential, useful and ultimately understandable patterns in data [1]. It brings a set of tools and techniques that can be applied to the processed data to discover hidden patterns. According to a broad view of data mining functionality, data mining is viewed as the process of discovering interesting patterns and knowledge from large amounts of data [6].

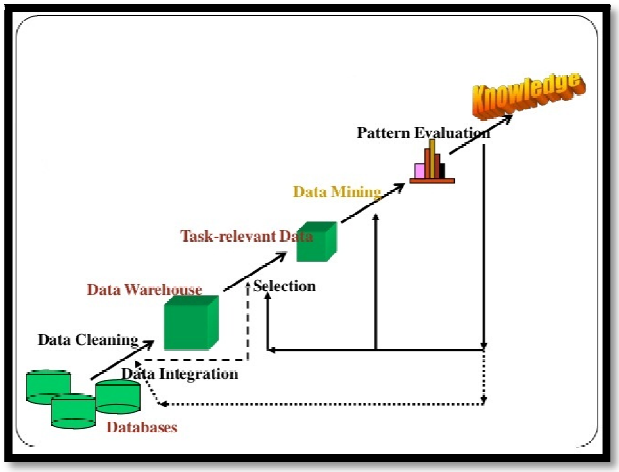

Fig. 1. Steps in Knowledge Discovery Process

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

B. Medical Data mining: Data mining in medical helps to extract useful knowledge and provide scientific decision-making for the diagnosis and treatment of disease from the medical database. Applying data mining techniques in medical data to extract meaningful patterns and knowledge is called Medical Data mining [5].The medical information has characteristics of redundancy, multi-attribution, incompletion and closely related with time. The problem of effectively utilizing the massive volumes of data is becoming a major problem for healthcare industry. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. [2]

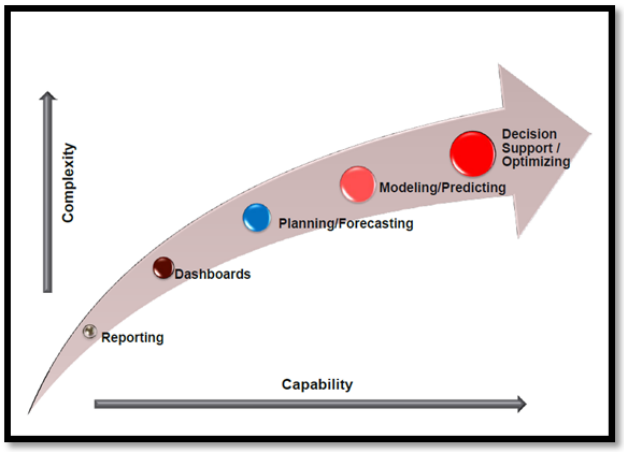

Fig. 2. Medical Data Mining Spectrum

II. LITERATURE REVIEW

Table 1. Various data mining techniques used for heart disease prediction:

Author	Purpose	Techniques Used	Tool	Accuracy
S.Florence, N.G.Bhuvaneswari Amma, G.Annapoorani, K.Malathi [22]	This system is used to predict the heart attack and also discussed various uses of various data mining algorithm for disease prediction.	Convolutional Neural Network and Decision Tree	Rapid Miner	82%
Milan Kumari [18]	The objective is to analyze various Data mining techniques on cardiovascular disease dataset.	Decision Tree	Weka3.6.6	79.05%
		Neural Networks		80.06%
		SVM		84.12%
Abhishek Taneja [11]	The purpose is using various data mining techniques an attempt to assist in the diagnosis of the heart disease.	Naive Bayes	Weka 3.6.4	86.53%
		Decision tree		89%
		Neural Networks		85.53%
Paulo Cortez[25]	It presents the use of data mining algorithms, in classification and regression tasks.	Neural Network	R tool	86%
		SVM		81%
Velu C. M.et.al[26]	The main objective is Diagnosis of Heart Disease Using Multiple Kohenen Self Organizing Maps	SVM	Orange	97.5%
		KSOM		99.1%
Nilakshi P. Waghulde[17]	The focus is propose a Genetic Neural Approach for Heart Disease Prediction	Genetic-Neural Network	Matlab	98%

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

III. AN OUTLINE OF HEART DISEASE

The Heart is a hollow, muscular organ that pumps blood, rich in oxygen and nutrients, to all part of the body. A heart attack occurs when one or more coronary arteries that supply blood to your heart muscle become blocked off. Medically, it is referred to as a myocardial infarction. [8]

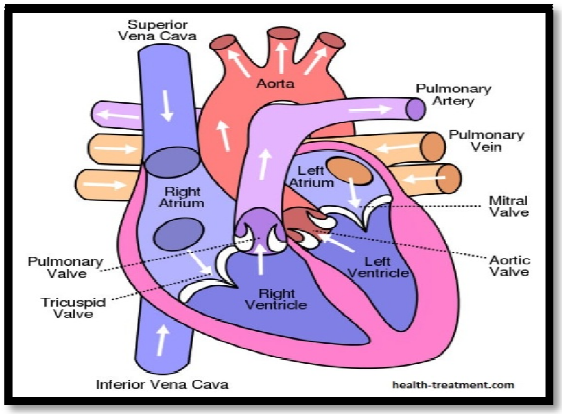

Fig. 3. Human Heart Structure

A. Heart Attack Risk Factors: [8]

Some risk factors can't be controlled: Family history, Increasing age, Ethnicity, Being male. But other risk factors can be prevented or controlled. These include: Smoking, Diabetes, High cholesterol, High blood pressure, Not being physically active, Being overweight or obese.

IV. PROPOSED HEART DISEASE PREDICTION SYSTEM

A. Data Source: A Dataset can often viewed as a collection of data objects. These data objects are described by a no.of attributes that capture the basic characteristics of an object. A total of 300 records with 16 medical attributes (factors) were obtained from the UCI Machine Learning Data Repository [7]. Following are the 16 main used in proposed work.

Table 1.2 Shows Arrtributes:

Attributes	**Range**
Age	>=18
Gender	
Blood pressure (mm Hg)	90/60 to 150/90
Cholesterol(LDL) (mg)	100 to 159
Heredity	YES/NO
Blood sugar (mg/dl)	80-120
PQ value	21.6-76
ST value	9.89-27.63
QT value	22.022-154
QRS value	8.532-68
R value	104-324
Heart beat rate	60-100
BMI	MEN<27 and WOMEN<25
Smoking habit	YES/NO
Alcohol Intake	YES/NO
Mental stress	YES/NO

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

V. PROPOSED MODEL

Proposed Model for Heart Disease Prediction System is created using Deep learning approach. The Flowchart for proposed work is shown in Fig.4.

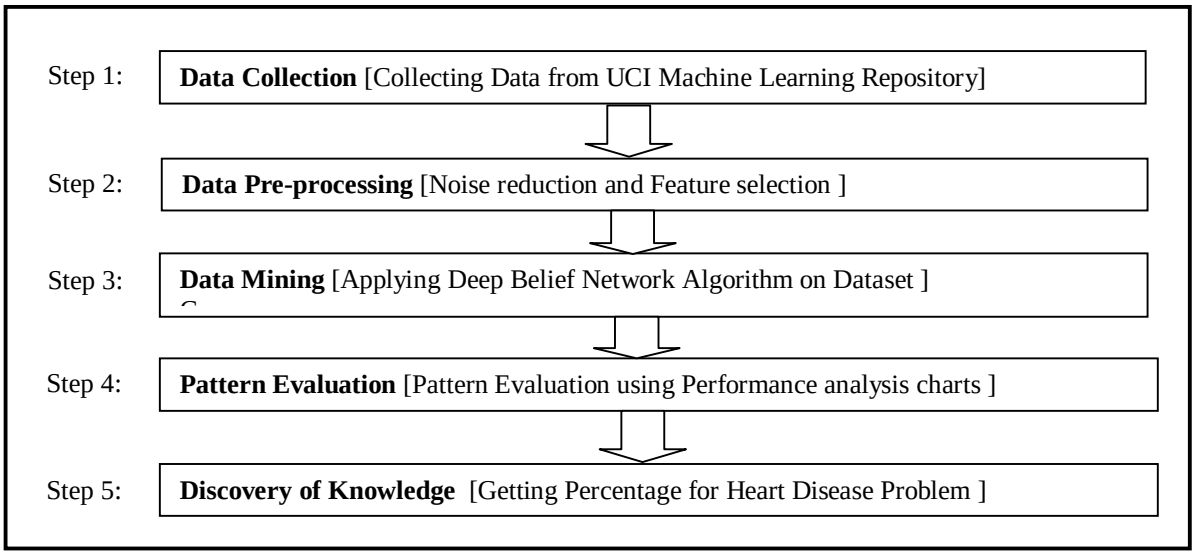

Fig. 4. Flowchart for Proposed Work

A. DEEP LEARNING - Deep learning is a subfield of machine learning that is based on learning multiple levels of representation and abstraction. Deep learning contains computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. Deep Neural Network [DNN] is a neural network with several layers of nodes between input and output. The series of layers between input and output do feature identification and processing in a series of stages. Deep learning methods aim at learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower level features.

B. DEEP BELIEF NETWORKS (DBN) - Deep Belief Networks are graphical models which learn to extract a deep hierarchical representation of the training data. It is a generative graphical model, of multiple layers of latent variables, with connections between the layers but not between units within each layer. When trained on a set of examples in an unsupervised way, Deep Belief Network can learn to probabilistically reconstruct its inputs. In contrast to perception and back propagation neural networks, DBN is unsupervised learning algorithm. A basic belief network is composed of layers of stochastic binary units with weighted connections. [10]

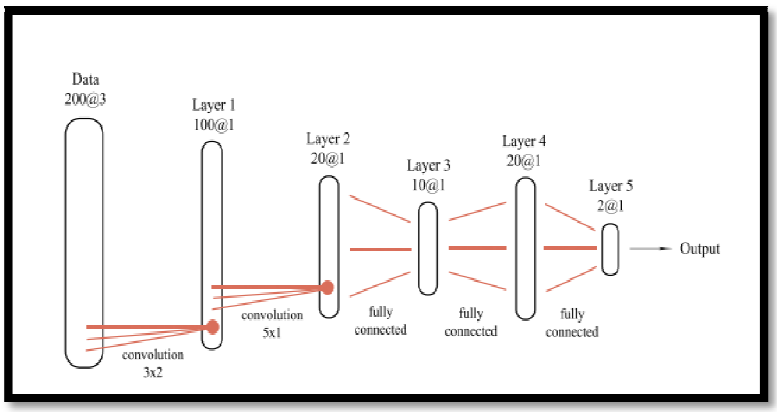

Fig. 5. Model architecture of Deep Belief Network Classifier

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 1 , January 2017

The Model architecture of Deep Belief Network Classifier is deployed in the above Fig 6. This proposed architecture has two convolutional and three fully-connected layers. the proposed architecture prevents over fitting. The last three layers are fully-connected and they perform a classification of the features obtained after the convolutional layers. The output of the final layer is fed to a softmax which produces a distribution over the class labels.

C. UNSUPERVISED TRAINING PHASES: Three basic approaches are primarily used, Those are,

- Feature selection
- Restricted Boltzmann Machines
- Autoencoders

Phase-I - Feature selection

Feature selection is the first and foremost step in the unsupervised learning phase. In this feature selection method is used to select 16 attributes from the whole datasets.

Phase-II – Restricted Boltzmann Machines and Autoencoders

These two approaches have layer wise architecture represented by the visible layers that correspond to the data and the hidden layers corresponding to the inherent features of the data. The main difference between these approaches is that in Boltzmann Machines hidden units are considered as latent random variables and in autoencoders they are considered as computational nodes.

D. DEEP LEARNING TRAINING

Step 1: Convolutional layer

In this layer, a dot product (or a convolution) of each sub region of the input data with a kernel is computed, and the result values form the output of this layer. The layer is parameterized by the size and the number of kernels, steps of the convolution in the width and height dimensions, and the activation function h applied to introduce a non-linearity to this layer.

Step 2: Max-pooling layer

This layer follows a convolutional layer and performs a down sampling operation in order to reduce the feature size. It takes small rectangular blocks of the data and produces a singular output for each block. This can be done by several ways, but the common one is taking a maximum in the block. Thus, if the block size is 2 × 2, then the number of features will be reduced by 4 times.

Step 3: Auto encoder

An Autoencoder is a symmetrical neural network that is used for unsupervised feature learning. Its training is done by minimizing the reconstruction error between the input data and its reconstruction at the output layer, and the activation values of the hidden layer are considered as a feature vector corresponding to the input data. Encoding of an input vector x ∈ R N in the autoencoder is done by applying a linear transformation and a nonlinear activation function to x.

$$\mathbf{h= \sigma(W_1x + b_1)}$$

where $W_1 \in R\ H\times N$ is a weight matrix, $b_1 \in R\ H$ is a bias, $\sigma(t)= (1 +\exp(-t))-1$ is a logistic sigmoid function, and $h \in R\ H$ represents activation values of the hidden layer. Decoding of the obtained vector h is done by another transformation with a separate weight matrix W_2 and bias b_2, Where x' is a reconstruction of the vector x.

$$\mathbf{X' = \sigma\ (W_2^T h + b_2)}$$

Step 4: Restricted Boltzmann Machine

Restricted Boltzmann Machine (RBM) is a two-layer bipartite graphical model with a set of visible units v, a set of hidden units h, and symmetrical connections between these two layers represented by a weight matrix W. The joint distribution between the hidden and visible variables is given by:

$$\mathbf{P(v,h) = \frac{1}{Z}exp(-E(v,h))}$$

$$\mathbf{E(v,h) = -h^T W_V - b^T v - c^T h}$$

where E(v, h) is an energy function, b and c are visible and hidden units biases respectively, Z is the partition function, and vi , bi ∈ {0, 1}. To fasten the learning for a RBM, contrastive divergence algorithm is used and the general idea is

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

to update all the hidden units in parallel starting with visible units, reconstruct visible units from the hidden units, and finally update the hidden units again. The Contrastive Divergence algorithm for the RBM is depicted in the following Fig 7. The learning rule is:

$$\Delta W_{i,j} = < v_i h_j >_0 - < v_i h_j > \infty$$

The simple greedy learning algorithm works for training DBN. This is because that training RBM using CD algorithm for each layer looks for the local optimum and the next stacked RBM layer takes those optimally trained values and again look for the local optimum.

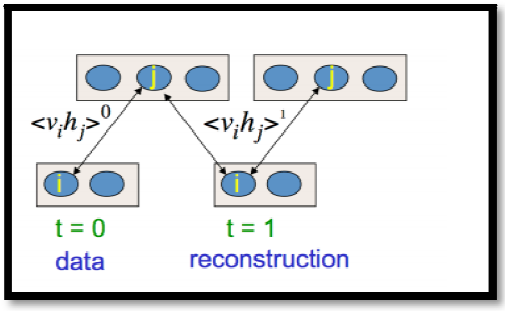

Fig. 6. Contrastive Divergence algorithm for RBM

E. DEEP LEARNING CLASSIFICATION

Finally, after several convolutional and max-pooling layers, the obtained features are converted into a single one-dimensional vector that is used for the classification. The classification layers are fully connected and usually use one output unit per class label.The gradient based learning and Deep Belief Network algorithm are useful for getting the accurate percentage level of patient's heart disease. For deep neural networks, it is important to initialize all weights to small random values. The biases may be initialized to zero or to small positive values. The iterative gradient-based optimization algorithms used to train deep networks.[22]

V. EXPERIMENTAL RESULTS

A. HEART DISEASE PREDICTION SYSTEM – GUI: MATLAB GUI result interface consists of 3 panels: Those are, The User data input panel prediction percentage display section and prediction performance display section. This MATLAB GUI environment is easy to design and easily adopt any kind of modifications. It gives an attractive GUI window for getting patient details from Users.

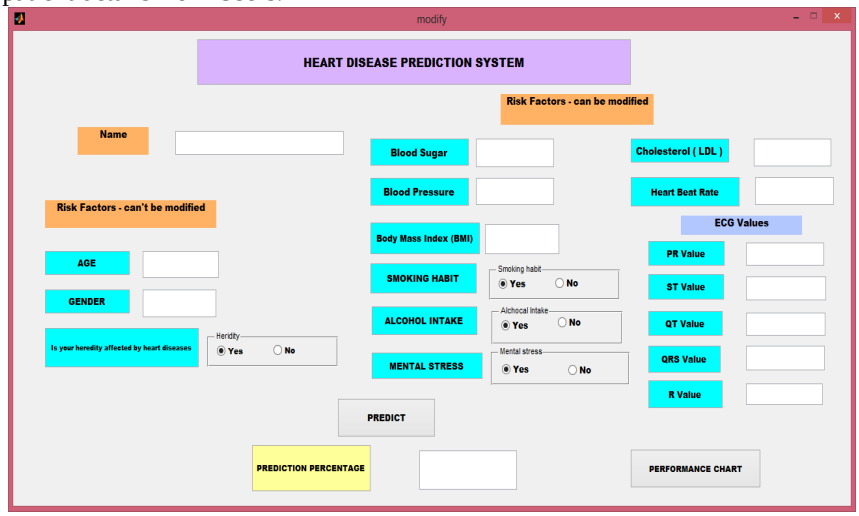

Fig. 7. Heart Disease Prediction System – User Interface

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 1 , January 2017

B. PERFORMANCE MEASURE:

Accuracy - Deep Belief Network uses unsupervised training process. So the training and test data will not mismatch there is no information loss in the dataset so the prediction percentage of the data is more accurate compare to Convolutional Neural Network.

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

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative
The True Positive (TP) is the number of data is correctly identified as relevant data. False Positive (FP) is the number of non-relevant data that are incorrectly identified, True Negative (TN) refers to the number of non-relevant data that are correctly identified as non-relevant data and False Negative (FN) is the number of data that are incorrectly identified as non-relevant data.
Accuracy of the Deep Belief Network is 90%.

C. PERFORMANCE ANALYSIS COMPARISON CHART

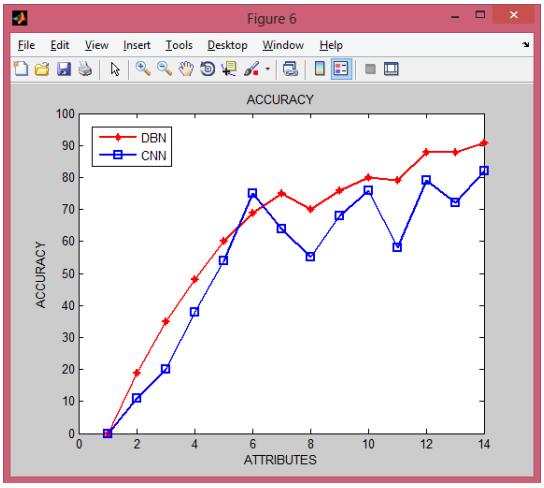

Fig. 8. Performance Analysis Chart

The above Fig 8.is a performance comparison chart for the Deep Belief Network[DBN] and the Convolutional Neural Network [CNN] classification algorithms. The X axis point outs the no.of attributes and the Y axis indicates the accuracy value of the heart disease prediction result. This chart entails that proposed system Deep Belief Network algorithm has the accuracy percentage of 90% and the Convolutional Neural Network algorithm has only 82%. It plots that the proposed Deep Belief Network algorithm outperforms compared to the Convolutional Neural Network algorithm.

VI. CONCLUSION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. It brings a set of tools and techniques that can be applied to the processed data to discover hidden patterns. Medical data mining is a promising field of Data mining is very helpful for the medical experts to analyze the medical data. This proposed model Heart Disease Prediction System with deep belief network classification algorithm introduces the deep learning approach into the data mining process for the prediction of likelihood percentage of heart disease. The proposed work makes use of 16 essential medical attributes for the prediction of heart disease. This Deep architecture has more number of hidden layers. So using this architecture we are getting accurate prediction percentage for the prediction of the heart disease. Convolutional Neural Network forms lot of unwanted structures, so it increases the complexity. It provides 82% of accuracy in the prediction of heart diseases. But the proposed system uses the deep learning approach with Deep Belief Network algorithm which has more number of hidden layers. So it gives more accurate results compared to the CNN algorithm. Accuracy of DBN classifier is 90% This proposed heart disease prediction system gives an user friendly environment and precise results.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 1 , January 2017

REFERENCES

1. Arun K Pujari , "Data Mining Techniques", University press , Edition 2001.
2. Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", *Journal of Healthcare Information Management* — Vol. 19, No. 2
3. David Crockett, Ryan Johnson and Brian Eliason, , "What is Data Mining in Healthcare?" *Health Catalyst*, 2004.
4. Jaiwei Han, Michelinne Kamber , "Data Mining : Concepts and Techniques"
5. Brian R. Hunt, Ronald L. Lipsman, Jonathan M. Rosenberg, "A guide to MATLAB for beginners and experienced users", Cambridge University Press.
6. V.A. Kanimozhi, Dr. T. Karthikeyan, "A Survey on Machine Learning Algorithms in Data mining for Prediction of Heart Disease", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 5, Issue 4, April 2016,
7. DB Statlog database: http://archive.ics.uci.edu/ml/machinelearning- databases/statlog/heart
8. Deep Learning Tutorial, Release 0.1, LISA lab, University of Montreal September 01,
9. Abhishek Taneja, "Heart Disease Prediction System Using Data Mining Techniques", *Oriental Journal Of Computer Science & Technology*, ISSN: 0974-6471 Vol. 6, No. (4), Dec 2013.
10. Yoshua Bengio , "Learning Deep Architectures for AI", Foundations and TrendsR_ in Machine Learning Vol. 2, No. 1 (2009) 1–127 _c 2009 Y. Bengio DOI: 10.1561/2200000006.
11. M. Durairaj, V. Ranjani , "Data Mining Applications In Healthcare Sector: A Study". *International Journal of Scientific & Technology Research*, volume 2, issue 10, october 2013 issn 2277-8616.
12. Nilakshi P. Waghulde1, Nilima P. Patil, " Genetic Neural Approach for Heart Disease Prediction" *International Journal of Advanced Computer Research* (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-4 Number-3 Issue-16 September-2014.
13. Milan Kumari, Sunila Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction" , *International Journal of Computer Sci ence and Technology*, IJCST Vol. 2, Iss ue 2, June 2011, I S S N : 2 2 2 9 - 4 3 3 3 (P r i n t) | I S S N : 0 9 7 6 - 8 4 9 1 (On l i n e).
14. 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.
15. Ian Goodfellow, Yoshua Bengio, Aaron Courville, "Deep Learning", w.deeplearningbook.org.
16. S.Florence, N.G.Bhuvaneswari Amma, G.Annapoorani, K.Malathi "Predicting the Risk of Heart Attacks using Neural Network and Decision Tree", *International Journal of Innovative Research in Computer and Communication Engineering*, ISSN(Online): 2320-9801 ISSN (Print): 2320-9798 Vol. 2, Issue 11, 2014 Nov.
17. Dr. T. Karthikeyan1, V.A.Kanimozhi, "A Review on Heart Disease Prediction System Using Data Mining Tools", *International Journal of Innovative Research in Computer and Communication Engineering* (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 9, September 2016, ISSN(Online): 2320-9801, ISSN (Print): 2320-9798
18. Paulo Cortez, 'Data Mining with Neural Networks and Support Vector Machines using the R/rminer Too'l, FCT grant PTDC/EIA/64541/2006.
19. Velu C. M. and Kashwan K. R., IEEE Member , 'Heart Disease Diagnosis Using Multiple Kohenen Self Organizing Maps'.
20. Vikas Chaurasia, Saurabh Pal, " Early Prediction of Heart Diseases using Data Mining Techniques", *Caribbean Journal of Science & Technology*, ISSN 0799-3757.

AUTHOR'S BIOGRAPHY

Dr. Thirunavukarasu Karthikeyan received his doctorate in Computer Science from Bharathiyar University in 2009. Presently he is working as an Associate Professor in Computer Science Department of P.S.G. College of Arts and Science, Coimbatore. His research interests are Image Coding, Medical Image Processing, Data Mining and Software Engineering. He has contributed as a program committee member for a number of international conferences. He is the review board member of various reputed journals. He is board of studies member for various autonomous institutions and universities.

Ms.Kanimozhi.V.A has completed her M.Sc degree at PSG college of Arts & Science, Coimbatore. Currently she is doing M.Phil in Computer Science at PSG College of Arts & Science. Her research interests are data mining and image processing