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 9, September 2016

# A Review on Heart Disease Prediction System Using Data Mining Tools

Dr. T. Karthikeyan[1,] V.A.Kanimozhi[2]

Associate Professor, Dept. of Computer Science, PSG College of Arts & Science, Coimbatore, India[1]

Research Scholar, Department of Computer Science, PSG College of Arts & Science, Coimbatore, India[2]

**ABSTRACT**: Data is a set of values of qualitative or quantitative variables; Data is measured, collected and reported, and analysed, whereupon it can be visualized using graphs or images. Data mining is the process of analysing data from different perspectives and summarizing it into useful information. Most of the data is unstructured and hence it entails a process and method to extract useful information from the data and transform it into understandable and usable form. Many data mining tools are available for data mining tasks. These tools are using artificial intelligence, machine learning, statistics, information retrieval, database technology and other techniques to extract knowledge. The proposed review gives a short overview of various data mining tools like Matlab, Weka, Rapid Miner, R and Orange used for Heart disease prediction and also it presents the advantages, significant features and limitations of each tool.

**KEYWORDS**: Data Mining Tools, Machine learning, Matlab, Weka, Rapid Miner, R, Orange

## I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information[2]. It is an iterative and interactive process of discovering novel, valid, useful, comprehensive and understandable patterns and models in MASSIVE data sources. Data mining brings a set of tools and techniques that can be applied to processed medical data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions [3].

Medical data mining is Applying data mining techniques and methods in medical data. The challenge faced by healthcare industry with regard to the massive data-rich but information-poor collection is to extract valuable information to be available at a particular time, place in the form needed to support the decision-making process [18].

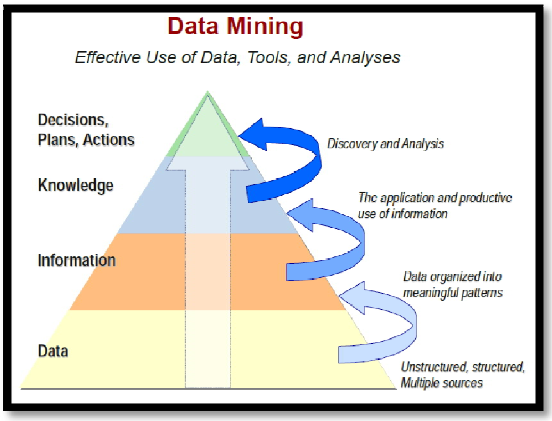

Data Mining
Effective Use of Data, Tools, and Analyses
Decisions, Plans, Actions
Knowledge
Information
Data
Discovery and Analysis
The application and productive use of information
Data organized into meaningful patterns
Unstructured, structured, Multiple sources

Fig 1: Transformation of data into decisions

Data mining consists of five major tasks: [11]

- Extract, transform, and load transaction data onto the data warehouse system.

DOI: 10.15680/IJIRCCE.2016. 0409090

- Store and analyze the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Present the data in a useful format, such as a graph or table.

## II. OVERVIEW OF HEART DISEASE PREDICTION SYSTEM

### A. *HEART DISEASE:*

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 or MI.** If the blood supply is cut off for more that a certain period of time, usually about 20 minutes, the muscle cells in the heart which are supplied by that artery may die [3].

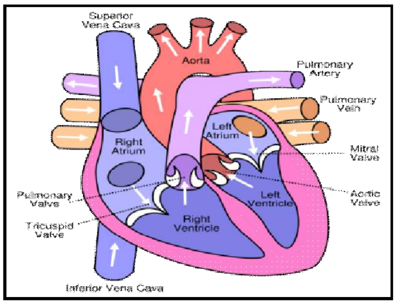


Fig. 2 Human Heart Structure

### B. *MEDICAL DATA MINING*

Applying data mining techniques in medical data to extract meaningful patterns and knowledge is called “Medical Data mining”. With the widespread use of databases and explosive growth in their sizes, the healthcare industry faced with the problem of information overloading. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making [28].

### C. *HEART DISEASE PREDICTION SYSTEM*

Heart Disease Prediction System aims to exploit the various data mining techniques on medical data set to assist in the prediction of the heart disease. Many data mining techniques like classification, clustering, regression and artificial neural networks can be used for heart disease prediction. Following are the 14 main attributes widely used for the prediction of heart disease.

Table 1: Shows important attributes used for heart disease prediction

| Attributes | Range |
|---|---|
| Age | |
| Gender | |
| Blood pressure | 100/70 to 150/90 |
| Cholesterol(LDL) | 100 to 189 |
| Heredity | 40% |
| Blood sugar | 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 | 70-100 |
| BMI | MEN<27 and WOMEN<25 |
| Smoking habit | YES/NO |
| Alcohol Intake | YES/NO |
| Mental stress | YES/NO |

## III. LITERATURE SURVEY

Ian H. Witten et.al [1] proposed a wide variety of machine learning methods in “Data Mining - Practical Machine Learning Tools and Techniques”.It describes these techniques and shows how they work. It presents how data are retrieved and visualized data mining techniques. Simple schemes are designed to explain clearly how the basic ideas work. It interprets machine learning as the acquisition of structural descriptions from simplest examples.

Dr. T. Karthikeyan et.al[2] proposed an analysis on various data mining classification algorithms used for the prediction of heart disease in “A Study on Data mining Classification Algorithms in Heart Disease Prediction”. It contains detailed study on data mining classification algorithms like, decision tree, Association rules, Naïve Bayes, Support Vector Machine, Neural Network also it sgives the advantages and disadvantages of each data mining algorithms.

Svetlana S. Aksenova [4] presented step by step explanation for WEKA data mining software in WEKA Explorer Tutorial. It contains descriptions to data mining tasks like preprocessing, classification, clustering, association, attribute selection, and visualization tools. At the end of each problem there is a representation of the results with explanations side by side. Each part is concluded with the exercise for individual practice.

Brian R. Hunt et.al [16] presented the real uses of matlab and a handy reference to the most useful features with pictorial representations in “A guide to MATLAB for beginners and experienced users”. It contains worked-out examples to interesting problems in mathematics and engineering. it contains explicit instructions for programming features and graphical capabilities.

Kalpana Rangra[23] employs the comprehensive and theoretical analysis of six open source data mining tools in Comparative Study of Data Mining Tools . It describes the technical specification, features, and specialization for each selected tool along with its applications. This paper presented the specific details and description of various open source data mining tools enlisting the area of specialization.

Table 2: Shows various data mining techniques used for heart disease prediction with accuracy

| Author | Purpose | Techniques Used | Tool | Accuracy |
|---|---|---|---|---|
| Chaitrali S.Dangare [20] | This paper presents prediction systems for Heart disease. | Decision Tree | Weka 3.6.6 | 90% |
| | | Naive Bayes | | 99.62% |
| Abhishek Taneja [19] | 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% |
| Roohallah Alizadehsan[24] | The focus is diagnosing Coronary artery disease. | Naïve Bays | Rapid Miner | 94.08% |
| Paulo Cortez[25] | It presents that R tool with algorithms, like neural Networks and support vector machines. | 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[27] | The focus is propose a Genetic Neural Approach for Heart Disease Prediction | Genetic-Neural Network | Matlab | 98% |

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 9, September 2016

## IV. DATA MINING TOOLS

Data Mining Tools are automatic collection and integration of data from a variety of internal data sources. Data mining software allows to analyze large volumes of raw data from healthcare systems, applications, database, websites and text based mediums. It extracts and manipulates information that is hidden in the raw data and is then reported in a well formed structure [17]

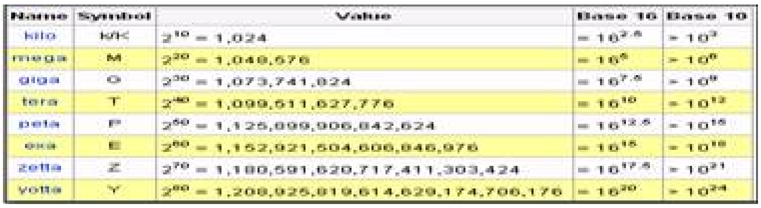
A. *Parameters for comparing different data mining tools:*

- Platforms supported
- Algorithms included
- Data input and model output options
- Usability ratings
- Visualization capabilities
- Modern automation methods

## V. KNOWLEDGE REPRESENTATION OUTLINE OF DATA MINING TOOLS

There are numerous different ways for representing the patterns that can be discovered by machine learning, and each one dictates the kind of technique that can be used to infer that output structure from data.[1]

*A)* ***TABLES* -** The rudimentary way of representing the output from machine learning is to make it as a *table*. It is an efficient format for comparative data analysis on categorical objects. The problem is, to decide which attributes to leave out without affecting the final decision.[1]



| Name | Symbol | Value | Base 16 | Base 10 |
|---|---|---|---|---|
| kilo | k/K | $2^{10}$ = 1,024 | = $16^{2.5}$ | ≈ $10^{3}$ |
| mega | M | $2^{20}$ = 1,048,576 | = $16^{5}$ | ≈ $10^{6}$ |
| giga | G | $2^{30}$ = 1,073,741,824 | = $16^{7.5}$ | ≈ $10^{9}$ |
| tera | T | $2^{40}$ = 1,099,511,627,776 | = $16^{10}$ | ≈ $10^{12}$ |
| peta | P | $2^{50}$ = 1,125,899,906,842,624 | = $16^{12.5}$ | ≈ $10^{15}$ |
| exa | E | $2^{60}$ = 1,152,921,504,606,846,976 | = $16^{15}$ | ≈ $10^{18}$ |
| zetta | Z | $2^{70}$ = 1,180,591,620,717,411,303,424 | = $16^{17.5}$ | ≈ $10^{21}$ |
| yotta | Y | $2^{80}$ = 1,208,925,819,614,629,174,706,176 | = $16^{20}$ | ≈ $10^{24}$ |

Fig 3. Data representation in table

*B)* ***LINEAR MODELS*** – It is a simple style of representation is a *linear model*, the output of which is just the sum of the attribute values, except that weights are applied to each attribute before adding them together before adding them together. These are easiest to visualize in two dimensions, where they are equivalent to drawing a straight line through a set of data points.[1]

*C)* ***TREES* -** A “divide-and-conquer” approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a *decision tree*. Nodes in a decision tree involve testing a particular attribute. Leaf nodes give a classification that applies to all instances that reach the leaf, or a set of classifications, or a probability distribution over all possible classifications.[1]

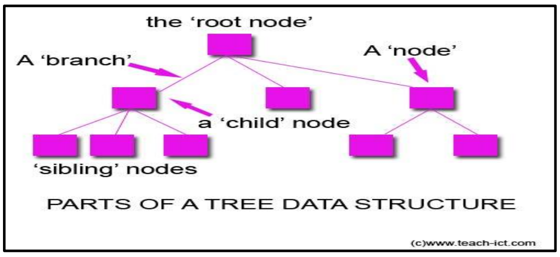


Fig 4. Tree Representation

*D)* ***RULES* -** Rules are a popular alternative to decision trees, and the *antecedent*, or precondition, of a rule is a series of tests just like the tests at nodes in decision trees, while the *consequent*, or conclusion, gives the class or classes that apply to instances covered by that rule, or perhaps gives a probability distribution over the classes. The preconditions are logically ANDed together, and all the tests must succeed if the rule is to fire.[1]

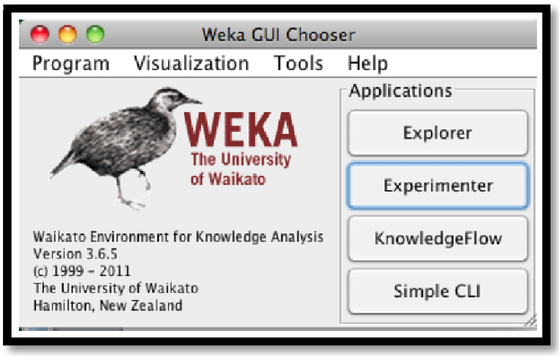
*E)* ***CLUSTERS*** **–** A cluster is learned, then the output takes the form of a diagram that shows how the instances fall into clusters. In the simplest case this involves associating a cluster number with each instance, which might be depicted by laying the instances out in two dimensions and partitioning the space to show each cluster.[1]

## VI. OUTLINE OF DATA MINING TOOLS

***A) WEKA*** **-** WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. It is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization [4].

The workflow of WEKA would be as follows:

**Data → Pre-processing → Data Mining →Knowledge**



Waikato Environment for Knowledge Analysis
Version 3.6.5
(c) 1999 - 2011
The University of Waikato
Hamilton, New Zealand

Fig 5. Weka Working Environment

The WEKA GUI contains **Explorer -** An environment for exploring data and supports data preprocessing, attribute selection, learning and visualization, **Experimenter –** It is used to performing experiments and conducting statistical tests between machine learning algorithms, **Knowledge Flow -** It has a drag-and-drop interface. It gives a visual design of the KDD process, **Simple CLI -** Provides a simple command-line interface for executing WEKA commands.[4]

***Features:***

- It is very easy to learn and use
- It contains an attractive GUI
- The supported data formats are ARFF, CSV, C4.5 and binary.
- WEKA has a very flexible combination of search and evaluation methods for the dataset's attributes. Search methods include Best-first, Ranker, Genetic-search, etc. Evaluation measures include InformationGain, GainRatio, ReliefF, etc. [4]

***Limitations:***

- Weka is much weaker in classical statistics.
- It does not have the facility to save parameters for scaling to apply to future datasets.
- It does not have automatic facility for Parameter optimization of machine learning/statistical methods [23].

B. ***RAPIDMINER*** **-** It is the world-leading open-source system for data mining. It is an environment for providing data mining and machine learning procedures including: data loading and transformation (ETL), data preprocessing and visualization, modelling, evaluation, and deployment [12]. Types of graphs and visualization techniques available in rapid miner are , Scatter matrices, Line, Bubble, Parallel, Deviation, Box, Contour, 3-D, Density, Histograms, Area, Bar charts, Stacked bars, Pie charts, Survey plots, Self-organizing maps, Andrews curves, Quartile, Surface plots [12].

***Features :***

- ➢ Flexible and affordable support options and fastest development environment
- ➢ Enterprise-ready performance, scalability for big data analytics and On-the-fly error detection
- ➢ Contains an organized logical GUI for business analytics success [12]

***Limitations:***

- It is useful to working with database files, such as in academic settings or in business settings. So this software requires the ability to manipulate SQL statements and files [23].

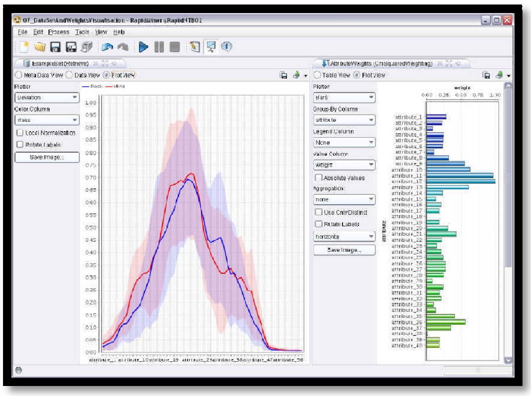
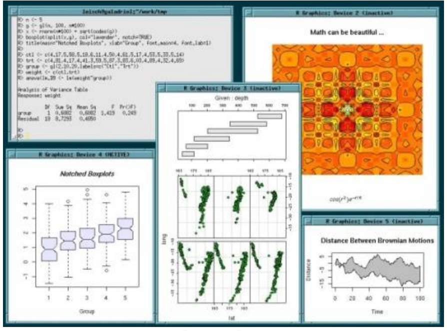


Fig 6.Visualization of Datasets in Rapid miner

B. ***R – TOOL*** - R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories by John Chambers and colleagues.[14] It is a software which provides an environment in which we can perform statistical analysis and produce graphics. It is actually a complete programming language[12]. R is widely used because the vast array of packages are available at the cran and bio conductor repositories. It is one of the standard tool in statistics. It offers wide variety of statistical and graphical techniques[14].



```
R> n <- 5
R> g <- gl(n, 100, n*100)
R> x <- rnorm(n*100) + sqrt(codes(g))
R> boxplot(split(x,g), col="lavender", notch=TRUE)
R> title(main="Notched Boxplots", xlab="Group", font.main=4, font.lab=1)
R>
R> ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
R> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
R> group <- gl(2,10,20,labels=c("Ctl","Trt"))
R> weight <- c(ctl,trt)
R> anova(lm.D9 <- lm(weight~group))

Analysis of Variance Table
Response: weight

          Df Sum Sq Mean Sq     F Pr(>F)
group      1 0.6882  0.6882 1.419  0.249
Residual  18 8.7293  0.4850

R>
R>
```

R Graphics: Device 4 (ACTIVE)
Notched Boxplots
Group
R Graphics: Device 3 (inactive)
Given : depth
R Graphics: Device 2 (inactive)
Math can be beautiful ...
R Graphics: Device 5 (inactive)
Distance Between Brownian Motions
Distance
Time

Fig 7. R-Tool Graphics

***Features:***

- ➢ R has over 4800 packages available from multiple repositories
- ➢ R is cross-platform. R runs on many operating systems and different hardware
- ➢ It contains extensive statistical library. It is a powerful elegant array language in the tradition of APL [14].

***Limitations:***

- Less specialized towards data mining.
- Programmer should familiar with array languages[23].

C. ***ORANGE*** - Orange is an Open source data visualization and analysis for novice and experts. It contains Components for machine learning. Add-ons for bioinformatics and text mining. Packed with features for data analytics. [5]. Orange can read files in native tab-delimited format, or can load data from any of the major standard spreadsheet file type, like CSV and Excel. Much of Orange is devoted to machine learning methods for classification, or supervised data mining. These methods rely on the data with class-labelled instances [8].

***Features:***

- simple data analysis with clever data visualization and effective Visual Programming
- Add-ons Extend Functionality and widget for assigning colors to entire schema[8].

***Limitations:***

- It contains only Limited list of machine learning algorithms.
- Machine learning is not handled uniformly between the different libraries [23].

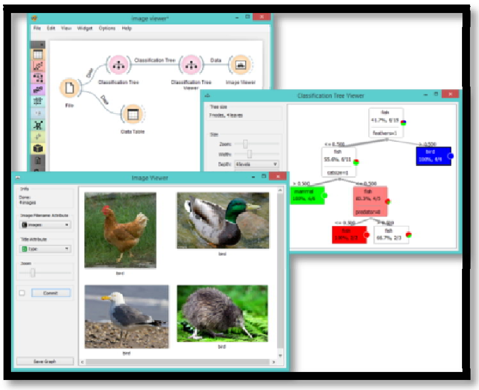


Fig 8. Data can contain references to images – Orange working environment

D. ***MATLAB*** - MATLAB is a high level language and interactive environment for numerical computation, visualization and programming. Using MATLAB we can analyze data, develop algorithms and create models and applications. The language, tools and built-in math functions enable us to explore multiple approaches and reach a solution faster than with spreadsheets or traditional programming languages [16].

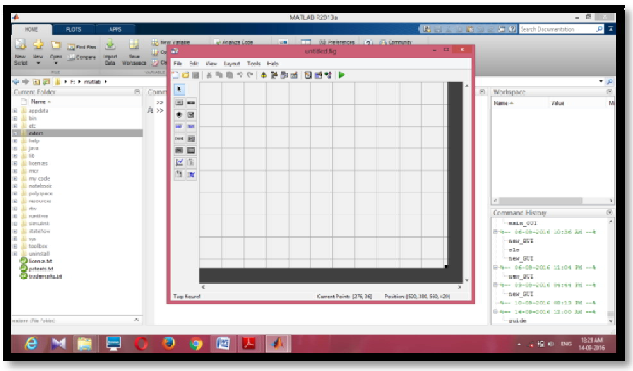


Fig 9. Working Environment of MATLAB

**Features:**

- It is useful to graph functions, solve equations, perform statistical tests and much more.
- It is very useful for Mathematical computations, Algorithm development, Modeling, simulation, and prototyping.
- It is an efficient platform for Data analysis, exploration, visualization, Scientific and engineering graphics
- It's functionality can be greatly expanded by the addition of toolboxes [16].

**Limitations:**

- It takes much CPU time for computation. It makes real time applications very complicated [23].

## VI.CONCLUSION

Data mining tools improve the detection of interesting and relevant patterns from database. Data mining gives a lot of techniques to extract the hidden potential information from the Healthcare industry. There are many efficient and prominent data mining tools available in the market. Some of the tools are available as a open source software. This proposed review given a brief overview of data mining visualization techniques and tools like Weka, Rapid Miner, R, Orange and Matlab. This review analyses advantages and limitation of each tools and provides concise information about the features of that tools and also helps to choose suitable tool for various data analysis tasks.

## REFERENCES

1. Ian H. Witten, Eibe Frank and Mark A. Hall, 'Data Mining - Practical Machine Learning Tools and Techniques", Third Edition, Morgan Kaufmann Publishers.
2. Dr.T.Karthikeyan, Dr.B.Ragavan and V.A.Kanimozhi 'A Study on Data mining Classification Algorithms in Heart Disease Prediction', International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol. 5, Issue 4, April 2016, ISSN: 2278 – 1323
3. V.A. Kanimozhi and 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, ISSN (Online) 2278-1021 ISSN (Print) 2319 5940.
4. Svetlana S. Aksenova , 'Machine Learning with WEKA - WEKA Explorer Tutorial for WEKA Version 3.4', 2004.
5. Mawuna Remarque Koutonin, 'The Best Data Mining Tools You Can Use for Free in Your Company', 2013.
6. Ralf Mikut, Markus Reischl , 'Advanced Review - Data mining tools', Volume 00, January / February 2011.
7. http://orange.biolab.si/
8. http://www.kdnuggets.com/2015/12/top-7-new-features-orange-3.html/2
9. Orange Data Mining, 'Orange Data Mining Library Documentation Release 3'.
10. http://blog.samibadawi.com/2010/06/orange-r-rapidminer-statistica-and-weka.html
11. http://www.rapid-i.com/downloads/brochures/RapidMiner_Fact_Sheet.pdf
12. Peter Dalgaard ,'Introductory Statistics with R', 2002
13. 'analyticstrainings.com' WordPress development.
14. https://www.r-project.org/about.html
15. www.knime.org
16. Brian R. Hunt, Ronald L. Lipsman and Jonathan M. Rosenberg, 'A guide to MATLAB for beginners and experienced users', Cmbridge University Press.
17. M. Usha Rani and R.J. Rama Sree, 'Superficial Overview of Data Mining Tools', Global Research Publications, 2008 edition.
18. Arun K Pujari , 'Data Mining Techniques', University press , Edition 2001.
19. Abhishek Taneja, 'Heart Disease Prediction System Using Data Mining Techniques', Oriental Journal Of Computer Science & Technology, Vol. 6, No. (4), ISSN: 0974-6471 December 2013.
20. Chaitrali S. Dangare and Sulabha, ' Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques', International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.
21. Vikas Chaurasia and Saurabh Pal, ' Early Prediction of Heart Diseases using Data Mining Techniques', Caribbean Journal of Science & Technology, ISSN 0799-3757.
22. Milan Kumari and Sunila Godara, ' Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction' , International Journal of Computer Science and Technology, IJCST Vol. 2, Issue 2, June 2011, ISSN : 2229-4333(Print)|ISSN:0976-8491(Online).
23. Kalpana Rangra and Dr. K. L. Bansal, 'Comparitive Study of Data mining Tools' International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, June 2014, ISSN: 2277 128X
24. Roohallah Alizadehsania, Jafar Habibia, Mohammad Javad Hosseinia, Hoda Mashayekhia, Reihane Boghratia, Asma Ghandehariouna, Behdad Bahadorianb and Zahra Alizadeh Sanib, 'A data mining approach for diagnosis of coronary artery disease', computer methods and programs in biomedicine' (2013), COMM-3519;
25. Paulo Cortez, 'Data Mining with Neural Networks and Support Vector Machines using the R/rminer Too'l, FCT grant PTDC/EIA/64541/2006.
26. Velu C. M. and Kashwan K. R., IEEE Member , 'Heart Disease Diagnosis Using Multiple Kohenen Self Organizing Maps'.
27. Nilakshi P. Waghuldel and 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.
28. David Crockett, Ryan Johnson and Brian Eliason, 'What is Data Mining in Healthcare?' Health Catalyst, 2004.

**BIOGRAPHY**

Prof. 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.