IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease

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

Research Scholar, Department of Computer Science, PSG College of Arts & Science, Coimbatore, India[1]

Associate Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore, India[2]

Abstract: Data mining can be viewed as a result of the natural evolution of information technology. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. 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. This survey analyzed different papers in which one or more machine learning algorithms of data mining used for the prediction of heart disease. Applying data mining techniques in medical field especially in heart disease treatment can give reliable performance.

Keywords: Data mining, Medical Data Mining, Data mining tools, Heart Disease prediction, Data mining techniques.

I. INTRODUCTION

The main objective of this paper is analyzing different papers in which one or more algorithms of data mining used for the prediction of heart disease. Data mining is an iterative and interactive process of discovering novel, valid, useful, comprehensive and understandable patterns and models in MASSIVE data sources (databases).
Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. By using data mining techniques it takes less time for the prediction of the disease with more accuracy. [25]

A. STEPS IN KNOWLEDGE DISCOVERY (OR) DATA MINING PROCESS:

1. Data integration - Where multiple data sources may be combined): First of all the data are collected and integrated from all the different sources.

2. Data selection - Where data relevant to the analysis task are retrieved from the database. We may not all the data we have collected in the first step. So in this step we select only those data which we think useful for data mining.

3. Data cleaning - To remove noise and inconsistent data): The data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data. So we need to apply different techniques to get rid of such anomalies.

4. Data transformation - Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance. The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.

5. Data mining - An essential process where intelligent methods are applied in order to extract data patterns. Data mining techniques are applied to discover the interesting patterns.

6. Pattern evaluation - To identify the truly interesting patterns representing knowledge based on some interestingness measure. This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.

7. Knowledge presentation - Where visualization and knowledge representation techniques are used to present the mined knowledge to the user. This step helps user to make use of the knowledge acquired to take better decisions.

Steps in Knowledge Discovery Process is shown in Fig.1.

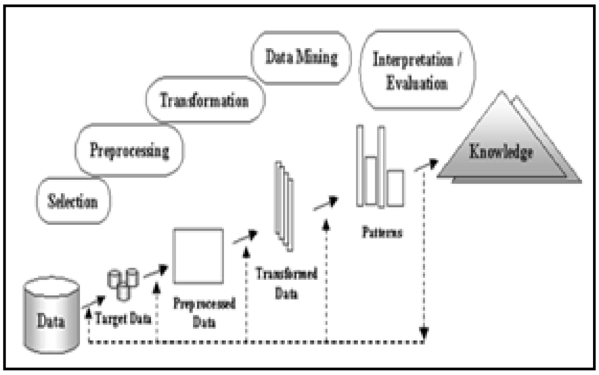

Fig. 1 The Data mining Process

B.FRAMEWORK FOR MEDICAL DATA MINING:

A general framework proposed by [23] for medical data mining is shown in Fig.2.

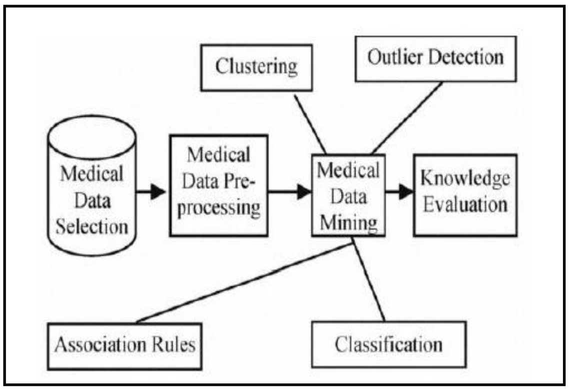

Fig. 2 Framework for medical data mining

II. AN OUTLINE OF 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.

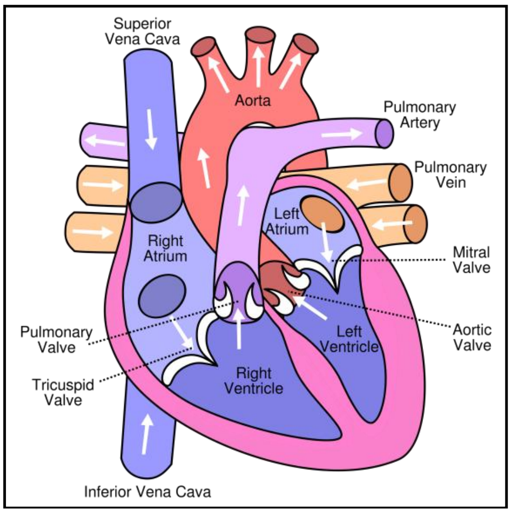

Fig. 3 Human Heart Structure

A) KINDS OF HEART DISEASE

- ❖ Coronary Artery Disease
- ❖ Heart Attack
- ❖ Angina
- ❖ Heart Failure
- ❖ Arrhythmias

Coronary Artery Disease (CAD)

When a substance called plaque builds up in the arteries that supply blood to the heart, CAD occurs. This Process can cause the arteries to narrow and harden over time, which is known as atherosclerosis.

Heart Attack

During a heart attack, flow of blood to the heart is severely reduced or stopped and heart muscle is damaged and soon begins to die. The more time passes without treatment, the greater the damage. Some heart attacks are sudden and intense, but most heart attacks start slowly, with mild pain or discomfort.

Angina

A symptom of Coronary Artery Disease, angina is chest pain or discomfort that occurs when the heart does not get enough blood. People describe angina as discomfort, pressure, or pain in the chest, back, neck, shoulders, arms (especially the left arm), or jaw.

Heart Failure

When your heart can't pump enough blood to meet your body's needs, you have a condition known as heart failure. It does not mean your heart has stopped. Key symptoms of heart failure include shortness of breath, a dry and hacking cough, weight gain, swelling, and fatigue.

Arrhythmias

An arrhythmia is an abnormally fast or slow heartbeat. An arrhythmia can also mean that your heart beats irregularly (skips a beat or has an extra beat). At some time or another, most people have felt their heart race or skip a beat.

B) HEART ATTACK RISK FACTORS :

Some 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 (140/90 or higher. Optimal is less than 120/80.)
- ➢ Atherosclerosis (hardening of the arteries)
- ➢ Not being physically active
- ➢ Being overweight or obese
- ➢ Blood pressure - 140/90 or higher

Cholesterol under 200, LDL under 100 and HDL over 40 for men - over 50 for women.

III. LITERATURE SURVEY

Heart disease is one of the important and crucial health diseases which occur among both men and women and is also the leading cause of the deaths in the world. This survey paper aims at analyzing various data mining techniques applied for heart disease prediction.

Table 1 Shows various data mining techniques used in heart disease prediction.

Author	Year	Techniques used	Total Medical Factors
Sellappan palaniappan Rafiah Awang	2008	Decision Tree, Naive Bayes	15

[15]		Neural Networks	
Jyoti Soni[6]	2011	K-Means Clustering, Decision Tree, Naive Bayes, Neural Networks, MAFIA algorithm	15
AH Chen, SY Huang, PS Hong, CH Cheng, EJ lin [8]	2011	Artificial Neural Networks	14
Jyoti Soni [9]	2011	Weighted Associative Classifiers	14
Milan Kumari [13]	2011	RIPPER classifier, Decision Tree, ANN, SVM	14
G.Subbalakshmi [17]	2011	Naive Bayes	15
Chaitrali S. Dangare [2]	2012	Decision Tree, Naive Bayes, Neural Networks	15
Nidhi Bhatla [5]	2012	Decision Tree, Naive Bayes Neural Networks	15
Shadab Adam Pattekari and Asma Parveen [7]	2012	Naive Bayes	15
M.Akhil jabbar [16]	2012	Associative Classifier, Genetic Algorithm	7
Ms. Ishtake S.H , & Prof. Sanap S.A. [1]	2013	Decision Tree, Naive Bayes, Neural Networks	15
Abhishek Taneja [4]	2013	Decision Tree, Naive Bayes Neural Networks	15
Vikas Chaurasia, [10]	2013	CART, ID3, Decision Table	11
Nilakshi.P. Waghulde [12]	2014	Genetic Algorithm, Neural Networks	12
Hlaudi DanielMasethe, Mosima Anna Masethe [14]	2014	J48,Naive Bayes, REPTREE, CART, BayesNet	11
Ms.Rupali R.Patil [19]	2014	Naive Bayes &Jelinek-mercer smoothing	13
Aditya Methaila [20]	2014	Neural Networks, Weighted Association with Apriori algorithm, Decision Tree, Naive Bayes, MAFIA algorithm	14
Andrea D'Souza[11]	2015	Neural Networks, K-Means Clustering, Frequent Itemset Generation using Apriori.	14

IV. MACHINE LEARNING ALGORITHMS USED IN DATA MINING

A.MACHINE LEARNING:
This is the algorithm part of the data mining process. It provides computers with the ability to learn without being explicitly programmed. This taxonomy or way of organizing machine learning algorithms is useful because it forces us to think about the the roles of the input data and the model preparation process and select one that is the most appropriate for our problem in order to get the best result.

Supervised Learning:
Input data is called training data and has a known label or result. A model is prepared through a training process where it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data.

Unsupervised Learning:
Input data is not labeled and does not have a known result. A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity.

DOI 10.17148/IJARCCE.2016.54137

Semi-Supervised Learning:
Input data is a mixture of labelled and unlabelled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions.

B) ALGORITHMS
Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases. [2]

1. Classification: Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large.
2. Clustering: Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes.
3. Predication: Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables.
4. Association rule: Association and correlation is usually to find frequent item set findings among large
5. Neural networks : Neural network is a set of connected input/output units and each connection has a weight present with it. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

Table 2. Shows various data mining techniques used in heart disease prediction with their accuracy

Author		Techniques Used	Accuracy
Ms. shtake S.H , &Prof.Sanap S.A. [1]		Decision Tree	94.93%
		Naive Bayes	95%
		Neural Networks	93.54%
Chaitrali S. Dangare [2]		Decision Tree	90%
		Naive Bayes	99.62%
		Neural Networks	100%
Jyoti Soni[6]	DM with ANN	Decision Tree	89%
		Naive Bayes	86.53%
		Neural Networks	85.53%
	DM with GN	Decision Tree	99.2%
		Naive Bayes	96.5%
		Classification via Clustering	88.3%
AH Chen, SY Huang, PS Hong, CH Cheng, EJ lin [8]		Neural Networks	80%
Vikas Chaurasia, [10]		CART	83.49%
		ID3	72.93%
		Decision Table	82.50%
Andrea D'Souza [11]		Neural Networks	79.38%
		K-Means Clustering	63.29%
Nilakshi P. Waghulde [12]		Genetic-Neural Network	98%
Milan Kumari [13]		RIPPER	81.08%
		Decision Tree	79.05%
		Neural Networks	80.06%
		SVM	84.12%
Abhishek Taneja [4]		Naive Bayes	86.53%
		Decision tree	89%
		Neural Networks	85.53%
Hlaudi Daniel Masethe, Mosima Anna Masethe [14]		J48	99.07%
		REPTREE	99.07%
		Naive Bayes	97.22%
		Bayes Net	98.14%
		CART	99.07%
Sellappan Palaniappan Rafiah Awang [16]		Naive Bayes	86.53%
		Neural Networks	85.53%
		Decision Tree	89%

V.OPEN SOURCE TOOLS AVAILABLE FOR DATA MINING

Rapid Miner - Rapid Miner is unquestionably the world-leading open-source system for data mining. It is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. It offers integrated environment useful in machine learning, text mining, data mining, business analytics and predictive analytics. The tool supports various steps useful in data mining including result optimization, visualization and validation.

Weka - Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. It shows you various relationships between the data sets, clusters, predictive modelling, visualization etc.

Orange - Orange is an Open source data visualization and analysis for novice and experts. Data mining through visual programming or Python scripting. Orange incorporates various components useful in data preprocessing, feature filtering and scoring, model evaluation, exploration and modeling techniques.

Tanagra - TANAGRA is a free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. The main purpose of Tanagra project is to give researchers and students easy-to-use data mining software to analyse either real or synthetic data.

Software – R - R-Software is another popular GNU open-source data mining tool. It already has a number of pre-defined modules and functions. Users need to write scripts for their operations. This data mining tool is more used by scientists, researchers and students for data mining and analysis requirements. It also provides tools for linear and non-linear modelling, statistical tests, classification, clustering etc.

MATLAB - A proprietary programming language developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, Fortran and Python. When doing data mining, a large part of the work is to manipulate data. Indeed, the part of coding the algorithm can be quite short since Matlab has a lot of toolboxes for data mining.

Table 3.Shows different data mining tools used on heart disease prediction with accuracy

Author	**Tool**	**Techniques Used**	**Accuracy**
Chaitrali S.Dangare [2]	Weka 3.6.6	Decision Tree	90%
		Naive Bayes	99.62%
		Neural Networks	100%
Abhishek Taneja [4]	Weka 3.6.4	Naive Bayes	86.53%
Nidhi Bhatla [5]	.NET platform	Naive Bayes	86.53%
		Decision Tree	89%
		Neural Networks	85.53%
Jyoti Soni[6]	Tanagra	Decision Tree	52.01%
		Naive Bayes	52.33%
		Neural Networks	45.67%
	Weka 3.6.0	Decision Tree	99.2%
		Naive Bayes	96.5%
		Classification via Clustering	88.3%
AH Chen, SY uang, PS Hong, CH heng, EJ lin [8]	Java	WAC	75.84%
Shadab Adam Pattekari and Asma Parveen [7]	C and C#	Neural Network	80%
Vikas Chaurasia[10]	Weka	CART	83.49%
		ID3	72.93%
		Decision Table	82.50%
Nilakshi P. Waghulde [12]	Matlab	Genetic-Neural Network	98%
Milan Kumari [13]	Weka	RIPPER	81.08%
		Decision Tree	79.05%
		Neural Networks	80.06%
		SVM	84.12%
Hlaudi Daniel Masethe, Mosima Anna Masethe[14]	Weka	J48	99.07%
		REPTREE	99.07%
		Naive Bayes	97.22%
		Bayes Net	98.14%
		CART	99.07%

VI. CONCLUSION

The objective of this survey work is to provide a study of different data mining techniques that can be employed in automated heart disease prediction systems. Various techniques and data mining classifiers are defined in this work for efficient and effective heart disease prediction. The analysis shows that different technologies are used in all the papers with taking different number of attributes. So, different technologies and algorithms used shown the different accuracy to each other.

REFERENCES

[1] Ms. Ishtake S.H , Prof. Sanap S.A. “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, April 2013.

[2] Chaitrali S. Dangare 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.

[3] Jyoti Rohilla, Preethi, “ Heart Disease Prediction Using Data Mining Techniques”, International Journal of Computer Science and Mobile Computing, A Monthly Journal of Computer Science and Information Technology,ISSN 2320–088X , IJCSMC, Vol. 4, Issue. 7, July 2015.

[4] Abhishek Taneja, “Heart Disease Prediction System Using Data Mining Techniques”, Oriental Journal Of Computer Science & Technology, ISSN: 0974-6471 December 2013, Vol. 6, No. (4).

[5] Nidhi Bhatla, Kiran Jyoti , “An Analysis of Heart Disease Prediction using Different Data Mining Techniques”, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181.

[6] Jyoti Soni, Ujma Ansari, Dipesh Sharma, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[7] Shadab Adam, Pattekari and Asma Parveen, “ Prediction System For Heart Disease Using Naive Bayes”, International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624. Vol 3, Issue 3, 2012.

DOI 10.17148/IJARCCE.2016.54137

[8] AH Chen, SY Huang, PS Hong, CH Cheng, EJ lin, "HDPS : Heart Disease Prediction System, Computing in cardiology", 2011 : 38:557-560.

[9] Jyoti Soni, Uzma Ansari, Dipesh Sharma, Suita soni, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers", Journal on Computer Science and Engineering (IJCSE), Volume: 3,No.6, June 2011.

[10] Vikas Chaurasia, Saurabh Pal, " Early Prediction of Heart Diseases using Data Mining Techniques", Caribbean Journal of Science & Technology, ISSN 0799-3757.

[11] Andrea D'Souza, "Heart Disease Prediction Using Data Mining Techniques", International Journal of Research in Engineering and Science (IJRES) ISSN (Online): 2320-9364, ISSN (Print): 2320-9356.

[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, ISSN : 2229-4333 (Print) | ISSN : 0976-8491 (Online).

[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] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques" IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

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

[17] Mrs.G.Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao , "Decision Support in Heart Disease Prediction System using Naive Bayes", Indian Journal of Computer Science and Engineering (IJCSE) .

[18] K.Srinivas, G.Raghavendra Rao and A.Govardhan, " Analysis of Attribute Association in Heart Disease Using Data Mining Techniques", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue4, July-August 2012.

[19] Ms.Rupali R.Patil, "Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing " , International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 5, May 2014.

[20] Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, "Early Heart Disease Prediction Using Data Mining Techniques", Computer Science & Information Technology (CS & IT), © CS & IT-CSCP 2014.

[21] K.Srinivas B.Kavihta Rani Dr. A.Govrdhan , "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010.

[22] Beant Kaur, Williamjeet Singh , "Review on Heart Disease Prediction System using Data Mining Techniques" , International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 10 3003 – 3008.

[23] Salha M. Alzahani, Afnan Althopity, Ashwag Alghamdi, Boushra Alshehri, and Suheer Aljuaid : " An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction", Lecture Notes on Information Theory Vol. 2, No. 4, December 2014.

[24] Bharati M. Ramageri "Data Mining Techniques And Applications", Indian Journal of Computer Science and Engineering Vol. 1 No. 4.

[25] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques" Second Edition, University of Illinois at Urbana-Champaign.

BIOGRAPHIES

Ms. Kanimozhi. V.A has completed her Msc 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.

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.

DOI 10.17148/IJARCCE.2016.54137