

A Survey Paper on Early Diagnosis Ofcardio-Vascular Disease Using Data Mining And Neural Network

Zarna Parekh¹ Avaniba Parmar²

^{1,2}Hasmukh Goswami College of Engineering, Ahmedabad, India.

Abstract— Cardiovascular diseases are major cause of humanity and death-rate in modern civilization. Medical assessment of these diseases is very important but the tortuous task should be performed efficiently and accurately. In this era of computing and intelligence it is an easy but complicated task to estimate the probability of disease on the basis of data and fact provided to the system. With the growing research it has become important to categories the research outcomes and provides readers with an overview of the existing heart disease prediction techniques in each category. The data mining algorithms are typically used to recognize the disease that occurs in original form the database. Neural Networks is a data mining analytical tools that is used to make prediction for medical data. In this paper, we present a survey report on the use of data mining algorithms & techniques and Neural Network Learning for the detection of Cardiovascular Disease.

Key words: Data Mining, Data Mining Techniques, Neural Network, Heart Disease Diagnosis

I. INTRODUCTION

Cardiovascular disease which is also called as heart disease comprises a class of disease which consists of heart the blood vessels or both. Heart diseases are the major cause of death globally: more people die annually from heart diseases than from any other disease [1]. An estimate is done that 17.3 million people died because of heart disease in 2012, which comprise of 30% of all global deaths. From these death an estimate was done that resulted with 7.3 million deaths were due to coronary heart disease and 6.2 million were due to stroke[1]. So to make a study and research on prevention of heart disease risk has become a important task nowadays for the researchers. Health care industry generates a large amount of complex data comprising of patients, electronic patient records, disease diagnosis etc. These large amount of data collected is a key source which is to be processed and analyzed for knowledge extraction which reduces the cost and helps for decision making. Extracting necessary information and providing scientific decision-making for the diagnosis and treatment of disease from the dataset becomes important. Data mining in medical field brings a set of tools and techniques which can be used to process the data to discover unseen patterns which provides health care professionals an additional useful source of information for making decisions. The database produced by the biological researchers can be massive [2]. However, using data mining techniques to develop a suitable treatment for prediction of heart disease has received a less attention so researchers have been identifying the effect of hybridizing more than one technique giving enhanced results indagnosis of heart disease. Artificial neural network has emerged as an efficient tool giving accurate and efficient results when hybridized with data mining[7].The rest of this paper is organized as follows. Section 2 Gives background knowledge about Data Mining and its techniques. Section 3

gives knowledge about ANN. Section 4 presents literature review on methods of data mining and ANN to predict cardiovascular disease. Conclusion of this survey is given in Section 5.

II. DATA MINING

Data Mining and Knowledge discovery is the process of getting important and valuable information that has been unknown in the raw data previously. It is a process of capturing knowledge with the help of computer. Data Mining is an exploration of large dataset to extract hidden and previously unknown patterns, relationships and knowledge which are not easy to detect with traditional statistics [14]. Researchers are trying to obtain satisfactory results in a reasonable time with help of searching techniques because many problems are difficult to be solved in a feasible time by analytically. Data Mining is divided into two tasks such as Predictive Tasks and Descriptive Tasks. The value of a specific attribute based on other attribute is been predicted by Predictive task. Classification, Regression and Deviation Deduction come under Predictive Tasks. Descriptive Tasks derive pattern that summarize the relationship between data and consists of Clustering, Association Rule Mining and Sequential Pattern Discovery. Data Mining involves few steps from raw data collection to some form of new knowledge. The iterative cycle consists of steps like Data cleaning, Data Integration, Data Selection, Data transformation, Data Mining, Pattern Evaluation, and Knowledge Representation. The Figure 1 shows that the Data Mining is the core for Knowledge Discovery processData Mining techniques like Naïve Bayes, Decision Tree, SVM(Support Vector Machine), K-NN, CMAR(Classification Based on Multiple Association Rules) etc are used for prediction of heart disease.[12]



Fig. 1: Data Mining is the core of Knowledge Discovery Process[12]

III. ARTIFICIAL NEURAL NETWORK

Artificial neural network is one of the most used method which helps to extract patterns in an intelligent and reliable way and is been greatly used as a way to find models which help to describe data relationship. A Neural Network (NN) is made up of many Processing Elements (PEs), loosely

called “neurons” and weighted interconnections among the PEs. Each PE performs a very simple computation, such as calculating a weighted sum of its input connections, and computes an output signal that is sent to other PEs[13]. The training (mining) phase of a NN consists of adjusting the weights (real valued numbers) of the interconnections, in order to produce the desired output. In learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of input samples [13]. The Artificial Neural Network (ANN) is a technique that is commonly applied to solve data mining applications. Neural Network is a set of processing units which when assembled in a closely interconnected network, provides rich structure exhibiting some features of the biological neural network. The structure of neural network gives an opportunity to the user to implement parallel concept at each layer level. They can be divided into two types based on the training method: Supervised training and unsupervised training. Networks that are supervised require the actual desired output for each input where as unsupervised networks does not require the desired output for each input. An iterative learning process is main feature of neural networks in which data cases are offered to the network one at a time, and the weights linked with the input values are attuned each time [13]. After all cases are presented, the process often starts over again. Network is been structured for a particular application when it is ready to be trained.

IV. LITERATURE SURVEY

A. Feature Selection Using Artificial Bee Colony For Cardiovascular Disease Classification [2]:

In this paper B. Subanya et al has planned feature selection approach in which ABC (Artificial Bee Colony) algorithm which is used to optimize the process of feature selection and gives best optimal feature subset that is used to raise the predictive accuracy of the classifier. The use of algorithm is done as a feature selector to generate the feature subset and a classifier is used to evaluate every feature subset that is made by the onlookers. Here the use of SVM is done with two key concepts to solve the problem: large margin separation and kernel function. The idea of large margin separation can be motivated by classification of points in 2 dimensions by drawing a straight line and call points on one side as positive and other as negative. Kernels are used for non-linear classification. Accuracy obtained is 86.76% with only 7 attributes.

B. Genetic Neural Network Based Data Mining In Prediction Of Heart Disease Using Risk Factors [3]:

In this paper Syed Umar Amin et al initially analyzed the dataset from various sources [15]-[16] and then composes a dataset of 12 important risk factors. Then by using the genetic algorithm which is specialized global searching algorithm [17] initial weights of neural network were optimized and use of back propagation algorithm was done to train the network using weight optimized by GA. Lastly a multilayered Feed Forward network is used. Here the 1st step is to initialize the weights of network using 'configure' function of MATLAB and then this weights are passed through GA to optimize according to fitness function and after weights are optimized BP algorithm is used train and

learn them. The predicted output will tell us the presence or absence of heart disease. 89% accuracy is obtained.

C. An Empirical Study On Prediction Of Heart Disease Using Classification Data Mining Technique [4]:

In this paper t John Peter et al has made use of pattern reorganization and data mining technique into risk prediction models in clinical domain of cardiovascular medicine and has proposed the system in which two methods are proposed: (1)ARFF creation and (2)attribute selection and classification.(1) In Attribute –Relation File Format file is an ASCII text file that describes a list of instances sharing a set of attributes. (2) A heart disease dataset consists large quantity of data and applying classifier to this data is time consuming and gives less accuracy so in this paper dimensionality of data is reduced using various attribute selection methods and then the data is classified using various classification methods.

D. Intelligent And Effective Heart Disease Prediction System Using Weighted Associative Classifiers [5]:

In this paper Jyoti Soni et al has proposed a new concept that uses Weighted Association Rule for classification. Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from the dataset. WAC has been proposed to get significant rule instead of flooded with insignificant relation. The major steps are:

- initially the heart disease dataset is pre-processed in order to make it suitable for mining process.
- each attribute is assigned a weight ranging from 0 to 1. Attributes having more impact will be assigned a highweight(0.9) and those having low impact will be assigned less weight(0.1).
- once the processing gets over WARM algorithm is applied to generate interesting pattern. It uses the concept of Weighted Support and Confidence Framework. Rules generated in this step are known as Classification Association Rules(CAR).
- these rules will be stored in rule base
- when a new patient’s record is available, CAR rule from the rule base is used to predict the class label.

E. Back Propagation Neural Network For Prediction Of Heart Disease [6]:

In this paper Nabeel Al-Milli et al has proposed a heart disease prediction system using BP algorithm which uses 13 attributes to predict heart disease. Algorithm uses two passes to pass through different layers of network: forward pass and backward pass. An activity pattern is given to input nodes of network in forward pass and its effect is propagated through the network layer by layer which produces a set of output which is actual response of the network. The actual response obtained is subtracted from a desired response to produce error signal which is propagated backward through the network. The synaptic weights are adjusted in such a way to make actual response of network come closer to desired response in true sense.

F. Feature Selection In Ischemic Heart Disease Identification Using Feed Forward Neural Network [7]:

In this paper K Rajeswari et al proposed Feed Forward Neural Network trained with BP algorithm to identify a person affected by IHD or not and further the classification

if IHD is done as 'high', 'medium' and 'low' risk levels. Two phases are included: 1) learning phase where network learns by modification of weights and 2) testing phase where unknown input is tested for proper learning of neural network. Here initially the data was collected and analyzed for heart risk score prediction based on extensive study and expert opinion. Based on data collected classification of immediate risk analysis was done as 'no risk', 'low risk', 'medium risk' and 'high risk'. Next step was feature selection in which features were selected from given dataset. Use of multilayer perception network was done. Everytime combinations of some features were taken and accuracy was checked. In the end 12 features were selected with training accuracy 89.4% and testing 82.2%.

G. Neural Network Approach In Diagnosis Of Patient: A Case Study [8]:

In this paper Farhad Soliemanian Gharehchopogh et al has made use of a decision support model. Here the author has proposed Neural Network for patients information and has made use of four attributes. Data is collected of 40 persons from health center in Tabriz region. Here multilayer perceptron neural network architecture with 6 input nodes, 4 hidden nodes and 2 output nodes is used. Back propagation algorithm is used to find beneficial and effective information from data. 85% accuracy was obtained.

H. Coronary Heart Disease Diagnosis By Artificial Neural Network Including Genetic Polymorphisms And Clinical Parameters [9]:

The proposed method includes 487 patient records who underwent uniform standard clinical examinations, coronary angiography and genetic analysis. The information obtained from tests and genotyping helped to create the dataset which was used for diagnosis of CHD using ANN. Multilevel FFN taught by statistical BP of error variable parameter set were selected in order to adjust ANN models by co-relating pair wise between dataset parameters and CHD diagnosis. Accuracy of model was increased by using GA with different optimizing parameters [18]. The novelty of this research was inclusion of genes as risk factors in order to estimate capabilities of CHD prediction. Near about 90% accuracy was obtained.

I. Early Diagnosis Of Heart Disease Using Classification And Regression Trees [10]:

In this paper the author has proposed a method to automatically classify PCG (Phonocardiograms) and used classification and Regression trees (CART) to identify pathological murmur. 3 methods are used: 1) Pre-Processing which consists of filtering of heart sounds is performed with goal of removing unwanted noise and segmentation – identifies the heart sound components and timing interval between them. 2) Feature Extraction – this phase is focused on extracting signal features that better highlight the properties of the PCG signal, with the goal of identifying those that are more suitable for classification purpose. 3) Classification and Regression Trees – a step by step process in which decision tree is constructed by either splitting each node on tree in two daughter nodes. Objective of partitioning is to find partitions of data such that terminal nodes are such homogeneous as possible.

J. Context Aware Cardiac Monitoring For Early Detection Of Heart Diseases [11]:

In this proposed method initially the use of ECG waveforms was done to determine many cardiac arrhythmias [19]. When the values of the interval do not fall within expected range then detection of different cardiac abnormalities was done. Then by using wearable sensors some vital signs were measured continuously which can say the kind of illness which may have direct impact on some heart disease. When abnormal ECG is detected and classified the CMS picks the rules associated with arrhythmia, which were defined by experts and stored in service providers cloud repository. If context matched with rule then necessary action is taken to help user to overcome the problem.

K. HDPS : Heart Disease Prediction System [12]:

In this paper the author has proposed a tool which is built in C language to implement heart disease classification and prediction via ANN. Learning Vector Quantization (LVQ), a prototype based supervised learning algorithm is used. The clinical data obtained is separated into two equal parts randomly. One is used for training and the other is used for testing. An initial weight is assigned randomly to each feature. The weights of all features are adjusted using calculated errors. The final weight of every feature is determined when the errors meet with the termination conditions. The testing data are then used to calculate the performance of this model. The process is repeated for 100 times. The output results include the average value of accuracy, specificity, and sensitivity. Lastly, the ROC curve is calculated in order to check the decency of the model.

V. CONCLUSION

Heart diseases are the major cause of death globally: more people die annually from heart diseases than from any other disease. So to make a study and research on prevention of heart disease risk has become an important task nowadays for the researchers. Data Mining Techniques along with various Artificial Neural Network methods has played a vital role in early prediction of heart diseases. This survey paper focuses on many data mining techniques and Neural Network models that are used in early diagnosis of heart diseases.

VI. ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude to my guide, Asst Prof. Avaniba Parmar for her valuable guidance and useful suggestions.

REFERENCES

- [1] "Global atlas on cardiovascular disease prevention and control", WHO, 2011
- [2] B. Subanya, Dr. R.R. Rajalaxmi "Feature Selection using Artificial Bee Colony for Cardiovascular Disease Classification" IEEE (ICECS-2014)
- [3] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg "Genetic Neural Network based Data Mining in Prediction of Heart disease using Risk Factors" 978-1-4673-5758-6/13 2013 IEEE
- [4] T. John Peter, K. Somasundaram "An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Technique "IEEE-

- International Conference On Advances In Engineering, Science And Management (ICAESM -2012)
- [5] Jyoti Soni, Uzma Ansari, Dipesh Sharma “Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers” vol 3 no 6, IJCSE 2011
- [6] Nabeel al-milli “Back Propagation Neural Network for Prediction of Heart Disease” JATIT & LLS Oct 2013,vol 56, no 1
- [7] K.Rajeswari , Dr.V.Vaithyanathan , Dr. T.R. Neelakantan “Feature Selection in Ischemic Heart Disease Identification Using Feed Forward Neural Network “ Procedia Engineering 41 (2012) Science Direct
- [8] Farhad Soleimanian Gharehchopoghi, Zeynab Abbasi Khalifelu ”Neural Network Approach in Diagnosis of Patient: A Case Study “978-1-61284-941-6/11 2011 IEEE
- [9] Oleg Yu. Atkov, Svetlana G. Gorokhova, Alexandr G. Sboev, Eduard V. Generozov, Elena V. Muraseyeva, Svetlana Y. Moroshkinad, Nadezhda N. Cherniyc “Coronary Heart Disease Diagnosis by Artificial Neural Network including Genetic Polymorphisms and Clinical Parameters” Journal of Cardiology (2012) Science Direct
- [10] Amir Mohammad Amiri and Giuliano Armano “Early diagnosis of Heart Disease using Classification and Regression Trees “ IEEE
- [11] Abdul Forkan, Ibrahim Khalil, Zahir Tari “Context Aware Cardiac Monitoring for Early Detection of Heart Diseases” IEEE Computing in Cardiology 2013
- [12] AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin “HDPS Heart Disease Prediction System” IEEE Computing in Cardiology 2011
- [13] George Cybenk,, (1996)“Neural Networks in Computational Science and Engineering”, IEEE Computational Science and Engineering, pp.36-42
- [14] Liangxiao. J, Harry.Z, Zhihua.C and Jiang.S “One Dependency Augmented Naïve Bayes”, ADMA, pp 186-194, 2005.
- [15] “Centre for Disease Control and Prevention, http://www.cdc.gov/heartdisease/risk_factors.htm.”
- [16] “American Heart Association, <http://www.heart.org/HEARTORG/Conditions>”
- [17] Y. J. Lei, and X. W. Zhang, Genetic Algorithm Toolbox of MatLab and its Application, Xian University of Electronic Science and Technology Press, 2005.
- [18] Goldberg DE. Genetic algorithms in search, optimization and machine learning. New York: Addison Wesley; 1989.
- [19] Owis MI, Abou-Zied AH, Youssef A, Kadah YM. “Study of features based on nonlinear dynamical modeling in ecg arrhythmia detection and classification”. Biomedical engineering IEEE Transactions on 2002;49(7):733-736