CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

WHO (2012) reported Cardiovascular Diseases are the number one cause of death globally. More people die annually from CVDs than from any other cause. As the requirement increases, at a faster rate, effective and accurate identification method is imperative to decrease mortality from heart diseases. Data mining techniques are proved to be a convenient tool to assist physicians in the identification of disease by obtaining knowledge and information regarding the disease from patient’s data.

Kangwanariyakul et al. (2010) and Srinivas et al.(2010) applied different techniques in the clinical field through data mining. All the above said researches are proved to be some extend of accuracy in the identification of heart disease. To predict the occurrence of heart disease in an individual, the researchers obtained the life style traits (risk factors) of that person and applied that risk factors to various soft computing classification methods like Neural Networks, Decision Trees, Probabilistic models, Classification and Regression Trees and Association Rules. Clustering algorithms like the k-means algorithm were used. This is applied on the clinical data of patients to identify instances which are more relevant to heart attack.
P.Kiran Sree et al. (2011) focused on the prediction of heart attack in the early stage by the modeling tool called as cellular automata (CA). It precisely explains the usage of cellular automata of one dimension with three neighborhoods CA. The CA database includes the large number of patient’s records. The above systems are been implemented by using the clinical data which are stored in a data warehouse. This work also explained with a rule based system and the usage of artificial neural networks to predict the disease. More formally all the above said three works predict the disease only based on behavioral patterns. These patterns will not judge the disease for all humans and may fail some times.

2.2 A SURVEY ON RISK FACTORS INVOLVED IN HEART ATTACK

It is well known that electrocardiogram plays a vital role in the prediction of the heart disease. According to Andra L. Blomkalns M.D et al. (2003) the ECG findings will provide certain details like ST Segment Elevation, ST – Segment Depression, Left Bundle Branch Block and Right Bundle Branch Block. ECG data analysis is done by the specialist or the physicians, who encounters the patient, and the decision is taken based on the past experience and old ECG data. The prognostic values may depend on some other factors too. The risk factors can be categorized as demographics, Major Risk factors, Contributing factors, History of diseases and presenting factors. The major risk factors for the heart disease is High Blood Pressure (Hypertension), High Blood Cholesterol, Diabetes, Obesity and Overweight, Smoking, Physical Inactivity, Gender, Heredity and Age. Some contributing risk factors are Sex Hormones, Stress, Birth Control Pills, and Alcohol.
A range of literatures proved that there is vast number of Clinical decision support systems (CDSS). Different CDSS make use of different methodologies for a better medical diagnosis. The above said point is noted in Clinical Decision Support Systems by Mark A.Musen et al. (2006). This paper describes a variety of tools used for making medical decisions and some of them proved a high value of accuracy. Clinical Decision Support Systems that banner abnormal values or that provide lists of interpretation for those abnormalities which is defined by Tatro et al.(2010). Such programs provide an initial output which is based on a set of specific patient data. All these decision making systems can be used as medical assistant only and the pharmaceutical decisions cannot be taken. A motivating Clinical Decision Support System is MYCIN. Shortliffe(1967) explained about the MYCIN program, which does not use straight forward algorithms or statistical approach. The researchers of MYCIN used the field of Artificial Intelligence.

The system proposed by Trigo et al.(2007) is a telecardiology framework for Clinical Decision Support system. The system uses the Heart rate variability in patients undergoing haemodialysis. The system produce the output based on the ECG data of the patient, these data are processed and the generated result is given to the specialist to take further action. Implemented system lacks in decision making due to huge amount of data storage.Many literatures define a high value of accuracy in Medical diagnosis using clinical decision support system based on the use of fuzzy logic and association rules. A Fuzzy Expert System for Heart Disease Diagnosis by Ali.Adeli et al.(2010) provides a prominent clinical decision support system using fuzzy logic. It is well understood that the repository used in this system is fair and can also accept the
input models and provides the prompt output. The system uses some fuzzy values and linguistic variables and the system intends to read twelve input variables and provide a fuzzy output ranges from 0 – 4. An added clinical decision support system is proposed with the use of Class Association rules by Wenmin Li et al. (2001). The proposed system is based on a class association rule, which groups a set of rules into a class label. Finally a pattern mining techniques are used to match the input data.

2.4 ECG ANALYSIS FOR HEART ATTACK PREDICTION

Electrocardiogram contains a wealth of diagnostic information routinely used to guide clinical decision making. ECG remains the reference standard for diagnosis despite the advance of many other diagnostic techniques in the identification of heart disease. Each individual heartbeat in the cardiac cycle of the recorded ECG waveform shows the time evolution of the heart’s electrical activity, which is made of distinct electrical depolarization–repolarization patterns of the heart. Any disorder of heart rate or rhythm, or change in the morphological pattern, is an indication of an arrhythmia, which could be detected by analysis of the recorded ECG waveform. Many algorithms for automatic detection and classification of ECG heartbeat patterns have been presented in the literature, including signal processing techniques such as frequency analysis, wavelet transform, filter banks, statistical and heuristic approaches, hidden Markov models, support vector machines, artificial neural networks (ANN), and mixture-of-experts method.

Turker Ince et al. (2009) proposed a Generic and Robust System for Automated Patient-Specific Classification of ECG Signals. This paper recommends the performance of ECG pattern classification strongly depends on the
characterization power of the features extracted from the ECG data and the
design of the classifier. The proposed feature extraction technique employs the
translation-invariant dyadic wavelet transform in order to effectively extract the
morphological information from ECG data. Furthermore, the dimension of the
input morphological feature vector is reduced by projecting it onto a lower
dimensional feature space using principal component analysis.

Andra L. Blomkalns et al. (2003) introduced criteria to predict the
cardiac events through the ECG data. This study used data from a prospectively
collected, retrospectively analyzed Internet-based data registry of undifferentiated
chest pain patients. Logistic regression modeling was performed to determine the
ECG findings that were predictive of positive cardiac markers and short-term
adverse cardiac events. Samantha Poli et al. (2003) reviewed the methods of
extracting the various features found on the surface of ECG. The results obtained
using more complex signal processing tools, such as frequency domain analysis
or non-linear techniques did not yield indexes as useful as those obtained by
time-domain analysis. Since these methods do not require the accurate
localization of P wave onset and offset, they may overcome the limitations
suffered by the time domain techniques.

2.5 APPLICATION OF DATAMINING TECHNIQUES ON ECG DATA

A team of researchers has used data-mining and machine-learning
techniques to find subtle changes in electrical activity in the heart that can be
used to predict potentially fatal heart attacks. Researchers from the University of
Michigan, MIT, Harvard Medical School and Brigham Women’s Hospital in
Boston sifted through 24-hour electrocardiograms from 4,557 heart-attack patients to find errant patterns that until now had been dismissed as noise or were undetectable.

Kavitha et al. (2014) proposed a model of predicting the heart disease using data mining techniques. Minghao Piao et al. (2012) illustrated the usage of data mining techniques in the prediction of heart disease using ECG data analysis. They proposed an incremental decision trees induction method which uses ensemble method for mining evolutorial diagnostic rules for cardiac arrhythmia classification. The work justified that the proposed method works better than the traditional algorithms. Gupta. K.O. et al. (2012) proposed a method of ECG signal analysis and classification using datamining and artificial neural networks. The work focused on four types of heart beats which is common: Normal, Premature Ventricular Contraction, Atrial Premature Contraction and Left branch Bundle Block Beat. These heart beats show different variations and are nonlinear in nature. Artificial neural networks are very useful in nonlinear problems and gives very capable results.

Maheshet al. (2009) illustrated the development of a rule-based expert system that emulates the ECG interpretation skills of an expert cardiologist. In this proposed system the user needs to input the wave features as seen in the ECG of a patient. These are used to identify the rule from an already established rule-base for diagnosis by the system. The post diagnostic results are stored in a database date-wise for future reference. The expert system also accepts comments from the user which can be incorporated in the system after validation by experts. This system is developed using Visual Basic and it provides a very simple and friendly user interface. The work incorporated with four major modules, the facts, a rule base, an inference module and a database.
Rajiv Kumar Nath et al. (2010) proposed a method of mining the ECG signal to obtain new diagnostic information. In the proposed methodology, based on an automatic identification of RR interval the study of signal pattern within this RR interval is done. RR interval of the ECG signal mostly used in the calculation of the heart rate of the patient and is one of the markers of one cycle of the ECG signal, which is generated from S-A (Sino-atrial) node. The R peak of the ECG signal generated from the contraction of the ventricles and applied in different disease identification. Dayong Gao et al. (2004) also proposed the method of ECG analysis using ANN classifier based on a Bayesian framework. The ECG signals and other patient data are gathered in real-time and sent to the Client Machine. The information is transmitted using the wireless network to the server. This stored data can then be processed to detect firstly the various complexes and then detect specified arrhythmia.

Gari D. Clifford (2009) proposed a method of Spectral and Cross-Spectral Analysis of the ECG. Ten seconds of 125-Hz typical ECG in sinus rhythm recorded with a lead II placement and associated linear and log-linear periodograms. Changing heart rates and varying the finite Fourier transform window size and overlap will change the relative magnitude of this cross-coherence. The Spectral Nature of Arrhythmias, which manifests due to abnormalities in the conduction pathways of the heart, can generally be grouped into either atrial or ventricular arrhythmias. Zeng et al. (1990) described a model for AV node conduction which was able to accurately model many observations of the statistical distribution of the beat-to-beat intervals during atrial arrhythmias. This model-based approach was further extended in to produce a method of classifying beats based upon their statistical distribution.
2.6 ECG DATA ANALYSIS USING PCA

PCA is a mathematical transformation used in the process of identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences. This technique uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The same technique is used for data dimensionality reduction. Okin et al. (2002) describes a way of using PCA in the dimensionality reduction of ECG data as the interested patterns are T-Wave values. The major demerit of the PCA technique is the complexity and the more amount of memory utilization. Charles J. Everett et al. (2005) proposed a method of Multiple Lipid Measures in prediction of coronary heart disease using PCA. It creates new variables which are called as principal components that generate the pattern from the total variation. Each principal component is a weighted linear combination of the original variables that is uncorrelated with any of the other principal components.

2.7 FORMAL CONCEPT ANALYSIS

Formal Concept Analysis (FCA) provides a way of deriving a formal ontology from a collection of objects and their properties. In the area of data analysis, it uses formal contexts and concept lattices. Nicholas Jay et al. (2006) proposed a method of Using FCA for mining the patient flow in the health care. The work focused on the ice berg lattices to find the patient flow. To remove the redundant data in the medical examination, Anamika Gupta et al. (2005) proposed a method using FCA. The context reduction techniques and classification rules are been used to find redundancies among the various medical examination tests. It acts as a field of applied mathematics based on the
mathematization of concept and conceptual hierarchy and thereby it activates mathematical thinking for conceptual data analysis and knowledge processing.

2.8 PROBABILISTIC MODELS USING BAYESIAN NETWORK

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. This model encodes dependencies among all variables. It readily handles situations where some data entries are missing. A Bayesian network can be used to learn causal relationships and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. As the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge and data.

According to Jayanta K. Ghosh et al. (2000) Bayesian networks details a capable technique for clinical decision support and provide powerful capabilities for representing uncertain knowledge, including a flexible representation of probability distributions that allows one to specify dependence and independence of variables in a natural way through the network topology. The estimation of cardiovascular patient risk with a Bayesian network is illustrated by Jan Bohacik et al. (2011) by accessing the risk of individual patients. A dataset of 839 instances is considered with 17 different attributes in this work. The Bayesian network is learnt on the basis of a decision tree. The decision tree is generated according to the C4.5 algorithm used on collected data for all 17 describing attributes $k A$, $k=1, \ldots, 17$, class attribute $C$ and 839 instances.
2.9 NON LINEAR CELLULAR AUTOMATA

The invention of homogeneous structure of Cellular Automata has been employed for modeling physical systems with diversity. To get better insight of a physical system, in due course, the CA structure is simplified with a restriction to local interactions among the cells. The simplified structure of a 1-dimensional CA, each cell having two states (0/1) with uniform 3-neighborhood (self, left neighbor and right neighbor) dependencies among the CA cells. It effectively introduces the modularity in a CA structure. Usha devi et al. (2010) introduced a method of fast and effective heart attack prediction system using Nonlinear cellular automata. Kiran sree et al. (2011) proposed another method of predicting the heart attack using Nonlinear cellular automata. The proposed methodology performs in a way that the heart disease CA set is pre-processed successfully by removing duplicate records and supplying missing values. The refined heart disease CA set resultant from pre-processing. The neural network is trained with the selected significant patterns. The designed prediction system employed MLPNN with Back-propagation as training algorithm.

2.10 ARTIFICIAL NEURAL NETWORKS

According to Hand et al. (2001) artificial neural networks are one of a class of highly parameterized statistical models that have attracted considerable attention in recent years. One of the advantages of using neural networks is that they are quite robust with respect to noisy data. Because the network contains many nodes, with weights assigned to each connection, the network can learn to work around these uninformative examples in the dataset.
A new system Intelligent Heart Disease Prediction System (IHDPS) was proposed by Sellappan Palaniappan et al. (2008). The system proved to be sturdy as it provides healthier results and output than any decision making system. The techniques used in IHDPS are the common techniques like Decision Trees, Naïve Bayes and Neural Network. The system also encompasses patterns, relationships and medical factors connected with heart disease. Anamika Gupta et al. (2005) projected an effort that the medical data can be analyzed and clustered using the data mining techniques. In this system the medical diagnostic data are retrieved and passed through a method of formal concept analysis. A model is proposed based on data mining techniques to classify the medical data. Basic classification rules and a redundancy algorithm is used which provides more strength to the system. The proposed system implemented utilize the ANN for the decision making process. The multi layered perceptron feed forward network is used in this process. The information is distributed among the nodes and the flow of network is uni-directional as it is a feed forward network. The processing among the nodes is done parallel by the network. All outputs of the first layer will be forwarded to the next layer in an unidirectional way in the feed forward network.

2.11 ST SEGMENT ELEVATION ANALYSIS

Sheng Nan Chang et al. (2008) introduced a method of measuring the ST segment elevation Myocardial Infarction through data mining techniques. The work is based on the narrative reports in electronics discharge notes in the clinical information systems of National Taiwan University Hospital from 2002 to 2004. The patients who presented to the emergency room with AMI were identified. Ten parameters related to QC for AMI were measured through data-mining of electronics discharge notes alone as well as through an extensive
search of the CIS and paper medical records. Performance measures that could be readily retrieved from the discharge notes or the hospital laboratory information system were selected as target parameters. The same cardiologist acted as a reviewer. He read the electronic records and transformed the narration into coded data in an EXCEL spread sheet.

The feasibility of ST segment elevation myocardial infarction and the occurrence of thrombus aspiration is proposed by Paul W Armstrong et al. (2013), which also illustrates the guidelines of the disease development. An inspiring literature denoted by Aruna et al. (2015) which details the detection of various arrhythmias using neural networks. The detection of different arrhythmias can be made possible by the proposed system but at times it leads to a long computation time. The indicated research works proves some extend of accuracy in the identification of heart disease.

In linear prediction a large number of cardiovascular disease classification and prediction systems are generated by artificial neural networks and are effective. Many instances with nonlinear inputs the artificial neural network fails. Aruna et al. (2015) illustrates the automatic identification of RR interval in the study of signal pattern within this RR interval and the proposed system adopts the classification techniques inculcated with the data mining techniques. James W. Hoekstra et al. (2009) proposed a method of Acute Detection of ST-Elevation Myocardial Infarction Missed on Standard 12-Lead ECG With a Novel 80-Lead Real-Time Digital Body Surface Map which clearly identified the extraction of ST segment from an ECG signal pattern.
A general and strong model for Automated Patient-Specific Classification of ECG Signals is proposed by Turker Ince et al. (2009). The proposed feature extraction system incorporates the translation-invariant dyadic wavelet transform in order to effectually excerpt the morphological information from ECG signal. Minghao piao et al. (2012) elaborated the use of data mining techniques in the prediction of heart disease using ECG data analysis. An incremental decision tree induction method is used which uses ensemble method for mining evolotional diagnostic rules for cardiac arrhythmia classification.

2.12 APPLICATION OF CART IN THE FIELD OF HEART DISEASE PREDICTION

The decision tree structure has the logistic regression functions at the leaves. The leaf node has two child nodes which is branched right and left depending on the threshold. In quest of arrhythmia detection and classification work, many computer techniques have been developed. Notably, Palreddy et al. (1997) employed a multiple-classifier architecture composed of Self Organizing Maps and Learning Vector Quantization to classify premature ventricular contraction beats and the non-PVC beats.

Linh et al.(2003) selected the Hermite Function Expansion as the feature extraction method to represent the QRS complex. The projected method uses a fuzzy neural network where Hermite coefficients served as the features to classify only seven different types of arrhythmias. Mahesh et al. (2009) proposed a method of logistic regression trees using a Logic Boost Method. The proposed classifier follows steps of regression. The linear regression function is fitted using the Logit boost method to build a logistic model.
The Logit boost method uses 5 examples for the cross validation to determine the best number of iterations to run, when fitting the logistic regression function at a node of the decision tree. The logistic model is built using all data. The split of the data at the root is constructed using the threshold. This splitting is continued till some stopping criterion is met. For the evaluation of the proposed classifier, a total of 1281 segments, extracted from the MIT BIH arrhythmia database records were used. Five time-domains, one frequency domain and two non-linear parameters were derived from these segments. These eight parameters along with the corresponding output class (type of arrhythmia) forms a feature vector. Thus 1281 feature vectors comprise the dataset. 75% of each type from this dataset was used as the train dataset and the remaining 25% as the test dataset.

Improving EASI ECG Method Using Various Machine Learning and Regression Techniques to Obtain New EASI ECG Model is introduced by Wojciech Oleksy et al. (2012). Four different methods were tested to find a best fitting model, namely ANN Linear Regression, Pace Regression and Bagging Predictors. Each model calculation was 10 fold cross validated. All results are based on data from PhysioNet database and also on data that were artificially. Even the system is designed for improving the efficiency it fails to focus on the response time, which is required to take into consideration as the input are nonlinear.

2.13 APPLICATION OF BACTERIAL FORAGING OPTIMIZATION IN HEART ATTACK PREDICTION

The theoretical foundations, analysis and applications of Bacterial Foraging Optimization Algorithm (BFOA) were described by
Swagatam Das et al. (2009). The work also focused on the underlying biological features behind the foraging strategy of E.Coli bacteria which is emulated in an extraordinary way and used for the different optimization problems. Ismail Ahmed Al Hadi et al. (2011) applied the Bacterial Foraging optimization algorithm in the Neural networking learning environment. The author utilized five types of datasets and demonstrated the convergence rate and the classification percentage. Minakhi Rout et al. (2014) applied the BFOA algorithm in the forecasting of currency exchange rates. By absorbing the simple statistical features of the exchange rate the internal coefficients are trained and the maximum value is obtained. The assembly line balancing using Bacterial Foraging Optimization Algorithm is detailed by Yakup Atasagun et al. (2014). The simple straight and U shaped assembly line balancing is unerringly demarcated. Passino (2002) describes the novel idea about the picturing the biological structure of the bacteria into the computational algorithm. The inspiration yields many researchers to attract to the algorithm and diverse to vast area of optimization algorithm. Nithin Kumar Jhankal et al. (2011) described the BFOA as a free derivative technique.

2.14 SUMMARY

A detailed study from more than 180 literatures are carried out and illustrated. The literatures and work study related to the problem domain is focused. The literature review provides a good background about the problem domain. The pros and cons of the different techniques are expressed in this chapter. Most of the work listed in this chapter lacks in accuracy of prediction due to the computation complexity. All traditional algorithms along with the newly inscribed algorithms are compared and illustrated in this chapter.