Diagnosis of CAD and risk factors using DM and Soft Computing Techniques – A Survey

R. Rajalakshmi and R. Latha

1Department of computer Science, St. Peter’s University, Avadi Chennai, Tamil Nadu, India
2Department of Computer Applications, St. Peter’s University, Avadi, Chennai, Tamil Nadu, India

Abstract: Coronary Artery Diseases (CAD) is a most listening disease and death rate also high in modern society. Patient’s information handling and computing is vital task in clinical section. In this regard essential tools are required to process the medical data. Data Mining (DM) is a collection of data that can be process, finding or exploring new recognizable pattern from large datasets using statistics and soft computing. Now a days, care of good health is most important task. So every hospitals needs huge amount of sample critical data about patients, disease diagnoses and related files etc. DM tools can answer these business queries that traditionally taken too much time consuming to resolve. The mining massive data is useful resources that supports to the doctors to predict, analyze, aware about disease, cost saving and decision making. By using data mining the time is consumed also disease prediction from clinical record is more effective and we discuss about the risk factor of CAD. The main aim of this survey is analysis of the uniqueness of medical data mining with soft computing techniques used to classification, diagnosis, prediction and prognosis. Finally, some of the research DM technique is also addressed to proceed on the same direction.

Key words: Data Mining · CAD · Prediction · Soft Computing · Risk Factors

INTRODUCTION

Today life threatening diseases are much greater than those of accidents and natural disasters. The primary reason of death in industrialized countries was due to cardiovascular disease. The World Health Organization estimates that 17.8 million deaths worldwide each year occur due to cardiovascular diseases. A major type of such diseases is coronary artery disease (CAD), which is reported to account for 7.5 million deaths over the world per annum. WHO estimated by 2030, almost 23.6 million people will die due to Heart disease as written in [1]. Prediction by using data mining techniques gives us accurate result of disease. CAD diseases not only have a major impact on individuals and their quality of life in general, but also on public health costs and the countries’ economies. Risk factors for these pathologies include diabetes, smoking, family history, obesity, high cholesterol etc, [2]. Health information decision is enabled by the physician and takes decision based on the patient’s answers to questions and lab test results [3]. The block occurs in coronary due to blood flow to the heart muscles was decreased. Prediction of heart disease is based on medical aware from patients. Due to various problems, prediction of CAD is more costly. But CAD can struck suddenly and have to take quick decisions. The prediction can be done with various computer aided diagnosis methods. All these diseases predictions are done by proper patient record maintenance. So these assessment we need tool to preserve the patient data. Data Mining is a tool to do all knowledge based data analysis. Data mining is the nontrivial extraction of implicit previously unknown and potentially useful information about data [4]. The data can be utilized by clinical admin to improve the services and predict the suitable treatment decisions for the patients.

Data mining is the computer based process of analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict future trends, allowing business to make proactive, knowledge-driven decisions. Some of the exist sample mining techniques that have been applied to medical data include Apriori and FP Growth [5], unsupervised neural networks [6], linear genetic programming [7], Association rule
mining [8], Bayesian Ying Yang [9], decision tree algorithms like ID3, C4.5, C5 and CART [10], Outlier prediction technique [11], Fuzzy cluster analysis [12], classification algorithm [13], Bayesian Network algorithm [14], Naïve Bayesian [15], combination of K-means, Self Organizing Map (SOM) and Naïve Bayes [16], Time series technique [17], combination of SVM, ANN and ID3 [18], clustering and classification [19], SVM [20], FCM [17], k-NN [21] and Bayesian Network [10]. This manuscript provides the summary of the techniques in terms of problem identified using data mining with soft computing in the following parts: (1) Pertaining to CAD risk factors and analysis using DM. (2) Review of literature (3) Proposed System (4) Concludes the manuscript.

Pertaining to Cad Risk Factors and Analysis Using DM:
Heart diseases can be caused by various factors out of which certain common factors can be mentioned as, age, sex, family history, high blood pressure levels, high cholesterol levels, obesity, stress, diabetes etc. N. Deepika et al. proposed Association Rule for classification of Heart-attack patients [42]. The extraction of significant patterns from the heart disease data warehouse was presented. These are the three main causes which leads to heart diseases (1) chest pain (2) stroke and (3) heart attack. [22]. To avoid and identification of these diseases different techniques of data mining is used through this easily find out heart related diseases. Heart disease is the leading cause of death all over the world in the past ten years. Numerous researchers are using statistical and data mining tools to help health care professionals in the diagnosis of heart disease. [23]. Following are some of the risk factors for heart disease:

**Smoking**: Smokers risk a heart attack twice as much as nonsmokers.

**Cholesterol**: A diet low in cholesterol and saturated Tran’s fat will help lower cholesterol levels and reduce the risk of heart disease.

**Blood Pressure**: High Blood Pressure leads to heart Attack.

**Diabetes**: Diabetes if not controlled can lead to significant heart damage including heart attack and death.

**Sedentary Life Style**: Simple relaxation time activities like gardening and walking can lower our risk of heart disease.

**Eating Habits**: Healthy diet, intake of low salt in diet, saturated fat in body, Tran’s fat, cholesterol and refined sugars will lower our chances of getting heart disease. Stress: Poorly controlled stress a danger can lead to heart attacks and strokes.

The heart disease data warehouse contains the screening clinical data of heart patients. Initially, the data warehouse preprocessed to make the mining process more efficient. The first stage of Association Rule used preprocessing in order to handle missing values. Later applying equal interval binning with approximate values based on medical expert advice on Pima Indian heart attack data. The significant items were calculated for all frequent patterns with the aid of the proposed approach. The frequent patterns with confidence greater than a predefined threshold were chosen and it was used in the design and development of the heart attack prediction system.

The, Pima Indian Heart attack dataset used was obtained from the UCI machine learning repository. Characteristics of the patients like number of times of chest pain and age in years were recorded. M A. Jabbar et al. proposed Association Rule mining based on the sequence number and clustering for heart attack prediction [43]. The entire database is divided into partitions of equal size. The dataset with 14 attributes was used in that work and also each cluster is considered one at a time for calculating frequent item sets. This approach reduces main memory requirement. To predict the heart attack in an efficient way the patterns are extracted from the database with significant weight calculation. The frequent patterns having a value greater than a predefined threshold were chosen for the valuable prediction of heart attack.

In heart disease prediction system, the input attributes play a major role for efficient prediction. Certain factors that are considered are as follows:

- **Age**: age in years >28
- **Sex**: sex (m = male; f = female)
- **Chest pain type**: cp

**Value 1**: typical angina

**Value 2**: non-anginal pain

**Value 3**: asymptomatic

**Cholesterol (chol)**: serum cholesterol in mg/dl
**Lipid Levels:** Low high-density-lipoprotein (HDL) levels (<0.91 mmol/L [<35 mg/dL]) and high low-density-lipoprotein (LDL) levels are independently associated with CAD (HDL especially in women).

**Diabetes Mellitus:** Increased risk is related to hyperglycemia and hyperinsulinemia, both of which are atherogenic.

**Hypertension:** Systolic and diastolic blood pressures are independent risk factors, but the systolic is the preferred marker.

**Smoking:** Promotes atherogenesis, ischemia and thrombogenesis.

**Family History:** Especially premature disease (parent with MI before age 60), but the independent effect is difficult to quantify.

**Left Ventricular Hypertrophy:** A powerful independent risk factor for CAD (more so than diabetes mellitus or smoking).

**Homocysteine:** Higher levels are associated with a 20% to 40% increased risk of cardiovascular events.

**C-reactive Protein:** The prevalence of CAD is increased by 50% for each doubling of the C-reactive protein level. The pathophysiologic significance of this relationship remains unclear.

**Other Risk Factors:** These include obesity, high levels of uric acid, triglycerides, lipoprotein Lp, tissue plasminogen activator antigen, fibrinogen and leukocytes.

By applying these factors as input to the algorithms the newer set of attributes and a more efficient system is developed. In terms of, accuracy and performance are obtained in a fraction of some milliseconds. The system that is proposed helps the medical practitioners to improve their practice as well as provide quality service.

**Review of Literature:** Data Mining is main concerned with the analysis of data and Data Mining tools and techniques are used for finding patterns from the data set. The main objective of Data Mining is to find patterns automatically with minimal user input and efforts. Data Mining is a powerful tool capable of handling decision making and for forecasting future trends of market.

Data Mining tools and techniques can be successfully applied in various fields in various forms. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis. By using Mining tools and techniques, various fields of business get benefit by easily evaluate various trends and pattern of market and to produce quick and effective market trend analysis. Data mining is very useful tool for the diagnosis of diseases. Generally, Prediction of heart disease physician uses artificial intelligence techniques [25]. To predict, prognosis and prescription in medical domain puts more stress on real cases than other domains uses Case-based reasoning (CBR) [26]. Medical diagnostic problem are identified by Machine learning algorithm [27]. Coronary Artery Disease was diagnosed using two techniques called Binary Particle Swarm Optimization (BPSO) and Genetic Algorithm (GA) [25]. For the diagnosis of heart disease various classification and prediction processes was used.

Wu, et al. has proposed that integration of clinical decision support with computer based patient records could reduce the medical errors, which could automatically enhanced the patient safety and will decrease unwanted practice variation and improve patient outcome [24].

Jesmin Nahar and Tasadduq Imam et al. [28] have proposed a computer intelligent based diagnosis of heart diseases. Apriori, Predictive Apriori and Tertius were the three different rule mining algorithms used to present rule extraction experiment on heart disease data and showed as efficiency algorithm for diagnosis task. Cleveland dataset, a publicly available dataset and widely popular with data mining researchers, have been used for diagnosis because of the privacy problem related to medical data set.

Resul Das and Ibrahim Turkoglu, et al. [29] have proposed several tools and various methodologies to develop effective medical decision support system. Diagnosing of the heart disease was one of the important issue and many researchers investigated to develop intelligent medical decision support systems to improve the ability of the physicians. A method was introduced which uses Statically Analysis System (SAS) base software 9.1.3 for diagnosing of the heart disease. In this method neural networks ensemble model was used which enabled an increase in generalization performance by combining several individual neural networks train on the same task. SAS base software 9.1.3 supported all tasks in a within a single, integrated solution while providing the
They have used two hundred and two coronary heart immature patients with coronary heart disease.

Socioeconomic status and the course of quality of life problems in CHD diagnosis risk assessment.

Information on unclear character and uncertainty provides belief interval as final output. It provides precise reasonable function from previous stage result which problem opinions were calculated the belief and seeming and completed by fuzzy interference rules. The systems represented with fuzzy sets and fuzzy rules in first phase interfered in two phase. The problems data were unclear character and uncertainty modeling problems were designed using evidence and fuzzy set theories. Here assessment problem in coronary heart disease which was have proposed a hybrid engine to determine risk specificity values for heart disease diagnosis.

Classification accuracy, 80.95% sensitivity and 95.91% diagnosis the experimental result obtained 89.01% flexibility for efficient collaborations. For heart disease diagnosis the experimental result obtained 89.01% classification accuracy, 80.95% sensitivity and 95.91% specificity values for heart disease diagnosis.

Laercio Brito Gonçalves and Marley Maria Bernardes Rebuzzi Vellasco et al. [30] have determined that the Inverted Hierarchical Neuro-Fuzzy Binary Space Partitioning (HNFB-1) based on the Hierarchical Neuro-Fuzzy Binary Space Partitioning Model (HNFB) which gave an idea that recursive partitioning of the input space. The classification task of HNFB-1 has been evaluated with different benchmark databases such as heart disease data sets. They introduced an Inverted Hierarchical Neuro-Fuzzy BSP System. It was a neuro-fuzzy model which has been specifically created for record classification and rule extraction in databases. It allowed the extraction of knowledge in the form of interpretable fuzzy rules. The HNFB-1 model had showed better classification performance when compared with several other pattern classification models and algorithms and the processing time converged by HNFB-1 was very less.

Kemal Polat and SeralSahan et al. [31] have applied k-nearest neighbour (k-nn) weighting preprocessing and fuzzy resource allocation mechanism with AIRS on the task of diagnosis of heart disease. In this system, a new weighting scheme based on k-nearest neighbour (k-nn) method was utilized as a preprocessing step before the main classifier. The results strongly suggested that k-nn weighted preprocessing and fuzzy resource allocation mechanism with AIRS can aid in the diagnosis of cardiac arrhythmias. 87% of classification accuracy was obtained by their system.

Vahid Khatibi and Gholam Ali Montazer et al. [32] have proposed a hybrid engine to determine risk assessment problem in coronary heart disease which was designed using evidence and fuzzy set theories. Here unclear character and uncertainty modeling problems were interfered in two phase. The problems data were represented with fuzzy sets and fuzzy rules in first phase and completed by fuzzy interference rules. The systems problem opinions were calculated the belief and seeming reasonable function from previous stage result which provides belief interval as final output. It provides precise information on unclear character and uncertainty problems in CHD diagnosis risk assessment.

Giorgio Barbareschi et al. [33] have proposed socioeconomic status and the course of quality of life immature patients with coronary heart disease. They have used two hundred and two coronary heart disease patients based on community based survey. Data on patients’ quality of life were gathered prior to the diagnosis. Their outcome showed that high socioeconomic status patients notified better end result at the premorbid assessment with less depressive feelings and improved physical functioning. Eventually, they concluded that coronary heart disease modulates premorbid divergence in depressive feelings.

Ghannad-Rezaie et al. [34] presented an approach to integrate PSO rule mining methods and classifier on patient dataset. They used Particle Swarm Optimization technique as well. The results revealed that, their approach is capable of performing surgery candidate selection process effectively in epilepsy.

Mahmud Khan et al. [35] focused on decision tree data mining algorithm for medical image analysis. Especially they studied on lung cancer diagnosis through classification of x-ray images.

Cheng et al. [36] applied classification algorithm to diagnose cardio vascular diseases. For classification effectiveness they focused on two feature extraction techniques namely automatic feature selection and expert judgment.

Xue et al. [37] proposed and applied Bayesian Network algorithm for diagnosis of an ailment known as Coronary Heart Disease (CHD).

Xing et al. [38] developed data mining techniques for predicting the probability of survival of CHD patients. To achieve this they combined three prediction models such as SVM (Support Vector Machine), Artificial Neural Networks (ANN) and Decision trees using C4.5 or ID3, CART and C5.

HaiWang and Shouhong Wang [39] studied on the role of medical experts in medical data mining. Medical experts can give expert advice that can be used as input in medical data mining.

Drugs and health effects are mined by Froelich and Wakulicz-Deja [40] using adaptive FCM (Fuzzy Cognitive Maps). Their work has led to improved decision support and planning in Healthcare domain.

In the research [41], Coronary artery disease (CAD) shakescresores of people all over the world including a major portion in India every year. Although much progress has been done in medical science, but the early detection of this disease is still a challenge for prevention.

The objective of this paper is to describe developing of a screening expert system that will help to detect CAD at an early stage. Rules were formulated from the doctors and fuzzy expert system approach was taken to cope with uncertainty present in medical domain.
This paper describes the work on risk factors responsible for CAD, knowledge acquisition and knowledge representation techniques, method of rule organization, Neural network, fuzzification of clinical parameters and defuzzification of fuzzy output to crisp value. The proposed CAD methodology is developed to assist the medical practitioners in predicting the patient’s risk status from rules provided by medical experts. The present paper focuses on rule organization using the concept of modules, DM meta-rule base, rule address storage in tree representation and rule consistency checking for efficient search of large number of rules in rule base.

Proposed System: The work presented in this paper is the application of data mining for discovering hidden knowledge from a medical dataset. The Medical data is temporal in nature and therefore traditional data mining techniques are not appropriate. This dataset contains medical records of coronary artery disease patients. The structure of these medical records is chain of observations taken at different times. In each observation, a set of clinical parameter are saved. The aim of this paper is mining the rules from set of medical data that can be used in early prediction and of risk in the patients. This study conducts data mining algorithms on the Sample 505 dataset. In this paper, the effectiveness of various rule-mining algorithms on the dataset is investigated. The data mining methodologies are used to harvest the embedded information which is used as a cause of knowledge for decision building. The electronic data are preprocessed by the data or pattern analysis to construct the predictive modules. These modules are rooted with the algorithms such as K-means, SVM, CART, Naïve Bayes, Fuzzy, KNN and etc., to envisage the indispensable knowledge or information for decision making.

In the proposed method the survey describes, the dataset is properly designed according to the compatibility with the data mining algorithms, after observing the dataset, weight must be assigned to the feature and then the features are selected for the task of classification according to their weights and finally the association rules will be generated. The generated association rule data mining techniques that are used at Naïve Bayes, decision trees and neural networks. The Naïve Bayes classifiers have the minimum error rate in comparison to all other classifiers. Bayesian classifiers have exhibited high accuracy and speed when applied to large databases. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Decision trees can handle both categorical and numerical data. The decision tree inducers are the algorithms that involuntarily construct a decision tree from a specified dataset. The decision tree algorithm has experienced a lot in the world of data mining. These inducers algorithms such as CART, C4.5 and C5 are largely used in the predictions. A neural network usually involves a large number of processors operating in parallel, each with its own small sphere of knowledge and access to data in its local memory. Typically, a neural network is initially "trained" or fed large amounts of data and rules about data relationships. There are large database of information that has been stored in various electronic forms which may consist of curtained, noisy and inconsistent data.

Optimization: The analytical methods that are used to stumble on the optimum solution or unimpeded maxima or minima of constant and differentiable function are said to be the Classical optimization techniques. The formula behind these techniques is executed iteratively by comparing diverse solutions in order to acquire the expected optimal result. There are two discrete types of optimization algorithms generally used.

Deterministic Algorithms: Specific rules for moving one solution to other.

Stochastic Algorithms: Probabilistic translation rules for gaining popularity due to certain properties. One of such algorithm is the Swarm Intelligent (SI).

Swarm Intelligent Optimization: The synthetic intelligence which is based on the collective performance of decentralized and self-organized systems is known as Swarm Intelligent (SI). The SI is a loosely structured collection of interacting agents which can be distinguished, communicated and/or interrelated with each other.

Particle Swarm Optimization (PSO): A Swarm Intelligent technique that searches for a best possible solution in the computable search space based on a population. This stochastic optimal search has been inspired by the Swarms of Bees, Flocks of Birds and Schools of Fish. The individuals of the search attempts to improve themselves by observing and imitating their neighbors. PSO are exceedingly used to find approximate solutions to extremely complex or unfeasible numeric maximization and minimization problems.
The fuzzy logic is said to be the superset of Boolean logic that has been unmitigated to grip the concept of the partial certainty values between “completely true” and “completely false”. The fuzzy system logic recognizes further than simple true and false values. The expertise considers fuzzy logic as “a constitution of knowledge depiction appropriate for notions that cannot be defined accurately, but which depend upon their context”. The ultimate scenario of fuzzy system is the prospect for modeling of circumstances which are inherent and simultaneous numerical and linguistic data. Fuzzy systems are extensively used for modeling, simulating and replicating many genuine tribulations.

In pattern recognition, the \( k \)-Nearest Neighbors algorithm (or \( k \)-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the \( k \) closest training examples in the feature spaces.

\( K \)-NN has a number of applications in different areas such as health datasets, image field, cluster analysis, pattern recognition, online marketing etc. There are various advantages of KNN classifiers. These are: ease, efficacy, intuitiveness and competitive classification performance in many domains. If the training data is large then it is effective and it is robust to noisy training data. A main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample. If there is a big sample then its response time on a sequential computer will also large. Studies of various DM techniques used for classifying the disease symptoms, diagnosing, prediction and/or prognosing of coronary artery diseases,

**CONCLUSION**

The medical data is time based historical in nature but from our survey the traditional data mining techniques are not much suitable to classification and prediction of disease. The dataset cycled for the proposed work to handle the number of attributes for diagnose disease from the given CAD medical records. Mining of this record used various techniques discussed. All the techniques had unique specialty to classify, predict and prognoses the disease. This paper has provided the summary of data mining techniques used for medical data mining. It also throws light into the importance of locally frequent patterns and the mining techniques used for diagnosing, prediction purpose. After careful evaluation on correctness and analysis of the results by these experts positive feedbacks were received. The survey concludes that the data mining prediction techniques using soft computing methods gives more effective outputs with more attributes as inputs than the same direction of Matlab with soft computing.

**REFERENCES**


14. Weimin Xue, Yanan Sun and Yuchang Lu (n.d), Research and Application of Data Mining in Traditional Chinese Medical Clinic Diagnosis. IEEE., pp: 1-4.


33. Giorgio Barbareschi and Robbert Sanderman et al., 2009. “Socioeconomic Status and the Course of Quality of Life in Older Patients withCoronary Heart Disease”, International Journal of behavioral Medicine, 16: 197-204.


37. Weimin Xue, Yanan Sun and Yuchang Lu (n.d.), Research and Application of Data Mining in Traditional Chinese Medical Clinic Diagnosis. IEEE., pp: 1-4.
39. Hai Wang, Shouhong Wang 1. (n.d), Medical Knowledge Acquisition through Data Mining. IEEE., 0(0): 1-4.