Abstract—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. Data mining tools can answer business questions that traditionally taken much time consuming to resolve. The huge amounts of data generated for prediction of heart disease are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. By using data mining techniques it takes less time for the prediction of the disease with more accuracy. In this paper we survey different papers in which one or more algorithms of data mining used for the prediction of heart disease. Result from using neural networks is nearly 100% in one paper. So that the prediction by using data mining algorithm given efficient results. Applying data mining techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease. The healthcare industry is generally “information rich”, which is not feasible to handle manually. These large amounts of data are very important in the field of Data Mining to extract useful information and generate relationships amongst the attributes. The doctors and experts available are not in proportion with the population. Also, symptoms often be neglected. Heart disease diagnosis is a complex task which requires much experience and knowledge. Heart disease is a single largest cause of death in developed countries and one of the main contributors to disease burden in developing countries. In the health care industry the data mining is mainly used for predicting the diseases from the datasets. The Data Mining techniques, namely Decision Trees, Naïve Bayes, Neural Networks, Associative classification, Genetic Algorithm are analyzed on Heart disease database.

Keywords: Data Mining, Decision Support, Heart Disease, Naïve Bayes

I. INTRODUCTION

Data Mining is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining refers to extracting or “mining” knowledge from large amounts of data. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data. The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data. Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but are largely limited. They can answer simple queries like “What is the average age of patients who have heart disease?”, “How many surgeries had resulted in hospital stays longer than 10 days?”, “Identify the female patients who are single, above 30 years old, and who have been treated for cancer.” However they cannot answer complex queries like “Given patient records, predict the probability of patients getting a heart disease.” Clinical decisions are often made based on doctors’ intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

The main objective of our paper is to learn the different techniques of data mining used in prediction of heart disease by using different data mining tools. Life is dependent on efficient working of heart because heart is essential part of our body. If operation of heart is not proper, it will affect the other body parts of human such as brain, kidney etc. Heart disease is a disease that affects on the operation of heart. There are number of factors which increases risk of Heart disease. Nowadays, in the world Heart disease is the major cause of deaths. The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year due to the
Heart diseases. In 2008, 17.3 million people died due to Heart Disease. Over 80% of deaths in world are because of Heart disease. WHO estimated by 2030, almost 23.6 million people will die due to Heart disease as written. Prediction by using data mining techniques gives us accurate result of disease. IHDP (intelligent heart disease prediction system) can discover and extract hidden knowledge associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus help healthcare analysts and practitioners to make intelligent clinical decisions which traditional decision support systems cannot. In this paper analysis of various data mining techniques given in tables which were used and helpful for medical analysts or practitioners for accurate heart disease diagnosis.

II. HEART DISEASE

The heart is important organ or part of our body. Life is itself dependent on efficient working of heart. If operation of heart is not proper, it will affect the other body parts of human such as brain, kidney etc. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it.

There are number of factors which increase the risk of Heart disease

- Family history of heart disease
- Smoking
- Cholesterol
- Poor diet
- High blood pressure
- High blood cholesterol
- Obesity
- Physical inactivity
- Hyper tension

Symptoms of a Heart Attack

Symptoms of a heart attack can include:

- Discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone.
- Discomfort radiating to the back, jaw, throat, or arm.
- Fullness, indigestion, or choking feeling (may feel like heartburn).
- Sweating, nausea, vomiting, or dizziness.
- Extreme weakness, anxiety, or shortness of breath.

Types of Heart diseases

Heart disease is a broad term that includes all types of diseases affecting different components of the heart. Heart means ‘cardio.’ Therefore, all heart diseases belong to the category of cardiovascular diseases. Some types of Heart diseases are

1. Coronary heart disease

It also known as coronary artery disease (CAD), it is the most common type of heart disease across the world. It is a condition in which plaque deposits block the coronary blood vessels leading to a reduced supply of blood and oxygen to the heart.

2. Angina pectoris

It is a medical term for chest pain that occurs due to insufficient supply of blood to the heart. Also known as angina, it is a warning signal for heart attack. The chest pain is at intervals ranging for few seconds or minutes.

3. Congestive heart failure

It is a condition where the heart cannot pump enough blood to the rest of the body. It is commonly known as heart failure.

4. Cardiomyopathy

It is the weakening of the heart muscle or a change in the structure of the muscle due to inadequate heart pumping. Some of the common causes of cardiomyopathy are hypertension, alcohol consumption, viral infections, and genetic defects.

5. Congenital heart disease

It also known as congenital heart defect, it refers to the formation of an abnormal heart due to a defect in the structure of the heart or its functioning. It is also a type of congenital disease that children are born with.

6. Arrhythmias

It is associated with a disorder in the rhythmic movement of the heartbeat. The heartbeat can be slow, fast, or irregular. These abnormal heartbeats are caused by a short circuit in the heart's electrical system.

7. Myocarditis
It is an inflammation of the heart muscle usually caused by viral, fungal, and bacterial infections affecting the heart. It is an uncommon disease with few symptoms like joint pain, leg swelling or fever that cannot be directly related to the heart.

III. LITERATURE SURVEY

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

An Intelligent Heart Disease Prediction System (IHDP) is developed by using data mining techniques Naive Bayes, Neural Network, and Decision Trees was proposed by SellappanPalaniappan et al. [3]. Each method has its own strength to get appropriate results. To build this system hidden patterns and relationship between them is used. It is web-based, user friendly & expandable.

To develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) a novel technique was proposed by HeonGyu Lee et al. [5]. To achieve this, they have used several classifiers e.g. Bayesian Classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree) and SVM (Support Vector Machine).

The prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks was proposed by Niti Guru et al. [4]. The dataset contains records with 13 attributes in each record. The supervised networks i.e. Neural Network with back propagation algorithm is used for training and testing of data.

The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez [7]. The resultant dataset contains records of patients having heart disease. Three constraints were introduced to decrease the number of patterns [6]. They are as follows:

1. The attributes have to appear on only one side of the rule.
2. Separate the attributes into groups. i.e. uninteresting groups.
3. In a rule, there should be limited number of attributes.

The result of this is two groups of rules, the presence or absence of heart disease.


Kiyong Noh et al. [8] uses a classification method for the extraction of multiparametric features by assessing HRV (Heart Rate Variability) from ECG, data pre-processing and heart disease pattern. The dataset consisting of 670 peoples, distributed into two groups, namely normal people and patients with heart disease, were employed to carry out the experiment for the associative classifier.

ShrutiRatnakar et al. used genetic algorithm to reduce the set of attributes of Naive Bayes generate relationship amongst the attributes.

AkhilJabbar et al. proposes efficient associative classification algorithm using genetic approach for heart disease prediction. The main motivation for using genetic algorithm in the discovery of high level prediction rules is that the discovered rules are highly comprehensible.

IV. DATA MINING TECHNIQUES USED FOR PREDICTIONS

The three different data mining classification techniques, i.e. Neural Networks, Decision Trees, and Naive Bayes are used to analyze the dataset.

4.1. Neural Networks

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural network. In other words, it is an emulation of biological neural system [13]. A Multi-layer Perceptron Neural Networks (MLPNN) is used. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of Yj = f (Σ wji xi) Where f is a simple threshold function such as sigmoid or hyperbolic tangent function.
4.2. Decision Trees

The decision tree approach is more powerful for classification problems. There are two steps in this technique: building a tree and applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these, J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces the size of the tree by removing overfitting data, which leads to poor accuracy in predictions. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

4.3. Naive Bayes

Naive Bayes classifier is based on Bayes theorem. This classifier algorithm uses conditional independence, meaning it assumes that an attribute value on a given class is independent of the values of other attributes. The Bayes theorem is as follows: Let \( X = \{x_1, x_2, \ldots, x_n\} \) be a set of \( n \) attributes. In Bayesian, \( X \) is considered as evidence and \( H \) be some hypothesis means, the data of \( X \) belongs to specific class \( C \). We have to determine \( P (H|X) \), the probability that the hypothesis \( H \) holds given evidence i.e. data sample \( X \). According to Bayes theorem the \( P (H|X) \) is expressed as \( P (H|X) = \frac{P (X|H) P (H)}{P (X)} \)

![Image of Naive Bayes algorithm implementation](image)

**V. TECHNIQUES USED IN DATA MINING**

**A. Association:**

Association is one of the best known data mining techniques. It involves discovering patterns in a dataset, typically based on the relationships between items in the same transaction. For example, the association technique is used in heart disease prediction as it tells us the relationship of different attributes used for analysis and sorts out the patient with all the risk factors which are required for prediction of disease.

**B. Classification:**

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of predefined classes or groups. Classification methods use mathematical techniques such as decision trees, linear programming, neural networks, and statistics.

**C. Clustering:**

Clustering is a data mining technique that makes meaningful or useful clusters of objects that have similar characteristics using automatic techniques. Different from classification, clustering techniques define the classes and associate objects in them, while in classification objects are assigned to predefined classes. For example, in the prediction of heart disease, using clustering techniques we get clusters or we can say that the list of patients which have the same risk factors. This means this makes the separate list of patients with high blood sugar and related risk factors and so on.

**D. Prediction:**

The prediction as its name implies is one of the data mining techniques that discovers relationships between independent variables and the relationship between dependent and independent variables. For instance, prediction analysis techniques can be used in sales to predict profit for the future if we consider sales as an independent variable, profit as a dependent variable. Then based on the historical sales and profit data, we can draw a fitted regression curve that is used for profit prediction.

**VI. METHODOLOGY USED IN DATA MINING**

Data Mining is core part of Knowledge Discovery Database (KDD). Many people treat Data Mining as a synonym for KDD since it’s a key part of KDD process. Knowledge discovery as a process is depicted in Figure 1 and consists of an iterative sequence of the following steps:
The overall objective is to study the various data mining techniques available to predict the heart disease and to compare them to find the best method of prediction. The objective of our 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 which has emerged in recent years for efficient and effective heart disease diagnosis. The analysis shows that different technologies are used in all the papers with taking different number of attributes. So, different technologies used shown the different accuracy to each other. In some papers it is shown that neural networks given the accuracy of 100% in prediction of heart disease.

On the other hand, this is also given that Decision Tree has also performed well with 99.62% accuracy by using 15 attributes [6]. So, different technologies used shown the different accuracy depends upon number of attributes taken and tool used for implementation. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amounts of data, researchers are using data mining techniques in the diagnosis of heart disease. Although applying data mining techniques to help health care professionals in the diagnosis of heart disease is having some success, the use of data mining techniques to identify a suitable treatment for heart disease patients has received less attention. As a future work, the researcher has planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types.

REFERENCES


