Abstract— Cardio vascular disease is a major threat to half of the world population. The term heart disease is related to all the diverse diseases affecting the heart. The healthcare industry generates large amount of data that are too difficult to be analysed by traditional methods. Hence computer assisted methods are necessary to make correct decisions. The objective of this paper is to develop a heart disease prediction system using data mining techniques. This helps to identify useful patterns of information from the medical data for quality decision making. Association rules and Classification techniques like Decision Tree, Naive Bayes and Neural Networks are used in literature for disease prediction. This work concentrates on building a classifier model using Decision Tree and Naïve Bayes for predicting heart disease. The system is implemented in .NET platform. The result obtained from the classifier enables to establish significant patterns and relationships between the medical factors relating to heart disease.

Keywords— Association rules, Decision Trees, Naïve Bayes, Neural networks.

I. INTRODUCTION

The computerization of various fields like Science and Engineering, Medicine, Business and Society has led to the explosive growth of available data. Data mining is the process of discovering interesting patterns and knowledge from such data. Data mining tools performs data analysis and may also uncover important data patterns contributing greatly to Knowledge bases, Business strategies, and Scientific and Medical research. The Data mining tasks can be classified into two categories Predictive and Descriptive. A Predictive model makes a prediction about values of data using known results found from different data and its goal is to identify strong links between variables of a data table. Predictive model data mining tasks involves the classification, prediction, time series analysis and regression. Examples include Disease diagnosis, detection of credit card fraudulent transactions. Descriptive model identifies patterns or relationships in data. It simply summarizes data in convenient ways or in ways that will lead to increased understanding of the way things work. The Healthcare industry is one among the most information intensive industries where the medical information, data, knowledge keep growing on a daily basis. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analysed by traditional methods. A century ago, it was noted that a symptom was related to a single disease. But, now a symptom is related to a number of diseases. Hence, Computer assisted information retrieval may help to support quality decision making and to avoid human error. Data mining can be a convenient tool to assist physicians in detecting the disease by obtaining knowledge and information regarding the disease from patient’s data. Thus quality services can be provided at affordable cost as some diagnostic and laboratory procedures are costly and painful to patients. By recognising the patterns, Doctors can prescribe relevant tests to the patients thereby providing treatment at affordable cost. The increased mortality rate and the shortage of specialists in many developing countries have motivated many researchers to use data mining techniques in disease diagnosis. Several works are carried out in Naïve Bayes, Neural networks and Decision Trees.

A. Classification Techniques:

Many Classification Techniques of Data mining are used for diagnosing Heart disease. Classification is a form of data analysis that extracts models describing important data classes. Such Classifiers predict categorical class labels. Classification has numerous applications including fraud detection, Target marketing, manufacturing and medical diagnosis [14]. Data classification is two step processes and consists of a learning step and a training step. In the learning step, the Classifier is built describing a predetermined set of data classes. The Classification step is used to predict the class labels for given data. There are different classification techniques like Decision Tree and Naïve Bayes. The Naïve Bayes classifier uses the Bayes rule of conditional probability. It takes all the attributes contained in the data, and analyses them individually like its equally important and independent of each other. The Neural network has three layers -input, hidden and output layer. It is used to train large amounts of data using Back propagation method. This paper
proposes a user friendly approach for heart disease diagnosis using Decision Trees and Naive Bayes. A decision tree is easy to interpret and has a considerable accuracy.

B. Heart Disease :
Heart disease is a general name for a variety of diseases, conditions and disorders that affect the heart and the blood vessels. Symptoms of heart disease vary depending on the specific type of heart disease. Congenital heart disease refers to a problem with the heart's structure and function due to abnormal heart development before birth. Congestive heart failure is when the heart does not pump adequate blood to the other organs in the body. Coronary heart disease or in its medical term Ischemic heart disease is the most frequent type of heart problem. Coronary heart disease is a term that refers to damage to the heart that happens because its blood supply is decreased, it leads to the fatty deposits build up on the linings of the blood vessels that provide the heart muscles with blood, resulting in them narrowing. The paper identifies the risk factors for the different types of heart diseases. The paper is organised as follows: Section 2 discusses the related works in the literature for heart disease. Section 3 explains the materials and methods used in this work. Section 4 deals with the experimental part and results. Section 5 summarises the results along with future work.

II. RELATED WORK

Numerous works [1, 10] related to heart disease diagnosis using data mining techniques have motivated this study. The dataset, algorithms, methods used by the authors and the observed results along with the future work are studied for each paper. Based on the observations, the data mining technique for the work is identified. M.Karaolis and C.S.Pattichis[3,7] have developed a data mining system for the assessment of heart event related risk factors. It is found that the data mining technique could help in the identification of high and low risk subgroups of patients. Decision Tree was used for extracting rules based on the risk factors. Sellassapan Palaniappan and Rafiah Awang[6] have developed an Intelligent Heart Disease Prediction System(IHDPs) using Decision Trees, Neural Networks and Naive Bayes. A web based, user friendly, scalable, reliable and expandable system was developed on the .NET platform. Anchantha Khemphila and Veera Boonjing [2] introduced a classification approach using Multi-Layer Perceptron (MLP) with Back-Propagation learning algorithm and a feature selection algorithm along with biomedical test values to diagnose heart disease. The Artificial neural networks was used to classify the diagnosis of patients. Kemal Polat and SalihGunes[8] have developed a hybrid approach to medical decision support systems. They have combined feature selection, fuzzy weighted pre-processing and AIRS. Carlos Ordenz et al[9,12] have introduced an improved algorithm to discover constrained association rules. Similar works are being carried by other researchers with different data mining techniques. Pavlopoulos[11] has observed the heart sound features and then classified using Decision trees. Tsein et al[13] have found that the classification trees and logistic regression have almost the same accuracy. Based on the observations, it has been found that Decision Tree Classifier classifies with considerable accuracy. It is easy to interpret and plays a significant role in diagnosing the presence or absence of Heart disease.

III. MATERIALS AND METHODS

A. Problem statement :
Heart disease prediction using data mining is one of the most interesting and challenging tasks. The shortage of specialists and high wrongly diagnosed cases has necessitated the need to develop a fast and efficient detection system. The main objective of this work is to identify the key patterns or features from the medical data using the classifier model. The attributes that are more relevant to heart disease diagnosis can be observed. This will help the medical practitioners to understand the root causes of disease in depth.

B. Description of the data :
The dataset is taken from the Cleveland Clinic Foundation and was supplied by Robert Detrano, M.D., Ph.D. of the V.A.Medical Center, Long Beach, CA. It is part of the collection of databases at the University of California, Irvine(UCI) collected by David Aha. The original dataset contains real representation of patient data and is extensively used for many testing purposes. The dataset consists of 14 attributes with the last attribute being the class attribute. The class attribute shows the presence or absence of disease. The ID3, C4.5 algorithms adopt a greedy approach in which decision trees are constructed in a top down recursive divide and conquer manner. The categorical and numerical attributes are handled by the C4.5 algorithm. It is an enhancement of ID3 algorithm where continuous attributes are used.

B. C4.5 Decision Tree algorithm :
The Decision Tree classifier takes the training set, attribute list and the splitting criteria method as inputs. A Decision tree is generated from which rules are predicted. Different attribute selection measures like Information Gain, Gain ratio, Chi square statistics, Gini Index and Distance measure can be used. The attributes can be reduced and then given to the model. By reducing the number of attributes, the algorithm can perform faster and efficient. In this work, the Information gain ratio is used as the splitting criteria. The attribute with the highest information gain is taken as the root of the tree. Information gain is based on Claude Shannon’s work on information theory. Info Gain of an attribute A is used to select the best splitting criterion attribute. The highest Info Gain is selected to build the decision tree. The formula is given as follows- Info Gain (A) = Info (D) – InfoA (D)
Where
\( A \) is the attribute investigated.

\[
\text{Info} (D) = \sum_{i=1}^{m} p_i \log_2(p_i)
\]

Where
- \( p_i \) = probability (class \( i \) in dataset \( D \));
- \( m \) = number of class values.

\[
\text{Info}_A (D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Info} (D_j)
\]

Where
- \(|D_j|\) = number of observations with attribute value \( j \) in Dataset \( D \);
- \(|D|\) = total number of observations in dataset \( D \);
- \( D_j \) = sub dataset of \( D \) that contains attribute value \( j \);
- \( v \) = all attribute values

**D. Construction of Bayesian Classifier :**

According to Bayesian theorem in statistics, every attribute is independent, so that the classifier can be simple and fast.

According to Bayesian theorem

\[
P(A|B) = P(A) \times P(B|A)/P(B)
\]

Where
- \( P(B|A) = P(A \cap B)/P(A) \)

Based on above formula, Bayesian classifier calculates conditional probability of an instance belonging to each class, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. In knowledge expression it has the excellent interpretability same as decision tree and is able to use previous data to build analysis model for future prediction or classification.

**E. Classifier evaluation measures :**

The four terms used in computing evaluation measures are used for evaluating the model and are described here. The True positives (TP) refer to the positive tuples that are correctly labelled by the classifier, while True negatives (TN) are the negative tuples that are correctly labelled by the classifier. False positives (FP) are the negative tuples that are incorrectly labelled by the classifier. False negatives (FN) are the positive tuples that were incorrectly labelled by the classifier. The confusion matrix is a useful tool for analysing how well the classifier can recognize tuples of different classes. Sensitivity is referred to as the true positive rate that is the proportion of positive tuples that are correctly identified.

\[
\text{Sensitivity} = \frac{TP}{P}
\]

Specificity is the true negative rate that is the proportion of negative tuples that are correctly identified. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. It is a function of specificity and sensitivity.

\[
\text{Accuracy} = \frac{TP + TN}{P + N}
\]

The Precision and Recall measures are widely used in classification. Precision is a measure of exactness (what percentage of tuples labelled as positive are actually positive). Recall is the measure of completeness. It is the same as sensitivity. The Receiver Operating Characteristics Curve can be used to compare two models and measure the efficiency.

**IV. EXPERIMENT AND RESULTS**

The overall System Architecture is shown in the figure. The Heart disease dataset is pre-processed by identifying the missing values, removing duplicates. A GUI is developed in the Microsoft .NET platform. The interconnection with certain java libraries are done through the IKVM interface. The Classifier model is developed and decision tree is generated. The risk factors are identified and the patterns are observed for the presence of heart disease. Finally the performance is evaluated using confusion matrix. The modules are identified as follows: Data pre-processing, Creation of Graphical User Interface in Microsoft .NET framework 4.0 and its interconnection with java libraries, Classification and performance evaluation. The system aims to use the powerful features of java and .NET. The IKVM interface is used for this inter platform interoperability. The interface has the java class libraries implemented in .NET and a convertor from JAR to DLL. The risk factors are assessed from the patterns of the classifier. Hence medical practitioners can prescribe certain tests relating to those attributes. These help in early diagnosis of diseases and also the patients are provided treatment economically.
A. Graphical User Interface:
The Graphical User interface is created using .NET framework 4.0. The IKVM interface is used to interconnect Microsoft .NET with librat.IKVM.NET and has a Java Virtual Machine implemented in .NET and tools that enable Java and .NET interoperability. The user interface views the number of attributes and instances in a different way to be easily handled by the medical practitioners. This serves as a convenient tool to help them to identify and assess the risk factors.

B. Classification:
The C4.5 algorithm is applied to generate the decision tree. The pre-processed dataset is given as input to the developed system. The given dataset is converted to the Attribute Relation File Format (ARFF). The J48 classifier in is used for classification. The results are obtained by using the full training set, cross validation and percentage split to classify the data. The tree form is also visualized. The input dataset is given to the system and the J48 classifier is used. J48 implements Quinlan’s C4.5 algorithm for generating a pruned or unpruned C4.5 decision tree. A feature of the J48 algorithm is that it “prunes” leaves that do not contribute greatly to the predictive accuracy of the tree. This creates smaller trees, which may be more resistant to over fitting. With IKVM.NET the compiled Java code (byte code) can be run directly on Microsoft .NET the byte code is converted on the fly to CIL and executed. The corresponding java class libraries are converted into appropriate dynamic link libraries and they are called sequentially. Thus, the user can open an existing ARFF file through the GUI and visualise the attributes and values. The count of instances and attributes are displayed. The graphical user interface helps the user to view the attributes in a convenient way. The tree structure and the graphical structure can also be visualised.

C. Tree Visualisation:
The tree is dynamically generated and the nodes along with the captions are displayed. It has been found from the result that the asymptotic chest pain type is more vulnerable to heart disease. The root and the type are also displayed. The tree is also displayed in the user interface. The patterns and risk factors are observed from the tree. The decision tree aids easy interpretability and hence medical practitioners can easily observe the patterns.

D. Bayesian Classifier:
The identified risk factors are classified by the naïve Bayesian Classifier. The probabilities for both classes are calculated and the class with the highest probability is returned. Thus the classifier predicts if the person with specified symptoms will have heart disease or not.

E. Performance evaluation:
The confusion matrix for the model is obtained for the classifier. A graph is plotted with the true positive rate in the X axis and false positive rates in the Y axis. Receiver Operating Characteristic curves are a useful visual tool in comparing two classification models [14]. An ROC curve for a given model shows the trade-off between the true positive rate and false positive rate. The ROC curve for the decision tree classifier is shown in the figure. The ROC is compared for the Decision tree and Naive Bayes models. It has been found that the area under the ROC of Decision tree is 0.807. It is found to be better than Naive Bayes. Hence this shows the accuracy of the classifier model. The sensitivity, specificity, precision and recall values are also measured and found to be considerably fair than Naive Bayes. The performance of the classifier is evaluated and thus the results are found.

V. CONCLUSION
Heart disease is a fatal disease by its nature and the misdiagnosis can cause serious and life threatening complications such as cardiac arrest and death. The aim of the study is to design a predictive model for heart disease detection using data mining techniques. The importance of Data mining in the Medical domain is realised and steps are taken to apply the relevant techniques in Disease prediction. The various research works done by different people are studied. The observations from the previous works have led to the deployment of the proposed system architecture for this work. Although various classification techniques are widely used for disease prediction, the Decision Tree Classifier is selected for its simplicity and accuracy. Large dataset can be applied in the classifier for future work. The Graphical
User Interface must enable data entry in format other than ARFF and automatic conversion to ARFF must be possible. Also further investigations in attribute selection measures before classification are still needed.

REFERENCES


