# A Survey on Predicting The Heart Disease Using Decision Tree Approach

**S.Vidyavathi[1],Mrs.D.Shona[2]**
[1]Dept of Computer Science
[2]Assistant Professor, Dept of Computer Science
[1, 2] Sri Krishna arts and science college,coimbatore-641008

***Abstract-*** *At the present time all the activities are done in the world through internet. The use of computer in the field of medicine are highly improved. The computerized hospital includes the activities like treatment of illnesses, maintain the patient's information up-to-date that leads to handle huge amount of data regularly. It is very difficult to handle the large amount of data for predicting the heart disease. Data mining approach is a good way for predicting the heart disease at minimum effort. This paper analyses the existing prediction system and discussing the various disputes on the existing system. The result of this paper helps the health consultants to diagnose the disease in less time and predict the probable problems well in advance and save the patient life.*

***Keywords***- Data mining, Heart diseases, classification.

## I. INTRODUCTION

Data mining turn into the ultimate method for finding the practical solutions to the day to day problems, and the health care is no exception to this. Most of the data mining methods are developed to help clinicians for making better decisions about the patient treatment purposes. Nowadays many data mining techniques are used in the areas such as healthcare organizations, health informatics, epidemiology, patient care and monitoring systems, assistive technology and large-scale image analysis to information extraction and automatic identification of unknown classes. The goal of this paper is to analyze several datamining techniques that are existed in recent years for the diagnosis of heart disease. Many researchers used datamining techniques in the diagnosis of diseases such as tuberculosis, diabetes, cancer and heart disease, in which several data mining techniques are used in the diagnosis of heart disease such as KNN, Neural Networks, and Bayesian classification. Classification based on clustering, Decision than just heart attacks. It applies to a number of conditions that affect the heart, including coronary artery disease, arrhythmias, atrial fibrillation, heart valve disease, congenital heart disease, cardiomegaly (enlarged heart), cardiomyopathy (heart muscle disease), and more. Heart disease is the number one cause of death in the United States for both men and women. Sometimes, symptoms can be subtle and go unnoticed until a major event like a heart attack occurs. Noticeable symptoms include: chest pain (angina), extreme fatigue, and shortness of breath. Certain lifestyle habits and risk factors contribute to heart disease. Some risk factors like age and gender can't be controlled. However, others can be. To keep your heart healthy, it's a good idea to lower good a blood pressure, eat a high-fiber, low-fat diet, exercise regularly, manage stress, and quit smoking.

### A. Types of heart diseases:

Heart disease is a word used to describe many different conditions affecting the heart. Coronary heart disease is a common type of heart disease. This condition results from a buildup of plaque on the inside of the arteries, which reduces blood flow to the heart and increases the risk of a heart attack Tree, Genetic Algorithm, Naive Bayes, Decision tree, WAC [Women's Army Crops] which are showing accuracy at different levels. Using medical profile such as age, sex, blood pressure and blood sugar can easily predict the likelihood of patients getting heart disease.

## II. HEART DISEASE

Heart disease labels that is a series of conditions that affects the heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems. The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease pointed or choked blood vessels that can leads several problem like a heart attack, chest pain (angina) or stroke. In most of the countries death is caused by the heart diseases that is surveyed by WHO (World Health Organization) and the CDC [Centre of Disease Control]. The adults of the countries UK, USA, Canada and Australia distressed from heart disease. The term heart disease refers to more and other heart complications. Other forms of heart disease include:

- irregular heartbeat (arrhythmias)
- congenital heart defects
- weak heart muscles (cardiomyopathy)

- heart valve problems
- Heart infections
- cardiovascular disease

**B. Heart disease statistics**

Approximately 610,000 people die from heart disease in the United States every year, according to the Centers for Disease and Control Prevention (CDC) .It's the leading cause of death in both men and women. Coronary heart disease is the deadliest of all heart diseases, just as it's the most common form. The Heart Foundation estimates 380,000 related deaths per year. The symptoms of heart disease vary between gender. Some are more obvious in men, who made up more than half of all heart disease-related deaths in the United States in 2009, according to the CDC. According to The Heart Foundation, 1 in 3 women die of heart disease every year in the United States. In 90 percent of these cases, women had at least one preventable risk factor.

**C. Symptoms of heart disease**

Heart disease is often called a silent killer. Your doctor may not diagnose the disease until you show signs of a heart attack or heart failure. Symptoms of heart disease vary depending on the specific condition. For example, if you have a heart arrhythmia, symptoms may include:

- a fast or slow heartbeat
- dizziness
- lightheadedness
- chest pains
- shortness of breath

Symptoms of a congenital (present at birth) heart defect may include skin discoloration, such as a bluish or pale color. You may also notice swelling in your legs and stomach. You might become easily tired or have shortness of breath shortly after beginning any type of physical activity.

**D. Heart Failure**

Heart failure happens when the heart isn't pumping enough blood to meet your body's needs. While many people believe the misconception that heart failure means an individual is about to die or that their heart has stopped, this is not true. Heart failure simply indicates that the heart is not squeezing as well as it should. It usually does not occur suddenly but gradually worsens over the time. Heart failure can be caused by:

- Cardiomyopathies
- Coronary Artery Disease
- Diabetes
- Diseases of the Heart Valves
- Heart Defects present at Birth
- High Blood Pressure
- Lung Disease such as Emphysema
- Past Heart Attacks

## III. DATA MINING

There is an unstoppable growth in the amount of electronic health records or EHRs being collected by healthcare facilities. It has been the norm for nurses to take responsibility in handling patient data input that was traditionally recorded in paper-based forms. Accuracy is extremely important when it comes to patient care and computerizing this massive amount of data enhances the quality of the whole system. Data mining has been used to uncover patterns from the large amount of stored information and then used to build predictive models. Improving the quality of patient care and reducing healthcare costs are the ideal goals of many programs. Data mining has helped these programs succeed.

**Benefits of Data Mining in the Healthcare Industry**

While other solutions might favor healthcare providers or insurance companies, data mining benefits everyone concerned, from healthcare organizations to insurers to patients. Patients receive more affordable and better healthcare services. This happens when healthcare officials use data mining programs to identify and observe high-risk patients and J. Shreve, H. Schneider, O. Soysal proposed a paper. A methodology for comparing classification methods through the assessment of model stability and validity in variable selection. Decision Support Systems. In this paper the author compares the performance of classification methods by using Monte Carlo simulations and illustrates that the variable selection process is integral in comparing methodologies to ensure chronic diseases and design the right interventions needed. These programs also reduce the number of claims and hospital admissions, further streamlining the process. HealthCare providers use data mining and data analysis to find best practices and the most effective treatments. These tools compare symptoms, causes, treatments and negative effects and then proceed to analyze which action will prove most effective for a group of patients. This is also a way for providers to develop the best standards of care and clinical best practices.

Insurers are now able to better detect medical insurance abuse and fraud because of data mining. Unusual claims patterns are easier to spot with this tool and it can identify inappropriate referrals and fraudulent medical and insurance claims. When insurers reduce their losses due to fraud, the cost of health care also decreases.

Healthcare facilities and groups use data mining tools to reach better patient-related decisions. Patient satisfaction is improved because data mining provides information that will help staff with patient interactions by recognizing usage patterns, current and future needs, and patient preferences.

## IV. LITERATURE SURVEY

In 2004, Carlos Ordonez [1], surveyed set of information which enveloped restorative records of the general population of coronary illness with qualities for the hazard elements, estimations of heart perfusion and limited conduit. The three imperatives were acquainted with decrease the quantity of examples, they are as per the following: 1) the attributes need to show up on one side of the as it was. 2) The rule isolates the characteristics into the uninteresting gatherings.3) The quantity of qualities from the rule is controlled by the medicinal records of the general population of coronary illness at long last. Additionally falling the running time according to the trials the limitations of indicating standards have been amazingly diminished the number. The author, Carlos Ordonez, foresaw the nearness or the nonappearance of coronary illness in four particular heart supply routes into the two gatherings of guidelines. Information mining techniques may help the clinicians in the expectation of the survival of patients and in the adaption of the practices subsequently.

In 2004, Franck Le Duff etal [2] build an efficient decision tree for executing medical procedure. Data mining methods may aid the clinicians in the prediction of the survival of patients, the comparison of traditional analysis and data mining analysis which sorted the variables. This provided the importance of data and variables for building a decision tree. In 2006, Kiyong Noh etal [3], has used classification method for extracting patterns from database. Assessing the heart rate variability from ECG of 670 people and grouped them into normal people and people with heart disease. This method is developed to identify people with heart disease based on the clinical data.

In 2006, Boleslaw Szymanski etal[4], "Using Efficient Supanova Kernel For Heart Disease Diagnosis" used heuristic method for computation of sparse kernel in SUPANOVA. It was applied to a benchmark Boston housing market dataset to detect the heart disease. The non invasive measurement which was based on magnetic field generated by human heart. Support vector machine was used for obtaining results which was accurate.

In year 2008, SellappanPalaniappan, etal[5] performed a work, “Intelligent Heart Disease Prediction System Using Data Mining Techniques". In this paper he created Intelligent Heart Disease Prediction System (IHDPS) utilizing data mining methods, i.e. Decision Trees, Naïve Bayes and Neural Network. Every strategy has its own particular energy to increase appropriate outcomes. The shrouded examples and connections among them have been utilized to build this framework. The IHDPS is easy to understand, web-based, scalable, versatile, and expandable.

RovinaDbritto, AnuradhaSrinivasaraghavan [11] in their paper, Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods ―concern with four algorithms like Naïve Bayes, Decision tree , K nearest neighbor and Support vector machine. the author suggest 3 phases. In this the size of the data set will be increased by using first two algorithms and predict the heart disease by using logistic regression [11].

In year 2012, Chaitrali S. Dangare, etal. [6] performed a work, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”. This paper has examined expectation frameworks for heart disease utilizing more number of attributes. The framework utilizes terms, for example, cholesterol, gender, blood pressure, like credits to anticipate the likelihood of patient getting a coronary illness. In this exploration work included two more properties i.e. obesity and smoking. The data mining classification techniques, namely Decision Trees, Naïve Bayes, and Neural Networks are analyzed on heart disease database.

In 2012, Akhil Jabbar etal [7], proposes proficient arrangement calculation utilizing hereditary approach for coronary illness forecast. The fundamental inspiration for utilizing hereditary calculation in the revelation of abnormal state forecast tenets is that the found standards are exceedingly intelligible, having high predictive precision and of high intriguing quality qualities. This paper goes for investigating diverse information mining methods that has been presented as of late for Heart Disease Prediction framework by various authors.

In 2012, R. Bhuvaneshwari and K. Kalaiselvi [8], used Naïve Bayes classifier method, and back propagation neural network. These algorithms which calculates the

probability of all objects from past experience. The posterior from prior is calculated by using Bayes rules based on the probability model.

In 2012, NidhiBhatla et al[9], used several data mining techniques namely Decision tree, Naïve Bayes and Neural Network. This helped in developing a prototype Intelligent Heart Disease Prediction system by using several attributes. The clinical data wasobtained from Cleveland heart disease database for the research purpose. The analysis show that the provides results accurately.

## V. TECHNIQUES USED FOR PREDICTION OF HEART ATTACK

**Decision Trees**

The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree & 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 size of tree by removing over fitting data, which leads to poor accuracy in predications. 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.

### *A. preprocessing*

The actions comprised in the preprocessing of a data set are the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields. Moreover it might be essential to combine the data so as to reduce the number of data sets besides minimizing the memory and processing resources required by the data mining algorithm [5]. In the real world, data is not always complete and in the case of the medical data, it is always true. To remove the number of inconsistencies which are associated with data we use Data preprocessing. The Pre-process panel has facilities for importing data from a database, a comma-separated values (CSV) file, etc., and for pre-processing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.

### *B. Classification*

The records with irrelevant data were removed from data warehouse before mining process occurs. Data mining classification technology consists of classification model and evaluation model. The classification model makes use of training data set in order to build classification predictive model. The testing data set was used for testing the classification efficiency. Then the classification algorithm like decision tree, naive Bayes and neural network was used for stroke disease prediction[3]. The performance evaluation was carried out based on Decision Tree algorithms and accuracy was measured. The Classify panel enables applying classification and regression algorithms (indiscriminately called classifiers in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, receiver operating characteristic (ROC) curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).

In classification method it is unable to predict the dataset accurately example (while testing old record of heart patient in current generation there occurs a problem).So, there are two main ways to predict the heart attacks;

1. By family Gene
2. By cholesterol

By using classification method, trained dataset in family gene affected by heart attack can predict the value by 68, and by cholesterol value 72.

### *C. Decision tree*

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making.

In Decision tree, while comparing the classification method dataset can be more efficiency and accurate.

1. Predicting the disease by family gene.

2. Predicting disease by cholesterol level.

**PROPOSED METHOD** Predicting the cholesterol level in decision tree, affected person can be get the heart attack by the value 70.

## VI. PROPOSED APPROACH

Proposed system is the medical sector application. This system helps to predict of heart disease, based on manually inserted inputs. Inputs are like patient's age, gender, chest pain, the resting blood pressure in mmHg, serum cholesterol in mg/d, hereditary, fasting blood pressure in mg/dl, thal and smoking. The database contains all the information of the user i.e. the user details and the admin uploaded files, these will be compared to produce an output. Output is consisting either patient has heart disease or not.

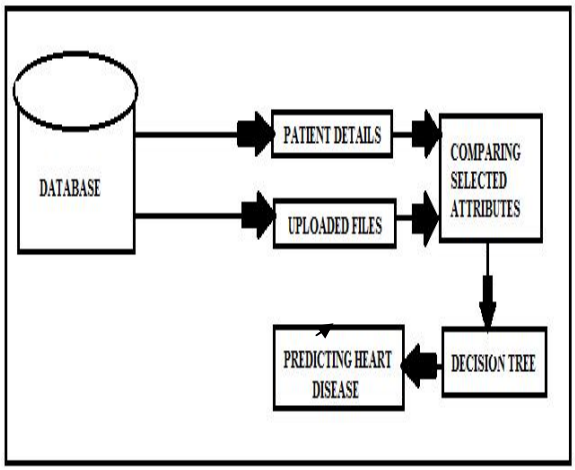


Fig1: Proposed system diagram

From the fig 1, the JSP java server pages will connect the user and admin to the database. Database consists of user details and admin uploaded files. Database is a storage device which has the capacity to hold huge data. In this proposed system it consists of patient details and uploaded files, and the attribute is selected. Then for selected the decision tree algorithm tree is applied to predict the heart disease.

## VII. CONCLUSION

This paper presented a review of disease prediction for healthcare system using data mining techniques. Medical related information's are volumetric in nature and it could be derived from different birthplaces which are not entirely applicable in a feature. In this work, we have performed a literature survey on various papers. In future, we are planning to propose an effective disease prediction system to predict the heart disease with better accuracy using different data mining classification techniques such as Decision Tree. Decision-tree algorithm is one of the most effective and efficient classification methods available. It has been shown that, by using a decision tree, it is possible to predict heart disease vulnerability in diabetic patients with reasonable accuracy. Classifiers of this kind can help in early detection of the vulnerability of a diabetic patient to heart disease. Preprocessing of a data set for the removal of duplicate records, normalizing the values used to represent information in the database. Clustering technique, simple k-means algorithm is used. Thus, the patients can be forewarned to change their lifestyles. This will result in preventing diabetic patients from being affected by heart diseases, thereby resulting in low mortality rates as well as reduced cost on health care for the state. This can be extended in future to predict other types of ailments which arise from diabetes, such as visual impairment. The proposed work can be further enhanced and expanded, to use stacking techniques to increase the accuracy of decision trees and reduce the number of leaf nodes.

## REFERENCES

[1] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.

[2] Franck Le Duff, CristianMunteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, pp. 1256-9, 2004.

[3] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee, and KeunHoRyu, "Associative Classification Approach for Diagnosing Cardiovascular Disease",Springer, Vol:345, pp: 721- 727, 2006.

[4] Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, KarstenSternickel, Lijuan Zhu, "Using Efficient Supanova Kernel for Heart Disease Diagnosis", proc. ANNIE 06,intelligent engineering systems through artificial neural networks, vol. 16, pp:305-310, 2006.

[5] SellappanPalaniappan, RafiahAwang, Intelligent Heart Disease Prediction System Using Data Mining Techniques; 978-1-4244-1968-5/08/$25.00©2008 IEEE.

[6] Chaitrali S. Dangareet. al., "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", (IJCA) (0975 – 8887), Vol.47, No. 10, June 2012, page no. 44-48.

[7] M.Akhiljabbar, Dr.Priti Chandra, Dr.B.LDeekshatulu, Heart Disease Prediction System using Associative Classification and Genetic Algorithm,InternationalConference on Emerging Trends

in Electrical, Electronics and Communication Technologies, 2012.

[8] R.Bhuvaneshwari and K .Kalaiselvi,"Naïve Bayesian Classification approach In Healthcare Applications", International Journal of Computer Science and Telecommunication ",vol 3,no 1,pp.106-112 , 2012.

[9] NidhiBhatla, KiranJyoti, " An Analysis of Heart Disease Prediction using Different Data Mining Techniques" International Journal of Engineering and Technology Vol.1 issue 8 2012.

[10] Shadab Adam Pattekari and AsmaParveen ,"Prediction System for Heart Disease Using Naïve Bayes", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624,Vol 3,Issue 3,2012,pp290-294.

[11] RovinaDbritto, AnuradhaSrinivasaraghavan, Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods.

[12] Lemke F, Mueller J-A. Medical data analysis using self-organizing data mining technologies, Systems Analysis Modeling Simulation. 2003; 43(10):1399–408.

[13] Parthiban L, Subramanian R. Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences.2008; 3(3).

[14] Li W, Han J, Pei J. CMAR: accurate and efficient classification based on multiple association rules. Proceedings of 2001 International Conference on Data Mining; 2001.

[15] Shreve J, Schneider H, Soysal O. A methodology for comparing classification methods through the assessment of model stability and validity in variable selection. Decision Support Systems. 2011; 52:247–57.

[16] Wanga T, Huang H, Tian S, Xu J. Feature selection for SVM via optimization of kernel polarization with Gaussian, ARD kernels.. Expert Systems with Applications.2010; 37:6663–8.

[17] Patil SB, Kumaraswamy YS. Extraction of significant patterns from heart disease warehouses for heart attack prediction. International Journal of Computer Science and Network Security (IJCSNS). 2009; 9(2):228–35.

[18] DeepikaN,.Chandrashekar K. Association rule for classification of Heart Attack Patients. International Journal of Advanced Engineering Science and Technologies. 2011; 11(2):253–57.

[19] Srinivas K, Rani KB, Govrdhan A. Application of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering. 2011; 2(2):250–5.

[20] P. Chandra, M. Jabbar, and B. Deekshatulu, Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection,‖ in 12th International
Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 628–634.

[21] Usha. K Dr, Analysis of Heart Disease Dataset using neural network approach, IJDKP, Vol 1(5), Sep 2011. [12]AnkitaDewan, Meghna Sharma, Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification, 2nd International Conference on Computing for Sustainable Global Development IEEE 2015 pp 704-706. [13].
www.kmd.ovgu.de/kdd15_medical_mining_tutorial.html
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2916206/

[22] www.medicalnewstoday.com/articles/237191.php [16]http://www.healthline.com/health/heart-disease/type.