Abstract — The diagnosis of heart disease from various symptoms is a major issue which is not free from false presumptions. The healthcare industry gathers large amount of heart disease data that is not mined to determine the useful information for effective decision making by healthcare practitioners. The effort to utilize knowledge and experience of specialists and data of patients collected in database is a valuable option. Data Mining using a variety of techniques for decision making knowledge in the database and extracting these in a way that they can use in areas such as decision support, predictions, estimation. This research will provide an intelligent heart disease prediction system (IHDPMS) able to help a physician as well as a health care system. In this research, the efficiency of heart disease system will enhance using if then else rules classification, fuzzy c means clustering and genetic algorithm. Various parameters like accuracy, time, specificity and sensitivity are calculated. The proposed algorithm provides better accuracy as compared to traditional algorithms.

Keywords: Data Mining, Disease Diagnosis, Heart Disease, classification using if then else rules, Fuzzy c means clustering, genetic algorithm.

I. Introduction

Predicting the outcome of disease is one of the most interesting and challenging task in data mining. The knowledge discovery database (KDD) process includes data mining techniques has become a popular research tool for medical researchers and it is able to predict the outcome of a disease using historical data records of patients. Firstly the heart disease was thought to be the problem of developed countries but now it is problem for developing countries too. Recent studies have documented that high cardiovascular diseases are reported in coal mining regions. The risk for cardiovascular disease is influenced by environmental, behavioural, health services variables. The automation of the system would be extremely beneficial for us. Regrettably all the doctors do not possess experience in every sub speciality and more ever there is shortage of resource persons at certain places [6]. Appropriate computer based information and decision system can aid in achieving test at the reduced cost. Today most countries face high and increasing rate of heart disease. It has become a leading cause of death. Due to increase in world’s population, the health care industries are facing many challenges and issues based on patient’s severity is to be reduced and detect it earlier in more efficient way. To save the life of patients and reduce the health care cost the medical error should be prevented [14].

Table 4 Description of attributes

<table>
<thead>
<tr>
<th>S.NO</th>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Age of patient in years</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
<td>Male, female</td>
</tr>
</tbody>
</table>
| 3    | Cp        | Chest Pain Type  
1. Angina=Typical Angina  
2. Abnang=atypical Angina  
3. Notang=Non Anginal Pain  
4. Asympt-asymtomatic |
| 4    | Trestbps  | Resting Blood pressure(in mm Hg on admission to the hospital) |
| 5    | Chol      | Serum cholesterol in mg/dl |
| 6    | Fbs       | Fasting blood sugar(fbs>120=true, fbs<120=false) |
| 7    | Restecg   | Resting electrocardiographic results are Hyper, Normal Abnm(abnormal) |
| 8    | Thalach   | Maximum heart rate achieved |
| 9    | Exang     | Exercise induced angina(Yes=true, No=false) |
| 10   | slope     | The slope of the peak exercise ST segment, Up, Flat, down |
| 11   | Family    | History True, false |
| 12   | ca        | Number of major vessels(0-3) colored by fluoroscopy |
| 13   | No. cigrates per day | No. of cigrates per day |
| 14   | No. of years a smoker | Number of years a smoker |
Heart Disease: The term heart disease encompasses the diverse disease that affects the heart. Before defining heart disease it is better to define what a heart is and functions performed by heart. Heart is one of the most important organs in our body. Essentially a pump, the heart is a muscle made up of four chambers separated by valves and divided into four chambers called an atrium and one called a ventricle. The arteries collect blood and the ventricles contract to push blood out of the heart. The right half of the heart pumps oxygen poor blood to the lungs where blood cells can obtain more oxygen. Then newly oxygenated blood travels from the lungs into the left atrium and left ventricle pumps the newly oxygen rich blood to the organs and tissues of the body. This oxygen provides body with energy and is essential to keep our body healthy.

Heart disease is a general name applies to number of illness, disorders, conditions that effect the circulatory system which consist of heart and blood vessels[4].Symptoms of heart disease depend on the specific type of heart disease. Classic symptom of heart disease is chest pain. Chest pain arises when the blood received by heart muscle is inadequate [4]. Sometimes there may be no symptoms in some people until life threatening complications occur. Early signs of heart disease are dizzy spell, discomfort following meals especially if long continued, shortness of breath, fatigue, pain or tightness in chest, palpitation etc[4]. There is common set of risk factors that influence whether someone will ultimately be at risk for heart disease or not. These risk factors include age, gender, cholesterol, obesity, sedentary life style etc.

Coronary heart disease: Most frequent type of heart disease is coronary heart disease. It is the main cause of the heart attack. coronary heart disease occurs when oxygen and blood supply to the heart is decreased. A sudden blockage of a coronary artery generally due to a blood clot results in a heart attack[4].There are few factors that are responsible for coronary artery disease high cholesterol that can increase fat concentration in our blood create build up of fatty deposits.

Fuzzy c means clustering: In hard clustering data is divided into distinct clusters where each data element can belong to one clusters like k means clustering. In fuzzy clustering data is divided into distinct clusters, one data element can belong to one or more clusters. Fuzzy c means clustering is also known as soft clustering. In fuzzy c means clustering each element is associated with set of membership level.

- Unlike K-means clustering where each data point belongs to exactly one cluster but in fuzzy c means data element can belong to more than one cluster.

- Gives best result for overlapped dataset and comparatively better than K-means algorithm.

II. Research Methodology

Diagnostic test are used to determine the presence or absence of disease. Confusion matrix is the primary source of measurement in classification problems. Classification instance classifies each instance into two classes either positive or negative. This gives four classifications for each instance: a true positive, a true negative, a false positive, a false negative. Given m classes ,a confusion matrix is a table of at least size m by m. When peoples are tested for a disease the test outcome can be positive(sick) or negative (healthy) while the actual status of the patient may be different then following four conditions may occur.

Accuracy: The overall accuracy of a classifier is estimated by dividing the total number of correctly classified instances by the total number of instances [26]

\[
\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FN+FP)}
\]

In this research Slope, Exercise induced angina and no. Of major vessels coloured by fluoroscopy are three main factors after testing these factors other factors are considered.

1. Healthy people correctly identified as healthy called "True negative (TN)".
2. Healthy people wrongly identified as sick called “false positive (FP)’’. 
3. Sick people correctly identified as sick called true “positive (TP)’’. 
4. Sick people wrongly diagnosed as healthy called “false negative (FN)’’.

<table>
<thead>
<tr>
<th>Test result</th>
<th>Actual condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>present</td>
<td>Positive</td>
</tr>
<tr>
<td>condition</td>
<td>Positive result = True Positive</td>
</tr>
<tr>
<td></td>
<td>Condition absent + Positive result = False Positive</td>
</tr>
<tr>
<td>absent</td>
<td>Negative</td>
</tr>
<tr>
<td>condition</td>
<td>Negative result = False Negative</td>
</tr>
<tr>
<td></td>
<td>Condition absent + Negative result = True Negative</td>
</tr>
</tbody>
</table>

Table 1 Confusion Matrix
The numbers along the diagonal from upper-left to lower-right represent the correct decisions made, and the numbers outside this diagonal represent the errors. Accuracy, specificity and sensitivity are the basic performance measurements.

**Sensitivity:** Sensitivity refers to the proportion of people with disease who have a positive test result. Specificity is ratio of number of true Positives to the number of true positives plus number of false negatives.

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

**Specificity:** Specificity refers to the proportion of people without the disease who have a negative test result. Specificity is the ratio of number of true negatives plus number of false positives.

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]

### III. Result

Experimental results deals with the output. The output result of proposed algorithm i.e. applied on Cleveland heart disease dataset is shown. Cleveland database has 303 records but in this dissertation only 58 records are used. Missing value records are discarded. In order to analyze the performance comparison will be made with existing algorithms.

![Fig 1.2 Accuracy comparison between proposed and existing algorithm](image)

![Fig 1.3 No. of instances used in existing algorithm and proposed algorithm](image)

### IV. Conclusions

In this research work performance of heart disease prediction system is evaluated. Data is collected from Cleveland heart disease dataset. This dataset contains total 303 instances and 76 attributes. In this research work 58 records and 14 attributes are used. Data mining Classification using if then else rules, fuzzy c means clustering, genetic algorithms are used in this research work to improve the accuracy of the system. The performance of the heart disease prediction system is evaluated using performance measures accuracy, Time, specificity, Sensitivity etc. The prediction system shows 86.6% accuracy, 32 milliseconds time, 0.44 specificity, 0.45 sensitivity.

### V. Future Work

In future to increase the performance of heart disease prediction system more datasets can be used. Sensitivity and specificity can be further improved. In future intelligent heart disease prediction system can be build that can work on specific type of heart disease.

### VI. References


[18] Usma Fayad, Gregory Piatetsky-Shapiro andPadhraic Smyth “The KDD process useful for extracting useful Knowledge from large volumes of data”.


[21] Daniel T. Larose “Discovering knowledge in Data”, An Introduction to data mining, Wiley


