CHAPTER V

APPLICATIONS OF FEATURE SELECTION TO HEART DATA

5.1 Introduction

Heart disease is any condition that impairs the functioning of the heart; it is a broad term that includes variety of heart conditions. Heart disease is a term that refers to several diseases of the heart and circulatory system including coronary heart disease, myocardial infarction, congestive heart failure, and other conditions. Despite decades of declining death rates, heart disease remains the first leading causes of death for men and women. Heart diseases can sneak up on human like a thief in the night. According to World Health Organization (WHO), heart disease is the leading cause of death in the United Kingdom, United States of America, Canada and Australia. 25.4% of all deaths i.e., one in every four deaths in the United States of America today are caused by heart disease. Traditionally, heart disease was thought to be the problem of developed countries, but now it is becoming a headache for developing countries too and it is especially devastating for the developing countries since they do not have adequate health care. The term “heart disease” refers to several types of heart conditions. The most common type is coronary artery disease, which can cause heart attack. Heart disease is an umbrella term for any type of disorder that affects the heart. Cardio Vascular Disease (CVD) refers to disorders of the blood vessels and heart, while heart disease refers to just the heart.
5.1.1 Shocking Statistics about Heart Disease:

- In the world, CVD deaths account for one third of all deaths (25-50% depending on the level of economic development) among which 50%: coronary deaths.
- CVD made up 16.7 million of global deaths in 2002, among which 7 million due to coronary heart disease, 6 million due to stroke.
- Global cardiovascular deaths in 2002: 16.7 million among which: coronary heart disease 7.2 million > stroke 6.0 million > 0.9 million hypertensive heart disease > 0.4 million inflammatory heart disease > 0.3 million rheumatic heart disease > 1.9 million other CVD.
- In 2005 CVD accounted for approximately 38 percent of all deaths.
- More that 2,500 Americans die from CVD each day.
- Among women, 1 in 2.6 deaths from CVD.

WHO estimated that by 2020 Cardio Vascular Disease will be the cause of over 40 per cent deaths in India as compared to 24 per cent in 1990. With over 3 million deaths owing to cardiovascular disease every year, India is set to be the 'heart disease capital of the world' in few years [61].

An Alarming Statistics from India said that one fifth of the deaths in India are from coronary heart disease. By the year 2020, it will account for one third of all deaths. Sadly, many of these Indians will be dying young. Heart disease in India occurs 10 to 15 years earlier than in the west. There are an estimated 45 million patients of coronary
artery disease in India. An increasing number of young Indians are falling prey to coronary artery disease. With millions hooked to a roller-coaster lifestyle, the future looks even grimmer.

5.1.2 Major Types of Heart Disease

There are many types of heart disease, but in this study the researcher chooses to discuss five types that are common to happen. As described by Cindy (2008) five common types of heart disease are discussed below.

- Atherosclerosis
- Coronary Heart Disease
- Pulmonary Heart Disease
- Congestive Heart Failure
- Congenital and Rheumatic Heart Disease
- Stroke

Atherosclerosis

Atherosclerosis is characterized by deposits of plaque such as fatty substances, cholesterol, cellular waste products, calcium, and fibrin in the inner lining of the arteries. It starts due to elevated levels of cholesterol, high blood pressure, and tobacco smoke.
Blood flow and oxygen supply to the heart can be reduced or even fully blocked with accumulating plague [136]. Due to Atherosclerosis, decreased blood supply particularly to the cerebrum and lower extremities. If the plaque prolongs and sustainability then at the end blood vessel may rupture. It is explicitly shown in Figure 5.2 and how the Atherosclerotic Plaques is Developed and deposited in artery in illustrated Figure 5.3.

**Coronary Heart Disease (Ischemic Heart Disease)**

Coronary heart disease or in its medical term Ischemic heart disease is the most frequent type of heart problem of all, and is also the leading reason of heart attacks. Coronary heart disease is a term that refers to damage to the heart that happens when its blood supply is decreased, fatty deposits build up on the linings of the blood vessels that provide the heart muscles with blood, resulting in them narrowing. These narrowing decreases the blood supply to the heart muscles and causes pain which is identified as angina. Factors those responsible causes of coronary heart disease are high cholesterol builds up of fatty deposits and cigarette and tobacco smoke.

**Pulmonary Heart Disease**

Pulmonary heart disease comes from a lung, or pulmonary, disorder, or a complication of lung problems where the blood flow into the lungs is slowed or even totally blocked, resulting in increased pressure on the lungs. There are a number of different symptoms that typically come with pulmonary heart disease, such as shortness of breath, syncope, dyspnea, and chest pain. It is a state which is often misdiagnosed, and has frequently progressed to late stages by the time that it is actually correctly diagnosed.
It has been previously chronic and untreatable with a poor survival rate. However, there are now numerous new treatments which are accessible which have extensively improved the overall prognosis of this disease.

**Congenital and Rheumatic Heart Disease**

Congenital heart disease affects 1 out of 125 children. may be due to hereditary factors, maternal diseases, or chemical intake (alcohol) during fetal development. Rheumatic heart disease results from rheumatic fever, which affects connective tissue. Rheumatic heart disease frequently derives from strep throat infections. This can be a reason for alarm for many because strep throat, while often preventable, is a quite common condition that affects many people who do not treat a minor sore throat infection in time. However, there is no reason to be because rheumatic heart disease that comes from strep throat is fairly rare. Actually, the sheer volume of cases of rheumatic heart disease has decreased considerably since 1960.

**Congestive Heart Failure**

Congestive heart failure, damaged or overworked heart muscle is unable to keep blood circulating normally. It affects over 5 million Americans. Damage to heart muscle may result from rheumatic fever, pneumonia, heart attack, or other cardiovascular problem. Lack of proper circulation may allow blood to accumulate in the vessels of the legs, ankles, or lungs.
Stroke

Stroke occurs when the blood supply to the brain is suddenly cut off, which can occur when a blood vessel in the brain or neck is blocked or bursts. Brain cells are then deprived of oxygen and die. A stroke can result in problems with speech or vision or can cause weakness or paralysis. Most strokes are caused by fatty deposits or blood clots—jelly-like clumps of blood cells—that narrow or block one of the blood vessels in the brain or neck. A blood clot may stay where it formed or can travel within the body. People with diabetes are at increased risk for strokes caused by blood clots. A stroke may also be caused by a bleeding blood vessel in the brain, called an aneurysm, a break in a blood vessel can occur as a result of high blood pressure or a weak spot in a blood vessel wall [56].

Percentage Breakdown of Deaths from Cardiovascular Disease in the United States is portrayed in Figure 5.1.

![Percentage of Deaths from Heart Disease in the United States](image)

**Figure 5.1: Percentage of Deaths from Heart Disease in the United States**
5.1.3 Risk Factors

Risk factors are conditions or habits that make more likely to develop heart disease. The more risk factors that the greater chances of getting heart disease [85], such as age or family history which cannot be changed and some risk factors are controllable. Risk factors are further classified into four categories. Reduction in these risk factors could reduce much of the burden and disability caused by heart disease. They are described below:

1. Major modifiable risk factors

   - High blood pressure
   - Abnormal blood lipids
   - Tobacco use
   - Physical inactivity
   - Obesity
   - Diabetes mellitus

2. Minor modifiable risk factors

   - Low socioeconomic status
   - Mental ill health (depression)
   - Psychosocial stress
   - Heavy alcohol use
   - Use of certain medication
   - Lipoprotein
3. Non-modifiable risk factors

- Age
- Heredity or family history
- Gender
- Ethnicity or race

4. Novel risk factors

- Excess homocysteine in blood
- Inflammatory markers (C-reactive protein)
- Abnormal blood coagulation (elevated blood levels of fibrinogen)

Heart disease is diagnosed based on the above risk factors. This research considers the dataset from UCI machine learning Repository to concern the heart disease diagnosis[126]. The dataset is explicated in the following section.

5.2 Heart Dataset

The data are collected from the Cleveland Clinic Foundation, and it is available at the UCI machine learning Repository [126]. Six instances containing missing values have been deleted from the original dataset. This dataset includes continuous, binomial, nominal, and ordinal features. A data frame with 297 observations of 13 conditional attributes and 1 decision attribute, which refers to the presence of heart disease in the
patient. The conditional attributes are age, sex, chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting electrographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels colored by flourosopy and thal. The decision attribute has the values 0 to 4, 0 denotes healthy, 1,2,3,4 denotes sick. The heart dataset description is tabulated in Table 5.1

### Table 5.1 Heart Dataset Description

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age in years</td>
</tr>
<tr>
<td>Sex</td>
<td>(1 = male; 0 = female)</td>
</tr>
<tr>
<td>Cp</td>
<td>Chest Pain Type (value 1: typical angina, value2: atypical angina, value3: non-angina pain, value 4: Asymptomatic)</td>
</tr>
<tr>
<td>Trestbps</td>
<td>Resting blood pressure</td>
</tr>
<tr>
<td>Chol(mg/dl)</td>
<td>Serum Cholesterol</td>
</tr>
<tr>
<td>Fbs</td>
<td>Fasting Blood Sugar (value 1: &gt;120 mg/dl; value 0:&lt;120 mg/dl)</td>
</tr>
<tr>
<td>Restecg</td>
<td>Resting electrographic results (value 0: normal; value 1: having ST-T wave Abnormality; value 2: showing probable or definite left</td>
</tr>
<tr>
<td>Thalach</td>
<td>Maximum heart rate achieved</td>
</tr>
<tr>
<td>Exang</td>
<td>Exercise induced angina (value 1: yes; value 0: no)</td>
</tr>
<tr>
<td>Old peak</td>
<td>ST depression induced by exercise relative to rest</td>
</tr>
<tr>
<td>Slope</td>
<td>Slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping)</td>
</tr>
<tr>
<td>CA</td>
<td>Number of major vessels colored by floursopy (value 0-3)</td>
</tr>
</tbody>
</table>
5.3 Performance Analysis

Data mining can be a useful tool in the health sector and healthcare. Organizations that perform data mining are better positioned to meet their long-term needs. Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. In recent years new research avenues such as knowledge discovery in databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables, and be able to predict the outcome of a disease using the historical cases stored within datasets.

Performance analysis of existing and proposed feature selection algorithms, exemplified in chapters 3 and 4 concerned on heart dataset. Experiments are carried out to evaluate the performance of the proposed algorithms. The Existing Genetic Algorithm, Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm have been implemented using MATLAB for heart dataset available in the UCI data repository.

In the dimension of feature selection, the data set initially has 13 attributes, Genetic Algorithm selected 6 features such as, Cp - Chest Pain Type, Trestbps (mmhg)- Resting blood pressure, Fbs - Fasting Blood Sugar, Exang - Exercise induced angina, Oldpk - Old peak CA - No. of vessels colored by floursopy and Thal - Maximum heart rate achieved and reduced input attributes by Compound Featuristic Genetic Algorithm are Cp - Chest Pain Type, Trestbps (mmhg)- Resting blood pressure, Exang - Exercise induced angina and CA - Number of vessels colored by floursopy. Among the thirteen
features, more important features are selected via the Core Featuristic Genetic Algorithm. Reduced attributes achieved for heart data set after applying the Core Featuristic Genetic Algorithm are: Cp - Chest Pain Type, Trestbps (mmhg)- Resting blood pressure, Fbs - Fasting Blood Sugar, Exang - Exercise induced angina and CA - Number of vessels colored by floursopy.

In heart data set, among 13 attributes, only 4 and 5 attributes had been chosen by Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm respectively for decision making. Even though, Compound Featuristic Genetic Algorithm has picked minimal number of attributes, Core Featuristic Genetic Algorithm has selected five attributes i.e. the extra attribute (Fbs - Fasting Blood Sugar) are required to predict heart disease. The Comparative Analysis of Genetic Algorithm with Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm are tabulated in Table 5.2.

Table 5.2: A Comparative Analysis of Genetic Algorithm with Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Number of Reduced attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic Algorithm</td>
<td>6</td>
</tr>
<tr>
<td>Compound Featuristic Genetic Algorithm</td>
<td>4</td>
</tr>
<tr>
<td>Core Featuristic Genetic Algorithm</td>
<td>5</td>
</tr>
</tbody>
</table>
Figure 5.2 shows the graphical representation of reduced number of features of prediction of heart disease diagnosis, based on the heart dataset.

Figure: 5.2: Performance Analysis of the Genetic Algorithm with Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm.
Discussion

The proposed feature selection algorithms are applied on Heart dataset. For heart data set, along with 13 attributes, Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm select 4 and 5 attributes by Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm correspondingly for predicting Heart disease. The proposed methods are commonly established to carry out better than existing algorithms in terms of reduced number of features.

5.4 Summary

To reiterate, Genetic Algorithm, Compound Featuristic Genetic Algorithm, and Core Featuristic Genetic Algorithm are contributed for feature selection crisis bumped into on Heart dataset. Core Featuristic Genetic Algorithm opts for predicting Heart disease with essential features. In toting up, Feature Selection is connected with classification. Chapter VI will be dealt on classification methods in terms of accuracy.