CHAPTER 3
COMPUTER-AIDED MI DETECTION SYSTEM

3.1 INTRODUCTION

The existing techniques for cardiovascular disease detection system were discussed in the previous chapter. CAD schemes using digital data processing techniques have the goal of improving the detection performance, and they are described in detail in this section. Typically, CAD systems are designed to provide a second opinion to aid rather than replacing the cardiologist.

![Figure 3.1 Intelligent MI prediction system](image)

Figure 3.1 shows that the proposed approach for detection of MI. Heart disease data analysis using a CAD system is an extremely challenging task. Since CAD systems are computer-directed systems, there is a need for a flawless system. The patient without external symptom can also have the possibility of MI. Hence there is a need to analyze the attributes of the HD data set and an intelligent classifier
to classify the abnormalities among the data. The system has been designed in a framework of MATLAB 7.10, and it aims at developing an intelligent system for MI detection.

3.2 DATA SET DESCRIPTION

Heart attack data set is obtained from UCI centre for machine learning and intelligent systems (Murphy 2004). The Cleveland heart disease data was obtained from V.A. Medical Center, Long Beach and Cleveland Clinic Foundation from Dr. Robert Detrano. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them.

The data collected from 270 patients are used for proposed work. The digitized data has 150 normal and 120 abnormal cases. In this data set, the first 13 attributes describe the risk factors of ischaemic heart disease, and the last attribute describes the output class. There are two output classes for the diagnosis of heart attack. In the selected data set, class 0 specifies the absence of heart attack, and class 1 specifies the presence of heart disease. The data set contains the data in the age range between 25 and 75, and it also contains the data of women as well as men.

A real-time clinical study is taken to analyze the abnormality of the ischaemic heart disease patient. The data is observed from the patient suffering from heart attack and normal patient. The real-time clinical data obtained are also distributed among various age group people and both genders. The data set having the same attribute of Cleveland data set is collected from 500 patients, and it has 300 normal and 200 abnormal cases. Traditionally, ischaemic heart disease has been divided into several symptoms. The 13 input attributes considered in this experimentation are age, gender, chest pain, resting blood pressure, cholesterol, blood sugar, resting
electrocardiographic results, maximum heart rate achieved, exercise induced angina, old peak, ST slope, number of major vessels colored and thallium test.

Age plays a major role in the prediction of MI. If age increases, the risk of damaged and narrowed arteries also increased. In this experimentation, the age is selected in the range 25-75. Men are generally at greater risk of heart disease. However, the risk for a woman increases after menopause. At age 75, a woman's risk for IHD is equal to that of a man's. IHD is the leading cause of death and disability in women after menopause. In the data set taken, ‘Male’ is encoded as ‘0’ and ‘Female’ is encoded as ‘1’.

The most common symptom of MI is angina or angina pectoris also known simply as chest pain. Stable angina is felt as predictable sensation of chest pain. Angina is said to be unstable when the symptom pattern worsens abruptly in terms of frequency and duration without an obvious cause of increased oxygen consumption. Asymptomatic chest pain occurs at any time even at rest. In the Cleveland data set, typical type 1 angina is encoded as ‘1’, typical type 2 angina as ‘2’, non-angina pain as ‘3’, and asymptomatic chest pain is encoded as ‘4’. Blood pressure is recorded as two numbers with a ratio like 120/80 mmHg.

Hypertension refers to high blood pressure. High blood pressure causes scarred arteries that fill up with plaque and become more prone to blood clots. In the selected data set, blood pressure is recorded from 90 to 190.

The cholesterol profile includes LDL cholesterol, HDL cholesterol, triglycerides and total cholesterol. High levels of HDL cholesterol indicate lower risk of developing cardiovascular disease. An HDL level of 60 mg/dL and over is considered excellent to provide optimal protection. Total cholesterol is a measure of LDL cholesterol, HDL cholesterol and other lipids. The desirable level of total
cholesterol is less than 200mg/dL. In Cleveland data set, serum cholesterol ranges from 160-410 mg/dL.

Diabetes mellitus is defined as fasting blood glucose of greater than 125 mg/dL or more. Diabetes increases the risk for developing cardiovascular disease. Diabetes can damage nerves as well as blood vessels, and so a heart attack can be silent that is lacking the typical chest pain. In the selected data set, 120 mg/dL is considered as threshold value, fasting blood sugar > 120 mg/dl is encoded as ‘1’, and fasting blood sugar < 120 mg/dl is encoded as ‘0’.

![Figure 3.2 ECG wave with its intervals and segments](image)

The ECG signal is a representation of the bioelectrical activity of the heart representing the cyclical contractions and relaxations of the human heart muscles. As seen in Figure 3.2 the first deflection termed as the P wave in ECG is due to the depolarization of the atria. The large QRS complex is due to the depolarization of the ventricles. This is the complex with largest amplitude and is easy to detect. The portion of the ECG from the end of the QRS complex to the beginning of another ECG is termed as the ST segment.
The ST segment should not be elevated or depressed. Any change in ST from baseline may indicate cardiac disease such as ischaemia. The normal ECG is encoded as ‘0’, ECG having ST-T wave abnormality is encoded as ‘1’, and probable or definite left ventricular hypertrophy is encoded as ‘2’.

An arrhythmia is an irregular heart rhythm. A normal heart rate is 50 to 100 beats per minute. Arrhythmias can occur with a normal heart rate or with heart rates that are slow or rapid. Abnormal heart rate reflects abnormal tone, and it is associated with higher risk. Heart rate is recorded between 71-202 in the selected Cleveland data set.

The exercise stress test has been used for decades as a diagnostic tool in the workup of patients with suspected coronary artery disease. In Cleveland data set, exercise-induced angina is represented as ‘1’, and ‘0’ represents the absence of angina during treadmill test.

The three types of modalities are dynamic, static and resistive. Exercise-induced ST segment depression does not localize the site of MI. The magnitude of ST depression at peak exercise also indicates the existence of severe disease. Major ST depression at low workload indicates severe disease.

A new paradigm that emerged with the result of thrombolysis in MI benefitted patients with ST elevation. Slowly up-sloping ST segment depression (>1.5 mm, 80msec from the J point) usually indicates MI. Horizontal ST segment depression is considered as abnormal response. Down-sloping ST segment depression represents severe MI. In the selected data sets, value 1 represents unsloping, value 2 represents horizontal ST wave, and value 3 represents downsloping ST.
Angiography or arteriography is a medical imaging technique used to visualize the inside or lumen of blood vessels and organs of the body with particular interest in the arteries, veins and the heart chambers. Angiography serves to investigate normal and pathological states of the vessel system particularly luminal narrowing and obstruction or aneurysmal widening. The number of major vessels (0-3) colored by fluoroscopy can be obtained from coronary angiogram. It is used to examine the entire coronary anatomy.

A thallium stress test is a nuclear imaging test that shows how well blood flows into the heart during exercise and at rest. It can identify areas of the heart that may have a reduced blood supply as a result of damage from a previous myocardial infarction or blocked coronary arteries. In the HD data set, normal value obtained in thallium stress test is coded as ‘3’, fixed defect is encoded as ‘6’, and reversible defect is encoded as ‘7’. Sample set of data attributes from Cleveland data set used for MI prediction is shown in Table 3.1.

<table>
<thead>
<tr>
<th>Data 1</th>
<th>Data 2</th>
<th>Data 3</th>
<th>Data 4</th>
<th>Data 5</th>
<th>Data 6</th>
<th>Data 7</th>
<th>Data 8</th>
<th>Data 9</th>
<th>Data 10</th>
<th>Data 11</th>
<th>Data 12</th>
<th>Data 13</th>
</tr>
</thead>
<tbody>
<tr>
<td>60.0</td>
<td>1.0</td>
<td>4.0</td>
<td>130.0</td>
<td>206.0</td>
<td>0.0</td>
<td>2.0</td>
<td>132.0</td>
<td>1.0</td>
<td>2.4</td>
<td>2.0</td>
<td>2.0</td>
<td>7.0</td>
</tr>
<tr>
<td>56.0</td>
<td>1.0</td>
<td>1.0</td>
<td>120.0</td>
<td>193.0</td>
<td>0.0</td>
<td>2.0</td>
<td>162.0</td>
<td>0.0</td>
<td>1.9</td>
<td>2.0</td>
<td>0.0</td>
<td>7.0</td>
</tr>
<tr>
<td>71.0</td>
<td>0.0</td>
<td>4.0</td>
<td>112.0</td>
<td>149.0</td>
<td>0.0</td>
<td>0.0</td>
<td>125.0</td>
<td>0.0</td>
<td>1.6</td>
<td>2.0</td>
<td>0.0</td>
<td>3.0</td>
</tr>
<tr>
<td>58.0</td>
<td>0.0</td>
<td>1.0</td>
<td>150.0</td>
<td>283.0</td>
<td>1.0</td>
<td>2.0</td>
<td>162.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>3.0</td>
</tr>
<tr>
<td>35.0</td>
<td>1.0</td>
<td>4.0</td>
<td>126.0</td>
<td>282.0</td>
<td>0.0</td>
<td>2.0</td>
<td>156.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>7.0</td>
</tr>
<tr>
<td>55.0</td>
<td>0.0</td>
<td>4.0</td>
<td>180.0</td>
<td>327.0</td>
<td>0.0</td>
<td>1.0</td>
<td>117.0</td>
<td>1.0</td>
<td>3.4</td>
<td>2.0</td>
<td>0.0</td>
<td>3.0</td>
</tr>
<tr>
<td>48.0</td>
<td>1.0</td>
<td>4.0</td>
<td>130.0</td>
<td>256.0</td>
<td>1.0</td>
<td>2.0</td>
<td>150.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>2.0</td>
<td>7.0</td>
</tr>
<tr>
<td>44.0</td>
<td>1.0</td>
<td>4.0</td>
<td>110.0</td>
<td>197.0</td>
<td>0.0</td>
<td>2.0</td>
<td>177.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>3.0</td>
</tr>
<tr>
<td>63.0</td>
<td>1.0</td>
<td>1.0</td>
<td>145.0</td>
<td>233.0</td>
<td>1.0</td>
<td>2.0</td>
<td>150.0</td>
<td>0.0</td>
<td>2.3</td>
<td>3.0</td>
<td>0.0</td>
<td>6.0</td>
</tr>
<tr>
<td>54.0</td>
<td>0.0</td>
<td>3.0</td>
<td>135.0</td>
<td>304.0</td>
<td>1.0</td>
<td>0.0</td>
<td>170.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>3.0</td>
</tr>
<tr>
<td>41.0</td>
<td>0.0</td>
<td>2.0</td>
<td>126.0</td>
<td>306.0</td>
<td>0.0</td>
<td>0.0</td>
<td>163.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>
3.3 PRE-PROCESSING

The goal of pre-processing the data is to simplify recognition of the patient with high risk factor without throwing away any important information. Real world data are generally incomplete, noisy and inconsistent. Normalization is one of the data transformation techniques used to scale attribute values to fall within a specified range.

In this work, CAD system is designed based on intelligent technique, and the tool used is neural network. Because neural networks work internally with numeric data, the binary data and categorical data must be encoded in numeric form. Additionally, experience has shown that in most cases numeric data such as a person's age should be normalized. In the proposed work, neural network is trained with back propagation algorithm. When using back propagation networks, depending on the activation function of the neurons it will be necessary to perform some pretreatment of data used for training. If logistic sigmoid function is used, the interval of variation of the output variables has to be accommodated to the maximum output range, i.e. from zero to one.

An adequate normalization not only for the network output variables but also for the input ones previous to the training process is very important to obtain good results and to reduce significantly calculation time. Hence the input attributes in the selected data set are normalized before classification.

The min-max normalization is used, and it performs a linear alteration on the original data. The values are normalized within zero and one. Normalization is one of the steps used in data pre-processing, and normalizing the data attempts to give all attributes an equal weight. If the neural network with back propagation algorithm is
used for prediction normalizing input values for each attribute measured in the training tuples will help speed up the learning phase.

The normalization technique used in intelligent MI prediction is min-max normalization. Let $A$ be a numeric attribute with $n$ observed values, $v_1, v_2, \ldots, v_n$. Min-max normalization performs a linear transformation on the original data. Suppose that $min_A$ and $max_A$ are the minimum and maximum values of an attribute $A$. Min-max normalization maps a value $v_i$ of $A$ to $v_i'$ in the range $new - min_A, new - max_A$ by computing

$$v_i' = \frac{v_i - min_A}{max_A - min_A} \left( new_{max} - new_{min} \right) + new_{min}$$

Min-max normalization preserves the relationship among the original data values. The values are normalized between the limit $[0, 1]$ and then Equation (3.1) becomes

$$v_i' = \frac{v_i - min_A}{max_A - min_A}$$

3.4 CLASSIFICATION

Classification is a form of data analysis that extracts models describing important data classes. Classification has numerous applications including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis. Data classification is a two-step process consisting of a learning step where a classification model is constructed and a classification step where the model is used to predict class labels for the given data.

In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the training phase where a classification algorithm builds the classifier by analyzing a training set made up of database tuples and their
associated class labels. A tuple \( X \) is represented by an \( n \)-dimensional attribute vector \( X = (x_1, x_2, x_3, \ldots, x_n) \), depicting \( n \) measurements made on the tuple from \( n \) database attributes \( A_1, A_2, A_3, \ldots, A_n \), respectively.

Each tuple \( X \) is assumed to belong to a predefined class as determined by another database attribute called the class label attribute. The class label attribute is discrete-valued and unordered. It is categorical in that each value serves a class. The individual tuples making up the training set are referred to as training tuples and are randomly sampled from the database under analysis, and this kind of learning is called as supervised learning. In the second step, the model is used for classification. First, the predictive accuracy of the classifier is estimated, and the test set made up of test tuples is used to analyze the classification accuracy.

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The associated class label of each test tuple is compared with the learned classifier’s class prediction for that tuple. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known. Classification is a classic data mining technique based on machine learning.

Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification one makes the software that can learn how to classify the data items into groups. Classification is most widely used by all the knowledge discovery approaches. Patterns that are extracted using machine intelligence can be used to predict which class the data falls under.

Classification falls under two categories, supervised and unsupervised. The time complexity of unsupervised classifier is high, and hence supervised
classification approach is used in the CAD system. The ultimate objective of automated methods for classification of abnormalities in HD data set is to provide a tentative diagnosis and the final decision about the MI produced by a human expert based on their physical attributes. These methods are incarnations of a generic model of supervised classification systems.

According to this model, a classifier is presented with attributes obtained from a selected database that are to be classified in a process known as training. The trained classifier can later label objects which were not used in its training, an ability known as generalization. The performance of classification methods depends on the type and quality of the features employed to train the classifier. Three most popular classifiers used for classification are FFNN, CNN and SVM.

ANN is a powerful tool for pattern recognition. Though ANN performs well when applied for classification, the global classifiers based on static ANN have not performed well in practice. The back propagation is most likely to get trapped into local minima, making it entirely dependent on the initial settings. Depending on the number of hidden layer neurons, the performance of ANN varies, and hence optimal selection is required. The momentum constant helps in accelerating convergence of error propagation algorithm. The learning rate also can improve the chances of reaching the global minimum when unusual pair of training patterns is presented. Hence soft computing optimization algorithms are used to optimize the hidden layer neurons, learning rate and momentum constant. The random population-based algorithms used for optimizing the ANN to improve the classification accuracy are Genetic Algorithm and Particle Swarm Optimization Algorithm.
3.5 SUMMARY

The proposed MI prediction system and the modules on it are discussed in this chapter. The input attributes considered for our research and their significance on MI prediction are also analyzed in this chapter. In order to improve the training process and accuracy, this research work investigates novel intelligent classifiers that use data attributes as input to classify the normal and the abnormal, and this is discussed in Chapter 4.