Fuzzy expert system approach for coronary artery disease screening using clinical parameters

Debabrata Pal a, K.M. Mandana b, Sarbajit Pal c, Debranjan Sarkar c, Chandan Chakraborty a,⇑

a School of Medical Science and Technology, IIT Kharagpur, India
b Cardiothoracic & Vascular Surgery, Fortis Hospitals Ltd., Anandapur, Bypass, Kolkata, India
c Variable Energy Cyclotron Center, Department of Atomic Energy, Kolkata, India

ABSTRACT

Coronary artery disease (CAD) affects millions of people all over the world including a major portion in India every year. Although much progress has been done in medical science, but the early detection of this disease is still a challenge for prevention. The objective of this paper is to describe developing of a screening expert system that will help to detect CAD at an early stage. Rules were formulated from the doctors and fuzzy expert system approach was taken to cope with uncertainty present in medical domain. This work describes the risk factors responsible for CAD, knowledge acquisition and knowledge representation techniques, method of rule organisation, fuzzification of clinical parameters and defuzzification of fuzzy output to crisp value. The system implementation is done using object oriented analysis and design. The proposed methodology is developed to assist the medical practitioners in predicting the patient’s risk status of CAD from rules provided by medical experts. The present paper focuses on rule organisation using the concept of modules, meta-rule base, rule address storage in tree representation and rule consistency checking for efficient search of large number of rules in rule base. The developed system leads to 95.85% sensitivity and 83.33% specificity in CAD risk computation.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Coronary artery disease (CAD) has become one of the mostly occurring diseases in the world and has increasing trend in its incidence in future. It is the cause of 20–30% of deaths in most industrialised countries [1]. According to WHO (World Health Organisation) report, CAD has become a modern epidemic. It is estimated that by 2010 India is expected to have 60% of world’s heart disease and that in India problems of heart are increasing in younger generation. According to recent research at least 20 million people suffer from heart attacks [2]. In India it is also reported that CAD tends to occur earlier in life in Indians than other ethnic groups. CAD has multi-factorial aetiology with many of the risk factors influenced by lifestyle.

CAD is the result of accumulation of plaques within the coronary arteries. These arteries supply the myocardium (the muscle of heart) with oxygen-rich blood. The plaque is made up of fat, cholesterol, calcium and other substances. This condition is called Atherosclerosis. When the plaque is growing it narrows the lumen of the coronary arteries. Consequently the blood flow to the heart muscle decreases. This causes a discomfort or a pain, the pain may be felt in chest, neck, jaws, abdomen, arms and shoulder also called Angina. During Angina the amount of the oxygenated blood flow decreases. But as the disease progresses the lumen of coronary arteries goes on narrowing due to the increased size of plaque. Hence the amount of blood supplied to this tissue becomes inadequate to supply the needs of the tissue. This condition is called myocardial ischemia and the tissues do not work at its fullest capacity. When the lumen of coronary artery has near-complete blockage, severely restricting the flow of oxygenated blood, the tissue in the areas of myocardium dies leading to myocardial infarction (particularly known as heart attack) which also accounts for sudden death. Although CAD has now become much familiar disease, but death rate is high due to the lack of awareness among the common people. Generally the patients neglect the initial symptoms and they only consult medical experts when those symptoms become severe or critical. But at that time the treatment becomes complicated and sometimes due to the acuteness and severity of the disease, the patients die before getting proper medication. Another situation is that there is not enough facility to diagnosis in the countryside like India and treat the patients soon.
And also the people living in the countryside become afraid of going to doctors of highly specialized hospitals. So the principal problem is that patients always go to the hospital later and the treatment becomes difficult at that time for preventing the disease.

In order to circumvent such a burning problem one of the possible solutions is to make the people aware of their respective CAD risks in advance and take preventive measures accordingly. It is only possible when an early detection of CAD occurs. According to the medical experts an early detection at the stage of angina may prevent the death due to CAD if proper medication is given thereafter. Here lies the importance of developing a CAD screening system.

In view of this, the goal is to develop CAD screening expert system based on primary risk factors. In the developing countries like India where population explosion is a major concern, manual screening system for CAD patients throughout the country is not an easy task, because the number of medical experts in comparison with number of patients is less. One way to deal with this problem is to build an intelligent decision support system which can mimic the reasoning of doctors called as expert system. Till date many medical expert systems have been developed like MYCIN [3], INTERNIST [4], CADIAG-2 [5]. But these are multiple disease expert systems. Research on risk factor analysis [6–11] has been done. Studies on 18-lead electrocardiogram (ECC) analysis [11], 12-lead ECC analysis [12–15], Spectrum analysis [16], wavelet transformed ECC analysis [17], Exercise stress testing analysis [18], Myocardial Infarction analysis [19], heart valve disease analysis [20,28] have been done extensively for detection of cardiovascular diseases. Single lead ECC classification methodology is proposed using time domain principal components [21–23] and discrete wavelet transformation domain principal components [24–26]. Various methods for Q point, R point, S point extraction from ECC are discussed in [27] for automated ECC analysis towards coronary heart disease detection. Different data mining approaches have been studied based on classification tree and logistic regression [30]. comparison of association rules and decision tree [31], association rules and rule filtering [32]. Kocurani and Aji [35] proposed a methodology for CAD diagnosis based on neuro-fuzzy model. In another study decision tree was used to find preliminary rules and based on the extracted rules a fuzzy model [36] was developed for CAD detection. Karolis et al. [29,37] developed a data mining system for investigating three coronary heart events—myocardial Infarction, Percutaneous coronary intervention and coronary artery bypass graft. A prototype intelligent heart disease prediction system was developed by Palaniappan and Awang [38]. Kurt et al. [39] did a comparative study of logistic regression, classification and regression tree, neural network, radial basis function and self organising feature maps for developing models to predict absence or presence of CAD. Karolis et al. [40] applied association rule analysis to develop a data mining system for CAD prediction. Kajabadi et al. [41] and Tu et al. [42] applied classification and regression tree and bagging algorithms respectively to develop CAD predicting models. Ji et al. [45] proposed a fuzzy k-prototype clustering algorithm and applied their methodology on heart data. An agent based approach to neural network training was proposed by Czarnowski and Je_drzejowicz [46] for CAD detection model. Sebastian and Patrick [47] proposed a methodology for domain driven KDD for extracting functionally novel rules. But from the available literatures, no clinical risk factors based expert system has been reported till now in India to screen the patients at an early stage. The present system has been developed based on the clinical risk factors of CAD, incorporating the experience of the medical experts. Here an attempt has been made to develop a knowledge base on cardiac risk factors and finally an inference engine to compute risk in terms of percentage.

The key contribution of this paper is development of a low cost dedicated fuzzy rule based expert system for CAD risk stratification using clinical and first level information from the patients. Every input parameter is fuzzified using linguistic variables; rules are framed based on cardiologist’s expertise on linguistic variables (like ‘low’, ‘medium’, ‘high’). Based on different conditions the fuzzy rules are fired, if multiple rules gets fired then all are aggregated using defuzzification strategy like centroid method.

2. Methodology

The proposed methodology here consists of knowledge acquisition, design of CAD screening system using fuzzy rule based inference, rule organisation and finally application of object oriented approach to the development of entire system. Each of these individual blocks is explained in a systematic manner as followings.

2.1. Knowledge acquisition

Knowledge acquisition (KA) is one of the most difficult and time-consuming part for developing an expert system. The main goal of KA in our work is to efficiently acquire the specific knowledge for CAD from a group of medical experts and representing the acquired knowledge in a computer perceptible form. The paper by Bose [33] describes different methods of knowledge acquisition. Structured interview method has been applied for the KA of the present work. Multiple phases of interviews with the domain experts from Advanced Medical Research Institute (AMRI), Salt Lake, Kolkata have been done by the knowledge engineers. The risk factors responsible for CAD are identified. A well defined questionnaire is developed advised by domain experts and sample data are collected from the patients.

2.1.1. Identification of CAD risk factors

The risk factors identified, are mainly based on Indian population and they have been identified from the past experience and the expertise of the medical experts. These factors are age, smoking habit, having obesity, hypertention, diabetes, and hyperlipidemia. Symptoms considered for first level information are angina pectoris or chest pain.

2.1.1.1. Age. Age is a non-modifiable risk factor. CAD risk increases with age significantly. The mortality of acute MI increases significantly for the aged people.

2.1.1.2. Smoking. It is quantified by the number of cigarettes per day. Toxins in the blood due to smoking of cigarettes contribute to the development of plaque in the coronary arteries resulting in obstruction of blood flow and causing heart attack. According to INTERHEART study smoking is a greater risk in younger men than women. Smoking more than 10 cigarettes carries an independent fourfold risk of acute myocardial infarction.

2.1.1.3. Obesity. It is one of the leading risk factors in Indian population. With the advent of economic changes, increased mechanisation and high rates of urbanisation major changes in lifestyle patterns have occurred leading to a trend towards decreasing physical activity resulting in increase in obesity. It is quantified by BMI (Body Mass Index) where it is defined as

\[
\text{BMI} = \frac{\text{Weight}}{\text{Height}^2}
\]

2.1.1.4. Hypertension. The prevalence of CAD is significantly higher among the hypertensive patient compared with non-hypertensive patient. The cause of hypertension is elevation of arterial blood pressure that leads to shortening of life expectancy. The high blood
pressure is commonplace affecting one in four men and one in five women. Hypertension rates increase with age, rising from 3.4% in the 20–39 age group to 51.6% in the 60–79 age group. Here to evaluate risk, both systolic blood pressure (SBP) and diastolic blood pressure (DBP) have been considered.

2.1.1.5. Diabetes. It is heterogeneous disease that results from a complex interplay between genes and environment. It is characterised by hyperglycemia and occurs due to impaired insulin sensitivity. CAD accounts for more than 50% of the mortality among diabetic patients. India alone is projected to experience the greatest global increase in diabetes by 2025. For our expert system design fasting blood sugar (FBS) level has been considered for quantification of diabetes.

2.1.1.6. Lipid profile. The lipid profile increases the risk of CAD manifold. There are different constituents of blood cholesterol along with low density lipoprotein (LDL) contributing to risk of CAD. High quantity of low density lipoprotein (LDL) increases the risk of CAD whereas high quantity of TC level and ratio of TC to high density lipoprotein (HDL) cholesterol increase the risk of CAD. Lipid profile has been quantified by TC, tri-glyceride (TGL), LDL and HDL.

2.1.1.7. Chest pain. Chest pain which is also known as angina pectoris, is an important symptom associated with CAD. It may be accompanied by other symptoms like radiation of pain to other parts of the body, shortness of breath, palpitation, etc.

2.1.2. Questionnaire development and sample data collection

A questionnaire has been developed under the guidance of medical experts. The questionnaire collects the information about the patients’ clinical parameters based on the risk factors of CAD. It stores all information from the patient about smoking status, height and weight for evaluating BMI, systolic and diastolic blood pressure for determining hypertension, fasting blood sugar (FBS) for determining blood glucose level, TC, TGL, LDL and HDL value for determining lipid profile, chest pain and its associated symptoms like radiation of pain to back, jaw, arms or abdomens, shortness of breath, palpitation, cough, sweating or fainting, swelling in legs, peptic ulcer, etc. Questions were asked verbally to the patients and the responses were encoded by the knowledge engineers. Total of 500 patients’ data were collected from Advanced Medical Research Institute (AMRI) hospital, Salt Lake, Kolkata. The participating patients’ minimum age was restricted to 30 years and the subjects with prior history of myocardial infarction (MI), arrhythmia and any type of coronary heart diseases were excluded from our investigation. The patients were broadly grouped into four categories viz patients admitted to the intensive care unit, those recovering from bypass graft surgery, patients undergoing medication for CAD but no prior history of MI and finally patients coming to the hospital for routine checkup.

2.2. Structure of CAD screening expert system

The present expert system has three major components – knowledge base (KB), inference engine (IE) and user input (UI). While KB contains the expert’s knowledge encoded in computer perceptible form, the IE incorporates the ability of reasoning for deciding which rules to be selected and fired for risk stratification. Although the structure of KB and the knowledge representation depend on the nature of the problem to be solved, the fuzzy inference is independent of KB and same IE can be used for different problems. Being a fuzzy based approach the present CAD screening system contains fuzzy KB, fuzzy IE and fuzzified inputs.

2.2.1. Fuzzification process

The input parameters in our fuzzy expert system [34] have been fuzzified. The fuzzy logic is a multi-valued logic. In the traditional crisp set approach an element belongs to a set either fully or it does not belong to the set. The concept of fuzzy set theory in the notion of partial belongingness of an element to one or more than one sets. The partial belongingness is denoted by membership function where the membership value ranges from 0 to 1. In fuzzy set if the membership of an object is 1 then the object fully belongs to the set, if it is 0, then the object does not belong to the set, if it is in between 0 and 1, then object has partial belongingness to the set. In crisp approach the membership grade of an object is either 0 or 1. Fuzzy logic is a representation scheme for uncertain or vague notions. This is why the fuzzification of input parameters and using them in fuzzy module will cope with the impreciseness of the medical domain. The risk factors discussed in the previous section are input parameters for our present system. They are smoking, obesity, hypertension, diabetes, lipid profile and chest pain. For the fuzzification of input parameters in the present CAD screening system the clinical parameters have been considered as fuzzy variables. Each fuzzy variable has three fuzzy values. The linguistic representation of the input parameters is shown in Table 1. The values of these fuzzy variables are defined by fuzzy membership grade which are determined by the membership function. Among various membership functions we consider here trapezoidal membership function because of its generality and capacity to contain more fuzzy information. In fact triangular membership function is its special form where two vertex points are same. The equation of trapezoidal membership function is defined as follows

$$\mu(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{g}, & \text{if } a \leq x < b \\ 1, & \text{if } b \leq x < c \\ \frac{d-x}{d-c}, & \text{if } c \leq x \leq d \end{cases}$$

where \(\mu(x)\) describes the membership function of variable \(x\) which is defined over the range of \([a, d]\) with \(a < b \leq c < d\). The values of \(a, b, c, d\) for different linguistic values of a linguistic variable have been determined based on doctor’s experience.

The interval \([b, c]\) is called core of the trapezoidal function, whereas \([a, d]\) is called support of the trapezoidal function and \([a, b]\) and \([c, d]\) are called boundaries of trapezoidal functions. In Fig. 1 a plot of trapezoidal membership graph has been given. For example support \([a, d]\) and core \([b, c]\) intervals are given as \([5, 25]\) and \([10, 20]\) respectively.

Fuzzification of the chest pain has been done using doctor’s perception and intuition as follows:

- \(C = \{C_1, C_2, C_3, \ldots, C_n\}\) is the set of symptoms associated with chest pain and \(|C| = n\), \(C_1, C_2, C_3\) are three discrete fuzzy sets whose elements are \(P(C) - \{\phi\}\), where \(P(C)\) is the power set of \(C\) and \(\{\phi\}\) is the empty set. \(C_1, C_2, C_3\) are discrete fuzzy sets indicating low significance of chest pain, medium significance of chest pain, high significance of chest pain respectively and are defined as follows:

$$C_i = \left\{ \frac{l_1}{C_1}, \frac{l_2}{C_2}, \ldots, \frac{l_k}{C_k}, \ldots, \frac{l_n}{C_n} \right\}$$

$$C_m = \left\{ \frac{m_1}{C_1}, \frac{m_2}{C_2}, \ldots, \frac{m_k}{C_k}, \ldots, \frac{m_n}{C_n} \right\}$$

$$C_h = \left\{ \frac{h_1}{C_1}, \frac{h_2}{C_2}, \ldots, \frac{h_k}{C_k}, \ldots, \frac{h_n}{C_n} \right\}$$

where \(l_1, l_2, \ldots\) are membership values of corresponding elements in \(C_i, m_1, m_2, \ldots\) are membership values of corresponding elements in \(C_m, h_1, h_2, h_3, \ldots\) are membership values of corresponding elements
Table 1
Linguistic representation of the fuzzy variables.

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Linguistic grading of risk parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>Young, Medium, Old</td>
</tr>
<tr>
<td>Smoking (number of cigarettes per day)</td>
<td>Low smoker, Medium smoker, High smoker</td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>Slightly increased, Medium increased, Highly increased</td>
</tr>
<tr>
<td>Systolic blood pressure (Hg-mm)</td>
<td>Slightly increased, Medium increased, Highly increased</td>
</tr>
<tr>
<td>Diastolic blood pressure (Hg-mm)</td>
<td>Slightly increased, Medium increased, Highly increased</td>
</tr>
<tr>
<td>Fasting blood sugar (mg/dl)</td>
<td>Slightly increased, Medium increased, Highly increased</td>
</tr>
<tr>
<td>Total cholesterol (mg/dl)</td>
<td>Slightly increased, Medium increased, Highly increased</td>
</tr>
<tr>
<td>LDL</td>
<td>Slightly increased, Medium increased, Highly increased</td>
</tr>
<tr>
<td>Tri-glyceride (mg/dl)</td>
<td>Slightly increased, Medium increased, Highly increased</td>
</tr>
<tr>
<td>HDL (mg-dl)</td>
<td>Low, Medium, High</td>
</tr>
<tr>
<td>Chest pain significance</td>
<td>Low significance, Medium significance, High significance</td>
</tr>
</tbody>
</table>

In the present work, the fuzzification of the clinical variables has been done based on doctor's experience. The normal and abnormal ranges were extracted based on literatures as well as doctors’ opinions. The range and vertex values of trapezoidal function for linguistic variables of each risk factor have been obtained through the consensus of group of doctors involved in the experiment. Several rounds of interviews were done to find the support and core intervals of trapezoids for linguistic variables of each risk factor. The ranges of linguistic variables are given in Table 2. In Table 4 the left chest pain fuzzification details has been shown. Based on cardiologists’ experience we have considered five symptoms associated with left chest pain. They are arm, jaw, back pain (s1), radiation of pain (s2), fainting (s3), black-out (s4) and shortness of breath (s5). Different combinations of these symptoms have different level of significance in three fuzzy sets –C0, Cm and Cb. The membership grades of significance of different combinations of symptoms have been obtained by knowledge engineers by interviewing them.

2.2.2. Fuzzy knowledge base
Knowledge base of our system has been developed with the help of the domain expert’s past experience and medical intuition. After multiple rounds of interviews with the cardiologists separately, the knowledge engineers extracted preliminary rules. The preliminary rules from each domain expert were verified by the other experts. Any confliction was resolved by the consensus. There are many techniques for knowledge representations like predicate calculus, frame, object-oriented technique, production rules, etc. The present module uses fuzzy production rules to represent the knowledge. The production rules are written in the format of IF<condition> THEN<conclusion>. In the present fuzzy system condition and conclusion are fuzzy variables. In Table 3 some sample fuzzy rules have been shown. These rules are diagnostic and they are selected by the inference engine of our fuzzy expert system.

2.2.3. Fuzzy inferenceing by Mamdami approach
Inference engine module is the heart of expert system. This module gets the input from the user and selects the rules from the rule base. The fuzzy inference engine of the present system selects fuzzy rules and uses Mamdami inferencing [34] to produce the fuzzy risk output. Mamdami approach is described in short below. The Mamdami (max–min) inference mechanism used in the present work is as follows

Let fuzzy rule base has r production rules like
If \( x_1 = A_{1i} \) and \( x_2 = A_{2j} \) and \( \ldots \) \( x_n = A_{ni} \) then \( y = B^k \) where \( k = 1,2,3,\ldots,r \).

\( x_1, x_2, \ldots x_n \) are causal factors and y is decision \( A_{1i}^1, A_{2j}^2, \ldots A_{ni}^n \) are fuzzy sets representing the kth rule risk factor and \( B(k) \) is the kth rule fuzzy risk set. For each k, Mamdami rule is described as

\[
\mu_{k}(y) = \min\{\mu_{1}^{k}(x_1), \mu_{2}^{k}(x_2), \ldots , \mu_{n}^{k}(x_n)\}
\]

where \( \mu_{k}^{n}(y) \) is the membership value of the kth rule risk factor. \( \mu_{n}^{k}(x_n) \) is the membership value of nth risk factor of the fuzzy set \( A_{ni}^n \). Let \( n_{1} \) (\( n_{1} \leq r \)) number of rules have been matched and fired and the fired rule set is denoted by \( R \). So aggregating \( n_{1} \) rules we have the fuzzy risk output as \( - \mu_{k}(y) = \max\{\mu_{k}(y)\} \), where \( \mu_{k}(y) \) is the kth rule risk membership value and \( k \in R \) and \( i = 1,2,3,\ldots,n_{1} \). \( \mu(y) \) is the fuzzy risk output which has to be defuzzified into crisp value by the defuzzification method stated in Section 2.2.4.

2.2.4. Defuzzification
Centroid method [34] is considered to defuzzify the fuzzy risk output to crisp risk percentage. The crisp output has been determined as

\[
\mu_{y}(y) = \frac{\sum_{k \in R} \mu_{k}(y) k}{\sum_{k \in R} \mu_{k}(y)}
\]
where $\int$ is the algebraic integration.

2.2.5. Rule organisation

Rule organisation is one of the important parts in our expert system. As the knowledge base is dynamic in nature its volume...
may increase in future with the addition of new rules or deletion of invalid rules in the rule base. It may become difficult to manage all the rules. Moreover, in practice another problem is that the patients may not know the values of all of his risk factors. The developed screening system deals with variable length input vectors i.e. a subset of total risk factors. To cope with the problem of variable length input vectors and large volume of rule base, significant focus has been given on rule organisation in the present system using the concepts of modules, meta-rules, rule indexing and rule checking technique.

### 2.2.5.1. Modules

To manage the large number of rules in the present expert system rules are grouped into different blocks. The

### Table 2

Range of the linguistic variables of risk factors.

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Young/low/slightly increased</th>
<th>Medium/medium increased</th>
<th>Old/high/highly increased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>Support interval Core interval</td>
<td>Support interval Core interval</td>
<td>Support interval Core interval</td>
</tr>
<tr>
<td>30–45</td>
<td>35–40</td>
<td>40–65</td>
<td>50–60</td>
</tr>
<tr>
<td>0–6</td>
<td>2–4</td>
<td>4–12</td>
<td>7–9</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>18.5–25</td>
<td>20–22</td>
<td>23–30</td>
</tr>
<tr>
<td>SBP (Hg-mm)</td>
<td>118–130</td>
<td>122–126</td>
<td>128–140</td>
</tr>
<tr>
<td>DBP (Hg-mm)</td>
<td>70–84</td>
<td>75–80</td>
<td>82–100</td>
</tr>
<tr>
<td>FBS (mg/dl)</td>
<td>98–120</td>
<td>108–112</td>
<td>118–150</td>
</tr>
<tr>
<td>TC (mg/dl)</td>
<td>180–210</td>
<td>184–194</td>
<td>196–240</td>
</tr>
<tr>
<td>LDL (mg/dl)</td>
<td>100–130</td>
<td>110–126</td>
<td>128–160</td>
</tr>
<tr>
<td>TGL (mg/dl)</td>
<td>130–156</td>
<td>135–145</td>
<td>150–200</td>
</tr>
<tr>
<td>HDL (mg/dl)</td>
<td>&lt;=45</td>
<td>&lt;=30</td>
<td>40–65</td>
</tr>
</tbody>
</table>

### Table 3

Sample fuzzy rules.

<table>
<thead>
<tr>
<th>Rule no.</th>
<th>Diagnostic fuzzy rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>If age is young and smoking is low then risk = low</td>
</tr>
<tr>
<td>R2</td>
<td>If age is medium and smoking is high and BMI is medium then risk = medium</td>
</tr>
<tr>
<td>R3</td>
<td>If age is old and smoking is high and BMI is high and SBP is high and DBP is high and FBS is medium increased and TC is slightly increased and LDL is low and TGL is high then risk = low</td>
</tr>
<tr>
<td>R4</td>
<td>If age is old and smoking is high and BMI is high and SBP is high and DBP is high and FBS is medium increased and TC is medium increased and LDL is slightly increased and TGL is slightly increased and HDL is high then risk = low</td>
</tr>
<tr>
<td>R5</td>
<td>If age is old and smoking is high and BMI is high and SBP is high and DBP is high and FBS is highly increased and TC is highly increased and LDL is medium increased and TGL is highly increased and HDL is high then risk = low</td>
</tr>
<tr>
<td>R6</td>
<td>If age is medium and smoking is low and BMI is high and SBP is high and DBP is high and FBS is highly increased and TC is highly increased and LDL is medium increased and TGL is highly increased and HDL is medium and chest pain significance is medium then risk = high</td>
</tr>
<tr>
<td>R7</td>
<td>If age is low and BMI is medium increased and SBP is slightly increased and DBP is medium increased and FBS is slightly increased and TC is medium increased and LDL is slightly increased and TGL is slightly increased and HDL is high and chest pain significance is low then risk = medium</td>
</tr>
<tr>
<td>R8</td>
<td>If age is old and BMI is medium increased and SBP is highly increased and DBP is highly increased and FBS is highly increased and TC is highly increased and LDL is slightly increased and TGL is slightly increased and HDL is medium and chest pain significance is low then risk = medium</td>
</tr>
<tr>
<td>R9</td>
<td>If age is old and BMI is medium increased and SBP is medium increased and DBP is medium increased and FBS is slightly increased and TC is slightly increased and LDL is slightly increased and TGL is slightly increased and HDL is low and chest pain significance is low then risk = low</td>
</tr>
<tr>
<td>R10</td>
<td>If age is young and BMI is slightly increased and SBP is slightly increased and DBP is slightly increased and FBS is slightly increased and LDL is low and chest pain significance is low then risk = low</td>
</tr>
</tbody>
</table>
grouping of rules has been done based on some criteria. Each block of rules is called module. Selection of a particular module is done by meta-rules that have been written in the meta-rule base. The design of modules depends on the nature of the problem for which the expert system is developed. In the present screening system the grouping criteria of the rules is governed by the number of disease markers used as a set of antecedents in the production rules. The present system has 10 modules. The particular module that has to be searched during diagnosis is based on the number of risk factors taken as input vector from the patients. Module one has the rules containing only one risk factor, i.e. any one of the ten risk factors. Module two contains the rules that involve two risk factors. Such a way module ten has the rules containing all the risk factors.

2.2.5.2. Meta-rules. Meta-rules are the special types of rules in our system. These govern the reasoning process of the expert system by determining which module will be selected based on the facts given by the patient. From the input vector given by the patient, the number of given risk factors is determined and the meta-rules select the particular module to be visited. The structure of meta-rules is like production rules. It also has IF–THEN structure. The structure of meta-rules used in the CAD screening system is given as follows

Let,

\[ S_p = \{ s_i \} \] denotes the set of risk factors for the patient \( P \), \( \forall i = 1, 2, \ldots, n \) where \( |S_p| = \text{cardinality of the set } S_p \).

Meta-Rule 1: IF \( |S_p| = 1 \) THEN < Visit Module 1 >
Meta-Rule 2: IF \( |S_p| = 2 \) THEN < Visit Module 2 >
Meta-Rule 3: IF \( |S_p| = 3 \) THEN < Visit Module 3 >
   .
   .
   .
Meta-Rule \( n \): IF \( |S_p| = n \) THEN < Visit Module \( n \) >

The meta-rules in the present system selects the appropriate block of rules even before the actual fuzzy reasoning starts in the inference engine. The main difference between the meta-rules and rules in the knowledge base is that meta-rules do not participate in the fuzzy reasoning mechanism whereas the rules written in the knowledge base are diagnostic rules. Meta-rules facilitate

### Table 5
Searching priority table.

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Searching priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>( p_1 )</td>
</tr>
<tr>
<td>Smoking</td>
<td>( p_2 )</td>
</tr>
<tr>
<td>BMI</td>
<td>( p_3 )</td>
</tr>
<tr>
<td>SBP</td>
<td>( p_4 )</td>
</tr>
<tr>
<td>DBP</td>
<td>( p_5 )</td>
</tr>
<tr>
<td>FBS</td>
<td>( p_6 )</td>
</tr>
<tr>
<td>TC</td>
<td>( p_7 )</td>
</tr>
<tr>
<td>LDL</td>
<td>( p_8 )</td>
</tr>
<tr>
<td>TGL</td>
<td>( p_9 )</td>
</tr>
<tr>
<td>HDL</td>
<td>( p_{10} )</td>
</tr>
<tr>
<td>CP</td>
<td>( p_{11} )</td>
</tr>
</tbody>
</table>

### Table 6
Rule address table.

<table>
<thead>
<tr>
<th>Module number</th>
<th>Root node</th>
<th>Address of root node</th>
<th>Module number</th>
<th>Root node</th>
<th>Address of root node</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Age</td>
<td>A1</td>
<td>M5</td>
<td>Age</td>
<td>A5</td>
</tr>
<tr>
<td></td>
<td>Smoking</td>
<td>B1</td>
<td></td>
<td>Smoking</td>
<td>B5</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>C1</td>
<td></td>
<td>BMI</td>
<td>C5</td>
</tr>
<tr>
<td></td>
<td>SBP</td>
<td>D1</td>
<td></td>
<td>SBP</td>
<td>D5</td>
</tr>
<tr>
<td></td>
<td>DBP</td>
<td>E1</td>
<td></td>
<td>DBP</td>
<td>E5</td>
</tr>
<tr>
<td></td>
<td>FBS</td>
<td>F1</td>
<td></td>
<td>FBS</td>
<td>F5</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>G1</td>
<td></td>
<td>TC</td>
<td>G5</td>
</tr>
<tr>
<td></td>
<td>LDL</td>
<td>H1</td>
<td>M6</td>
<td>Age</td>
<td>A6</td>
</tr>
<tr>
<td></td>
<td>TGL</td>
<td>I1</td>
<td></td>
<td>Smoking</td>
<td>B6</td>
</tr>
<tr>
<td></td>
<td>HDL</td>
<td>J1</td>
<td></td>
<td>BMI</td>
<td>C6</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>H1</td>
<td></td>
<td>SBP</td>
<td>D6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FBS</td>
<td>E6</td>
</tr>
<tr>
<td>M2</td>
<td>Age</td>
<td>A2</td>
<td>M7</td>
<td>Age</td>
<td>A7</td>
</tr>
<tr>
<td></td>
<td>Smoking</td>
<td>B2</td>
<td></td>
<td>Smoking</td>
<td>B7</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>C2</td>
<td></td>
<td>BMI</td>
<td>C7</td>
</tr>
<tr>
<td></td>
<td>SBP</td>
<td>D2</td>
<td></td>
<td>SBP</td>
<td>D7</td>
</tr>
<tr>
<td></td>
<td>DBP</td>
<td>E2</td>
<td></td>
<td>DBP</td>
<td>E7</td>
</tr>
<tr>
<td></td>
<td>FBS</td>
<td>F2</td>
<td></td>
<td>FBS</td>
<td>E8</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>G2</td>
<td>M8</td>
<td>Age</td>
<td>A8</td>
</tr>
<tr>
<td></td>
<td>LDL</td>
<td>H2</td>
<td></td>
<td>Smoking</td>
<td>B8</td>
</tr>
<tr>
<td></td>
<td>TGL</td>
<td>I2</td>
<td></td>
<td>BMI</td>
<td>C8</td>
</tr>
<tr>
<td></td>
<td>HDL</td>
<td>J2</td>
<td></td>
<td>SBP</td>
<td>D8</td>
</tr>
<tr>
<td>M3</td>
<td>Age</td>
<td>A3</td>
<td>M9</td>
<td>Age</td>
<td>A9</td>
</tr>
<tr>
<td></td>
<td>Smoking</td>
<td>B3</td>
<td></td>
<td>Smoking</td>
<td>B9</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>C3</td>
<td></td>
<td>BMI</td>
<td>C9</td>
</tr>
<tr>
<td></td>
<td>SBP</td>
<td>D3</td>
<td>M10</td>
<td>Age</td>
<td>A10</td>
</tr>
<tr>
<td></td>
<td>DBP</td>
<td>E3</td>
<td></td>
<td>Smoking</td>
<td>B10</td>
</tr>
<tr>
<td></td>
<td>FBS</td>
<td>F3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>G3</td>
<td>M11</td>
<td>Age</td>
<td>A11</td>
</tr>
<tr>
<td></td>
<td>LDL</td>
<td>H3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TGL</td>
<td>I3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>Age</td>
<td>A4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smoking</td>
<td>B4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>C4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SBP</td>
<td>D4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DBP</td>
<td>E4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FBS</td>
<td>F4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>G4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LDL</td>
<td>H4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the rule searching by finding a chunk of rules out of which some rules will be fired based on the fuzzy inference mechanism. The number of meta-rules changes only in case of addition or deletion of new modules in the present system.

2.2.5.3. Rule address storage in tree representation. In the expert system when the number of rules increases significantly, they become unmanageable and tedious for the knowledge engineers as well as for the domain experts. Moreover for a particular input vector, the search of the whole rule base is not needed always. The linguistic rules are based on different combinations of risk factors. In the present system at first every possible combination of risk factors has been generated and based on each combination linguistic rules have been formed. In the present system, number of combinations of risk factors is determined as \(11^3 + 11^2 + 11^3 + 11^4 + 11^5 + 11^6 + 11^7 + 11^8 + 11^9 + 11^{10} + 11^{11}\). Three linguistic variables have been used for each risk factor (see Table 1). For each combination of risk factors maximum number of linguistic rules are determined as \((3^2)^n\), where \(r\) = number of risk factors for a particular combination. For facilitating the rule searching mechanism, a rule address storage component has been developed in the present screening system. The combinations of risk factors have been represented by tree data structure. For module \(k\) the number of trees is determined as \((n - k + 1)\), where \(n\) = total number of risk factors. In the present study \(n = 11\). For a module \(k\) the level of tree is \((k + 1)\) assuming the level of the root node is one \((1)\). The leaf nodes at the level \((k + 1)\) contain the address of block of linguistic rules corresponding to the combination of risk factors represented by the path starting from the root node to the leaf node. In the present rule address storage component a master table has been designed by the knowledge engineers. The table stores the address of root nodes of each tree module wise. The rule address storage system has been shown in Table 6. Table 6 shows the module wise root node address. M1, M2, M3, M4, M5, M6, M7, M8, M9, M10, M11 are the module numbers. The table stores the address of all the root nodes per module. The address of the root nodes are represented as A1, B1, C1, D1, E1, F1, G1, H1, I1, J1. A new linguistic rule is added or deleted, the address in the leaf nodes change accordingly. The tree construction algorithm is given below:

Algorithm for construction of tree:

1. Assign the priority to each risk factor. The priority is an integer number given by the knowledge engineers for facilitating the rule searching mechanism. The priority values are shown in Table 5. Without loss of generality it is assumed that \(p_1 > p_2 > p_3 > p_4 > p_5 > p_6 > p_7 > p_8 > p_9 > p_{10} > p_{11}\), where

2. Sort the risk factors in descending order based on the priority values of the risk factors and store their indexes from the sorted list

3. Find the number of trees for module \(k\) as follows

4. Root nodes for the module \(k\) are determined by their indexes \(i\) in the sorted list of risk factors where \(i \leq k\)

5. For each root node construct a tree of level \((k + 1)\), assuming the level of the root node is one \((1)\). Each tree in a module \(k\) denotes all possible combinations of risk factors starting from the root node. For a module \(k\), for each root node tree is constructed as follows

\[ S = \text{sorted list of } n(n + 1) \text{ in the present work} \]

(b) empty list

\[ P_list = \text{empty list of parent node indexes} \]

\[ b = \text{empty list} \]

\(i\) in — index of the root node from the sorted list \(S\); \(l = 1\); initialize \(k\) (module number for which the algorithm is being implemented)

\(i\) Create a node \(Nd\).

\[ Nd = S(in); \]

\(ii\) Add the index \(in\) in the list \(p_list\)

\(iii\) while \(l < k\)

\(iv\) \(l = l + 1\); \(R = k - l\); top = 0;

\(v\) If the list \(b\) is not empty, delete the elements of the list \(b\).

\(vi\) For \(i = 1, 2, 3, \ldots, |p_list|\), where \(|p_list|\) denotes the cardinality of the list \(p_list\)

\(vii\) \(m = p_list(i)\)

\(viii\) For \(j = m + 1, m + 2, m + 3, \ldots, |S|\)

\(ix\) End // \(j\) loop

\(x\) End // \(i\) loop

\(xi\) Delete the elements of \(p_list\).

\(xii\) End//while loop

\(xiii\) For each node at level \(k\) create one child node. The nodes at level \((k + 1)\) are the leaf nodes.

Assign the address of the block of linguistic rules (fuzzy production rules) dealing with the risk factors of the corresponding path from root node to the leaf node. The address are represented as starting and ending row numbers of the block of linguistic rules corresponding to the path from the root node to the leaf node.

An illustration of tree structure has been shown in Appendix.

2.2.5.4. Consistency checking of rules. In the present screening system two types of consistency checking have been applied- (i) Contradictory rule checking (ii) Redundant rule checking. Contradictory rules generate discrepancies in the rule inference mechanism and the rule redundancy produces unnecessary and duplicate rules in the rule base. Different types of contradictory and redundant rules are defined in our system. These are described as follows:

(i) Contradictory rules

This type of consistency checking is further categorised into two types of contradictions checking – (a) intra-module contradictory rule checking (b) inter-module contradictory rule checking. Intra-module and inter-module contradictory rule checking are explained as follows:

(i) Contradictory rules

This type of consistency checking is further categorised into two types of contradictions checking – (a) intra-module contradictory rule checking (b) inter-module contradictory rule checking. Intra-module and inter-module contradictory rule checking are explained as follows:

Let, \(R = \{R_i\}\) is the set of rules in the rule base of a module, where \(i = 1, 2, 3, \ldots, n\).

(a) Intra-module contradictory rules

This consistency checking is held among the rules within a module. Two rules \(R_i\) and \(R_j\) of a module are contradictory if \(ANT(R_i) = \overline{ANT(R_j)}\) and \(CON(R_i) = \overline{CON(R_j)}\), \(\forall i \neq j\), where \(ANT(R_i)\) and
CON(Rt) denote antecedent and conclusion of the rule Rt respectively.

(b) Inter-module contradictory rules
This type of consistency checking is done between the modules.
Let \( S_k = \{ s_1, s_2, \ldots, s_p \} \) be the set of risk factors considered as antecedents and \( l_j \) the consequent of rule \( R_k \) of module \( M_m \). Let \( S_R = \{ s_1, s_2, s_3, \ldots, s_q \} \) be the set of risk factors considered as antecedents and \( l_j \) is the consequent of rule \( R_l \) of module \( M_n \), where \( l_j \in \{ \text{low, medium, high} \} \), \( l_j \in \{ \text{low, medium, high} \} \), \( M_m \neq M_n \).

Based on the severity of the disease we have low < medium < high. Two rules \( R_k \) and \( R_l \) are said to be inter-module contradictory if \( S_R \cap S_l \neq \emptyset \) and \( l_j > l_j \).

(ii) Redundant rules
Redundancy checking in the rule base is a special type of consistency checking in the present system. Our system has been given the flexibility to deal with the variable length input vector. Consequent of the each production rule is the output variable “risk” that has been described by linguistic variable low, medium and high. If \( S_R \subseteq S_l \) and \( l_j = \max(\text{low, medium, high}) \) and \( l_j \neq \max(\text{low, medium, high}) \) then \( R_l \) is called redundant rule for \( R_k \).

Consistency checking for rules is applied when a new rule has to be entered in the rule base. If the new rule is intra-module or inter-module contradictory in the present rule base the expert is asked for his opinion. Based on the advice of the experts either new rule is denied to be entered or the old rule is updated with the consequence of the new rule. If the new rule is redundant it is not entered in the rule base by the system.

3. Performance evaluation
In the present work three classes have been considered—low, medium and high risk patients. The low risk patients are considered as normal patients and both medium and high risk patients are considered as abnormal (CAD) patients while evaluating the performance of the fuzzy expert system model. For the performance evaluation of the present system we have used the measures of specificity, sensitivity and accuracy [29]. They are defined as follows:

(i) True positive (TP): It denotes the number of abnormal patients correctly classified by the model.
(ii) True negative (TN): It denotes the number of normal patients correctly classified by the model.
(iii) False positive (FP): It denotes the number of healthy patients wrongly classified as abnormal patients by the model.
(iv) False negative (FN): It denotes the number of abnormal patients wrongly classified as normal patients by the model.
(v) Specificity: It is defined as percentage of normal patients classified correctly by the model. It is determined as

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

(vi) Sensitivity: It is defined as the percentage of abnormal patients classified correctly by the model. It is determined as

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

(vii) Accuracy: It denotes the percentage of correctly classified patients. In the present work (3-class problem) it is determined as

\[
\text{Accuracy} = \left( \frac{l + m + h}{N} \right) \times 100
\]

where \( l \) is the number of correctly classified patients as low risk, \( m \) the number of correctly classified patients as medium risk, \( h \) the number of correctly classified patients as high risk, and \( N \) is the total number of patients examined.

4. Risk computation-example
For computational explanation of the system an example has been described on the basis of sample patient input parameters. After filling up the questionnaire form the following information has been collected.

Blood pressure (BP) = 148/80 mm Hg, height = 169 cm, weight = 84 kg, fasting blood sugar (FBS) = 147 mg/dl.

Using the methodology, discussed above, the risk computation for the patient P has been shown stepwise. For the present work risk stratification are taken like – Low risk = 0–30%, medium risk = 30–55%, high risk = 55–100%.

Step 1. It includes the fuzzification of the risk parameters. The fuzzified values of the risk factors of the patient P are shown below

\[
\begin{align*}
\mu_{\text{Slightly increased}}(\text{SBP}) &= 0 \\
\mu_{\text{Medium increased}}(\text{DBP}) &= 0 \\
\mu_{\text{High}}(\text{FBS}) &= 1 \\
\mu_{\text{High}}(\text{BMI}) &= 0 \\
\mu_{\text{Medium}}(\text{BMI}) &= 0.20 \\
\mu_{\text{High}}(\text{BMI}) &= 0.26 \\
\mu_{\text{Slightly increased}}(\text{FBS}) &= 0, \mu_{\text{Medium increased}}(\text{FBS}) &= 0.37, \mu_{\text{High}}(\text{FBS}) &= 0.18
\end{align*}
\]

Step 2. Fuzzy inference module selects the following rules from the Rule base.

R1: If SBP = High And DBP = slightly increased And BMI = Medium And FBS = Medium increased Then Risk = Medium.
R2: If SBP = High And DBP = slightly increased And BMI = Medium And FBS = High Then Risk = High.
R3: If SBP = High And DBP = slightly increased And BMI = High And FBS = High Then Risk = High.
R4: If SBP = High And DBP = slightly increased And BMI = High And FBS = High Then Risk = High.

After firing these rules Mamdami Inference is used to produce fuzzy output.

Step 3. Fuzzy output is defuzzified using centroid method and for the present example risk percentage is calculated as

\[
\text{Risk} = \frac{\int \mu_R(y) \cdot y \, dy}{\int \mu_R(y) \, dy} = 89.55\%
\]

Step 4. From the risk stratification the risk of 89.55 % is considered as High risk. So the patient P has high risk of coronary artery disease.

5. Results and discussion
The CAD screening expert system was applied to total of 500 subjects (low risk = 22.80%, medium risk = 33.60% and high risk = 43.60%) who came to AMRI hospital for check up. Information based on clinical risk factors was taken from the subjects and were labelled by a group of cardiologists. Simultaneously our screening system was applied to find the model predicted risk for those patients. For evaluation of the performance of the system, model predicted outputs were compared with the results given by the cardiologists. Confusion matrix is shown in Table 7. Overall accuracy, specificity and sensitivity of the system were found to be 84.20%, 83.33% and 95.85% respectively. From the performance evaluation it is obvious that our system has the high capability of detecting unhealthy (with respect to CAD) group of people and can be applied for mass screening for CAD detection. The Table 6 shows that the specificity is low relative to the sensitivity. It
indicates that more number of healthy people is detected as unhealthy. But we have used only very simple information (age, smoking, height, weight, chest pain associated with other symptoms) and very low cost laboratory tests for finding SBP, DBP, FBS, TC, LDL and HDL. The persons, categorised as CAD patients with their respective degree of severity by our system, will be prescribed for higher level tests like ECG, Angiogram, etc., and if they are wrongly classified as unhealthy, they will be properly detected as healthy after doing those higher medical tests. But higher sensitivity indicates that very less number of CAD patients is wrongly identified as normal patients. Higher sensitivity is required because if the sensitivity is low then a large number of abnormal subjects will be left out as normal and those apparently normal patients will grow CAD silently leading to fatal consequence as they will not be prescribed for further medical tests. Consequently the reliability of the system will decrease.

To compare the proposed methodology we have applied artificial neural network (ANN) and data mining algorithms viz-Iterative Dichotomiser 3 and classification and regression tree (CART). Table 8 shows the comparative results of ANN, ID3, CART and proposed fuzzy expert system. Our methodology has few advantages over the other classification technologies mentioned above. Although the overall accuracy of the ANN is 82.03% which is close to overall accuracy (84.20%) of our proposed methodology, but ANN acts as a black box classifier. Due to sophisticated mechanism of ANN, the reason of rule inference is not vivid. Decision tree extracts the rules [43] from the annotated dataset and rules are easily comprehensible for the cardiologists, but the accuracies are found to be less than our work. The proposed methodology is a rule based approach. Here the rules are represented in easily understandable format and the overall accuracies are higher than that of ANN, ID3 and CART. Moreover the rules have been developed by directly interviewing the cardiologists. So the data driven nature of the learning algorithms does not exist.

The fuzzification method used in this system has been done based on doctor’s perception and experience. The vertex values of the trapezoidal membership functions have been obtained from medical experts after interviewing them by the knowledge engineers. In Fig. 2 the graphs of membership functions of risk factors i.e. age, smoking, BMI, FBS, SBP, DBP, TC, LDL, TGL and HDL, have been shown. The ranges of values of risk factors are also plotted in the graph. The three linguistic variables of each risk factor correspond to three trapezoidal membership graphs. The vertex values of the functions are obtained from the group of doctors by interviewing the cardiologists. So the data driven nature of the learning algorithms does not exist.

Table 8
Comparative results with alternative approaches.

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Specificity (%)</th>
<th>Sensitivity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>81.08</td>
<td>92.30</td>
<td>82.03</td>
</tr>
<tr>
<td>ID3</td>
<td>67.33</td>
<td>72.85</td>
<td>69.66</td>
</tr>
<tr>
<td>CART</td>
<td>70.57</td>
<td>75.86</td>
<td>72.66</td>
</tr>
<tr>
<td>Proposed fuzzy system</td>
<td>83.33</td>
<td>95.85</td>
<td>84.20</td>
</tr>
</tbody>
</table>

In case of BMI (Fig. 2c) a number of doctors suggested that persons having BMI in between 18.5 kg/m² and 27 kg/m² can be considered as slightly obese (slightly increased BMI); whereas the other doctors gave opinion that the persons having BMI in between 23 kg/m² and 30 kg/m² should be considered as medium increased BMI. As experts’ opinions are divided in the interval range of [23–25], this ambiguity has been dealt with the fuzzy membership function. In this range both group of doctors’ opinions have been incorporated using trapezoidal membership function. Form the graph it is obvious that there are two membership values of BMI in the range of [23–25] because the BMI values in this range partially belong to both fuzzy sets-slightly increased and medium increased. Other cases of ambiguities for other risk factors in different ranges have been dealt similarly. The overlapping regions of graphs in Fig. 2 reflect the ambiguity in doctors’ opinions.

A special focus has been given on proper rule organisation technique in present expert system. All the clinical risk factors may not be abnormal for CAD patients. From doctors’ experience it is observed that sometimes if a subset of clinical risk factors exceeds their normal range but other remain normal, then also a person may be CAD patient. Keeping the view in mind, rules with different possible combinations have been formulated and clinically validated by a group of cardiologists. Another reason of making different combinations is that a patient may not be aware of his all clinical parameters’ values. We have considered ten clinical risk factors for our investigation. But at any instant a patient may not know about all his clinical parameters’ values. He may know some of these values. Let a patient P has the set of risk factors, $S_p = \{s_1, s_2, s_3\}$. He is only aware of these clinical risk factors’ values at any instant. The risk factors in $S_p$ are arranged in descending order based on their searching priority values. Let $s_1 = \text{age}$, $s_2 = \text{smoking}$ and $s_3 = \text{bmi}$. The sorted list of risk factors is age, smoking and bmi. Our system will consider these risk factors as input and here $|S_p| = 3$. After getting this set of facts from the patient, the reasoning process will be governed by the meta-rules. Meta-rule base is searched and the meta-rule 3 is fired (see Section 2.2.5). According to conclusion of meta-rule 3 the control of the program visits Module 3 that is uniquely identified by $M_3$. Based on the searching priority of the risk factors (Table 5) the address of the root node is retrieved for module 3. The tree is traversed to find a match with the path containing the risk factors age- > smoking- > bmi. At level $(3 + 1) = 4$ from the leaf node of the corresponding path the address of block of fuzzy linguistic rules are determined. This block of rules contains the diagnostic rules dealing with the risk factors of age, smoking and bmi. The inference engine of the system determines
which rules to be matched from the block and fired for finding the CAD risk. The tree storage addressing system reduces the number of comparisons for selecting the production rules during rule matching. For the above case instead of searching all the production rules of module 3, only the linguistic rules containing the risk factors – age, smoking and bmi are searched and no other rules containing more than or less than three risk factors are compared.

In this work we have proposed a cost-effective screening methodology for CAD patients by developing a fuzzy expert system in which degree of severity of disease is also classified. In literature various machine learning methodologies have been applied vigorously for investigation of coronary heart events. Ordonez [31] has applied constrained association rules for detecting healthy and sick arteries based on both clinical data and nine heart region data like Antero Lateral, Antero – septal, etc. Based on the both clinical and ECG parameters like ST elevation, ST depression waves, LV failure Tsien et al. [30] has shown a comparative study between logistic regression and classification and regression tree. A hybrid approach was developed by Pollat et al. [44] based on both clinical and ECG parameters. They applied decision tree for feature selection and fuzzy weighted pre-processing and artificial immune recognition system were done for classification into normal and CAD patients. Optimised fuzzy modelling methodology was developed by Markos et al. [36] based on clinical features for classification of CAD and non-CAD patients.

The above mentioned literatures do not focus on finding the degree of severity of CAD. They only categorise the patients into two groups – CAD and normal. Moreover they consider not only clinical features but also other higher level medical information like ECG, heart region data, etc. The proposed methodology uses easily available clinical parameters like age, smoking, height, weight, SBP, DBP, chest pain details and some low cost laboratory tests- TC, LDL, TGL, and HDL. No higher level medical tests like ECG, angiogram, etc. are needed for predicting severity level of the patients by this methodology at least in its early stage. Moreover the proposed methodology detects the degree of severity of subjects in terms of low (normal), medium and high risk stratification. The system is easy to handle and can be applied as an adjunct tool to the cardiologists for automated mass screening of apparently healthy CAD patients in the developing countries where population is a major concern. Our methodology is based on rule based approach and the rules have been developed by the knowledge engineers while interviewing a group of cardiologists. As the rules are not developed by applying learning algorithms on labelled dataset, there is no data driven nature in the rule base. These rules have been expressed in the form of production rules (IF–THEN) and linguistic variables (Table 1) have been included for easy comprehensibility. The application of fuzzy logic approach into the proposed methodology has made the expert system capable of dealing with the ambiguity present in decision making in the medical domain.

The present screening system is a single disease expert system. But the rule organisation has been developed in such a way that in future it can be extended to multiple-disease expert system also. The concept of modules can be extended to sub modules based on diagnostic rules for different number of diseases.

6. Conclusion

In the present paper the structure of CAD screening system has been described. The system has been developed for the early detection of CAD. Although it depends on the expert's domain
knowledge but it can be used as a supportive tool for the doctors as their domain knowledge is encoded into the system in computer perceivable form. Moreover it uses only easily available clinical parameters and data obtained from very low cost laboratory tests. The risk classification has been done even before obtaining the next higher level information like ECG, angiogram, etc. The methodology is designed with fuzzy linguistic labels, fuzzy rule base and a well defined rule organisation module. The rule organisation technique has been developed to facilitate the rule searching mechanism. For better rule organisation we have focused on the concepts of modules, meta-rules and consistency checking in the rule base. We have defined two types of consistency checking in our rule base. It is anticipated that the present screening system can identify the risk group of patients at an early stage and consequently they will be guided for proper medication. This will help in better patient management as well as reduction in mortality rate due to CAD, as the patients including apparently healthy persons but silent carrier of CAD, will be aware of their respective risk labels well in advance.

Acknowledgements

We are extremely thankful to VECC, Kolkata for their financial support and BioStatistics & Medical Informatics (BMI) Lab of School of Medical Science and Technology, IIT Kharagpur for providing computing facility.

Appendix A. Illustration of tree structure

Let us consider that we have to construct a tree for the module 2 (M2). In the present work, \( n = 11 \) and \( k = 2 \). Number of trees for module 2 (M2) is determined as \[ \text{tree no} = n + k - 1 = 11 + 2 - 1 = 10 \]. From the Table 5 and the step 1 of the algorithm for construction of the tree the sorted list of risk factors based on the priority values is given as \( S = \{ \text{age, smoking, bmi, sbp, dbp, fbs, tc, ldl, tgl, hdl, cp} \} \). In module 2 (M2) the root nodes are

- Age, smoking, bmi, sbp, dbp, fbs, tc, ldl, tgl, hdl, cp.

A tree can be constructed for each node. For example the algorithm is implemented on the root node ‘age’ i.e. the risk factor with index \( = 1 \).

The output of the algorithm on root node ‘age’ will construct a tree as follows

At \( l = 1, n = 1 \). A node is create to store the risk factor ‘age’.

At \( l = 2 \), the indexes of the risk factors are selected. The risk factors are smoking, bmi, sbp, dbp, fbs, tc, ldl, tgl, hdl, cp.

For these risk factors the nodes are created and each node is made the child node of the risk factor ‘age’. At \( l = 3 \) for each node one node created and address of the corresponding path from the root node to the leaf node is assigned.

At level three (3), the nodes are the leaf nodes. Each leaf node contains the block of linguistic rules. The address represents the starting row number and the ending row number of the block having the rules containing the risk factors of that particular path. For example, Address 1 contains the address of block of linguistic rules dealing with the risk factors of the corresponding path and they are age and smoking. Address 1 contains the starting row number and ending row number of the block of linguistic rules dealing with only age and smoking.

From other leaf nodes at level \( (2 + 1) = 3 \), address of other rules are found. Similarly trees with other root nodes of module 2 can be constructed and address of the linguistic rules dealing with different combinations of risk factors can be stored at the leaf nodes at level \( (2 + 1) = 3 \). The same concept applies to other modules also. When a set of clinical parameters are considered for input, the system sort them in descending order based on searching priority (Table 5). From Table 6 the address of the root node containing the risk factor of index one (1) in the sorted input list is found and the tree is traversed to find a match and consequently the address of block of linguistic rules is determined.

References

[1] K. Park, Park’s textbook of Preventive and Social Medicine, in: 18th Medicine, Banarasidas Banhot.


