Clinical Decision Support System: Risk Level Prediction Of Heart Disease Using Decision Tree Fuzzy Rules

Dr. Anooj P.K.

ABSTRACT - As people have interests in their health recently, development of medical domain application has been one of the most active research areas. One example of medical domain application is detection system for heart disease based on computer-aided diagnosis methods, where the data are obtained from some other sources and are evaluated based on computer based applications. At the earlier time, the use of computer is to build knowledge based clinical decision support system which uses knowledge from medical experts and transfers this knowledge into computer algorithms manually. This process is time consuming and really depends on medical expert's opinion which may be subjective. To handle this problem, machine learning techniques have been developed to gain knowledge automatically from examples or raw data. Here, a weighted fuzzy rule-based clinical decision support system (CDSS) is presented for the diagnosis of heart disease, automatically obtaining the knowledge from the patient’s clinical data. The proposed clinical decision support system for risk prediction of heart patients consists of two phases, (1) automated approach for generation of weighted fuzzy rules, and (2) developing a fuzzy rule-based decision support system. In the first phase, we have used the mining technique, attribute selection and attribute weightage method to obtain the weighted fuzzy rules. Then, the fuzzy system is constructed in accordance with the weighted fuzzy rules and chosen attributes. Finally, the experimentation is carried out on the proposed system using the datasets obtained from the UCI repository and the performance of the system is compared with the neural network-based system utilizing accuracy, sensitivity and specificity.

Keywords: Clinical Decision Support System (CDSS), Heart disease, Fuzzy logic, Decision tree fuzzy rules, attribute selection, Risk prediction, UCI repository, Accuracy, Sensitivity and Specificity

I. INTRODUCTION

To extract hidden patterns and relationships from large databases, Data mining merges statistical analysis, machine learning and database technology [4]. In several areas of medical services, including prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data, Data Mining techniques have been applied [5]. Modern-day medical diagnosis is a very composite process, entailing precise patient data, a philosophical understanding of the medical literature and many years of clinical experience. The healthcare data which, unfortunately, are not “mined” to discover hidden information for effective decision making are collected in a huge amount by the health care industry [2]. Clinical decisions are often taken on the basis of doctors’ perception and experience rather than on the knowledge rich data masked in the database [3]. However unfortunately every doctor is not evenly expert in every sub specialty and they are in several places a scarce resource.

Medical care and Reduce costs can be enhanced by a system for automated medical diagnosis [3]. The information afforded by the patients may entail redundant and interrelated symptoms and signs in medical diagnosis especially when the patients suffer from more than one type of disease of same category. The physicians may not be capable to diagnose it accurately [21]. Unfortunately, due to complex interdependence on a variety of factors, accurate diagnosis of disease at a premature stage is quite a challenging task [22]. To enhance health and health care, Clinical Decision Support (CDS) affords clinicians, staff, patients, or other individuals with knowledge and person specific information, intelligently filtered or presented at appropriate times. In enhancing outcomes at a few health care institutions and practice sites, CDS has been effective by making needed medical knowledge readily obtainable to knowledge users [1]. Addressing clinical requirements, such as ensuring accurate diagnoses, screening in a timely manner for preventable diseases, or averting adverse drug events, are the most general exploitations of CDS [6]. Nevertheless, CDS can also potentially lower costs, progress efficiency, and minimize patient inconvenience. In fact, CDS can occasionally deal with all three of these areas simultaneously- for instance, by alerting clinicians to potentially duplicative testing. For more complex cognitive tasks, such as diagnostic decision making, the intention of CDS is to support, rather than to replace, the clinician [7, 8], whereas the CDS may relieve the clinician of the burden of reconstructing orders for each encounter for other tasks (such as presentation of a predefined order set) [9]. The CDS possibly will suggest suggestions, but the clinician ought to filter the information, review the suggestions, and decide whether to take action or what action to take.

Clinical Decision Support Systems are widely categorized into two major groups namely 1) Knowledge based CDSS and 2) Non-knowledge based CDSS [12]. The Knowledge based Clinical Decision Support System comprises rules mostly in the form of IF-Then statements. Generally the data is associated with these rules. For instance, generate warning and more only if the pain intensity is up to a certain level. Generally the knowledge based encloses three main parts - Knowledge base, Inference rules and a mechanism to
communicate. To illustrate the result to the users as well as to afford input to the system, Knowledge base holds the rules, inference engine merges rules with the patient data and the communication mechanism is utilized. The adaptive guidelines from a knowledge base server prove to be much more effective than others in certain case, such as of chest pain management [27]. Vagueness, impreciseness and uncertainty are the fundamental and indispensable aspects of knowledge, so as in several practical problems, the experts face vagueness in feature vectors and uncertainty in decision-making. Basically, a symptom is an uncertain indication of a phenomenon since it may or may not occur with it. Especially, uncertainty characterizes a relation between symptoms and phenomena [28, 29]. In almost every stage of a clinical decision making process, Uncertainty occurs. Sources of uncertainties may comprise that patients cannot describe accurately what has happened to them or how they suffer, doctors and nurses cannot explain exactly what they detect, laboratories report outcomes may be with some degrees of error, physiologists do not precisely understand how the human body works, medical researchers cannot precisely characterize how diseases modify the normal functioning of the body, pharmacologists do not understand the mechanisms entirely accounting for the effectiveness of drugs, and no one can precisely determine one's prognosis [25, 26]. Decision Support Systems that are implemented with the support of Artificial Intelligence have the capability to espouse in new environment and to learn with instance [10, 11]. In Computer Aided Support Systems/ Expert Systems, various methods are exploited to congregate information used for the process of Decision making. Statistical Method, Neural Network, Knowledge Based Methods, Fuzzy Logic Rule Based, Genetic Algorithms and more are included in these methods [12]. Since the idea of computer-based CDSSs emerged at first, significant research has been made in both theoretical and practical. Nevertheless numerous obstacles persist to impede the effective implementation of CDSSs in clinical settings, among which representation and reasoning about medical knowledge predominantly under uncertainty are the areas that need refined methodologies and techniques [23, 24]. In the proposed work, we have proposed an effective clinical decision support system using fuzzy logic in which automatically generated weighted fuzzy rules are used. At first, data preprocessing is applied on the heart disease dataset for removing the missing values and other noisy information. Then, using the class label, the input database is divided into two subsets of data that is then used for mining the frequent attribute category individually. Subsequently, the deviation range is computed using these frequent attribute category so as to compute the relevant attributes. Based on the deviation range, the attributes are selected whether any deviation is there or not. Using this deviation range, the decision rules are constructed and these rules are scanned in the learning database to find its frequency. According to its frequency, the weightage is calculated for every decision rules obtained and the weighted fuzzy rules are obtained with the help of fuzzy membership function. Finally, the weighted fuzzy rules are given to the Mamdani fuzzy inference system so that the system can learn these rules and the risk prediction can be carried out on the designed fuzzy system.

II. RELATED WORKS
For devising Clinical Decision Support Systems, literature presents a number of researches that have made use of Artificial Intelligence and Data Mining techniques. Among them, to support decision makers in risk prediction of heart disease, a handful of researches have been presented. A few of the significant researches obtainable in the literature are explained below.
Using Dempster-Shafer theory of evidence and fuzzy sets theory, Vahid Khatibi and Gholam Ali Montazer [13] have proposed an inference engine named fuzzy-evidential hybrid inference engine. The hybrid engine functions in two phases. In the initial phase, through fuzzy sets, it models the input information’s vagueness. In following, it applies the fuzzy inference rules on the acquired fuzzy sets to generate the first phase results by extracting the fuzzy rule set for the problem. At the subsequent phase, the attained consequences of preceding stage were assumed as basic beliefs for the problem propositions and in this method, the belief and plausibility functions (or the belief interval) are positioned. They have afforded diverse basic beliefs which should be exploited to generate an integrative outcome by gathering information from diverse sources. It has yielded 91.58 percent accuracy rate for its accurate prediction by applying the proposed engine on the Coronary Heart Disease (CHD) risk assessment. The hybrid engine precisely models the information’s vagueness and decision-making’s uncertainty and through information fusion, affords further accurate results.
For the diagnosis of Coronary Artery Disease (CAD), Markos G. Tsipouras et al. [14] have proposed a fuzzy rule-based Decision Support System (DSS). Using a four stage methodology: 1) induction of a decision tree from the data; 2) extraction of a set of rules from the decision tree, in disjunctive normal form and formulation of a crisp model; 3) transformation of the crisp set of rules into a fuzzy model; and 4) optimization of the parameters of the fuzzy model, the system was automatically generated from an initial annotated dataset. The dataset utilized for the DSS generation and evaluation comprises 199 subjects, each one characterized by 19 features, in addition to demographic and history data, as well as laboratory examinations. Tenfold cross validation was applied, and using the set of rules extracted from the decision tree (first and second stages), the average sensitivity and specificity obtained is 62% and 54%, respectively, while the average sensitivity and specificity increase to 80% and 65%, respectively, when the fuzzification and optimization stages are exploited. Since it was automatically generated, the system suggests numerous advantages, it affords CAD diagnosis based on easily and non invasively acquired features, and was able to afford interpretation for the decisions made.
For the diagnosis of coronary artery disease based on evidence Noor Akhmad Setiawan et al. [15] have developed a fuzzy decision support system. The Coronary Artery Disease data sets obtained from University California Irvine (UCI) are utilized. By using rules extraction method based on Rough Set Theory, the knowledge base of fuzzy decision support system was taken. Based on information from discretization of numerical attributes, the rules then were selected and fuzzified. Using the information from support of extracted rules, Fuzzy rules weight was proposed. To verify the proposed system, UCI heart disease data sets collected from U.S., Switzerland and Hungary, data from Ipoh Specialist Hospital Malaysia is used. The results revealed that the system was capable to provide the percentage of coronary artery blocking better than cardiologists and angiography. The consequences of the proposed system were verified and authenticated by three expert cardiologists.

To investigate factors that contribute significantly to enhancing the risk of acute coronary syndrome Tahseen A. Jilani et al. [16] have utilized data mining techniques. They have presupposed that the dependent variable was diagnosis – with dichotomous values showing presence or absence of disease. They have applied binary regression to the factors distressing the dependent variable. The data set has been obtained from two diverse cardiac hospitals of Karachi, Pakistan. They have total sixteen variables out of which one was presupposed dependent and other 15 are independent variables. Data Reduction techniques like principle component analysis was applied for better performance of the regression model in predicting Acute Coronary Syndrome. They have considered only 14 out of sixteen factors on the basis of data reduction results.

Using neural network, Shantakumar B. Patil and Y.S. Kumaraswamy [17] have proposed an intelligent and effective heart attack prediction system. For the extraction of significant patterns from the heart disease warehouses for heart attack prediction, a proficient methodology has been proposed. Initially, in order to make it suitable for the mining process, the data warehouse was pre-processed. Once the preprocessing gets ended, the heart disease warehouse was clustered with the support of the K-means clustering algorithm, which will extract the data appropriate to heart attack from the warehouse. Consequently with the aid of the MAFIA algorithm from the data extracted, the frequent patterns applicable to heart disease are mined. In addition, on basis of the computed significant weight age, the patterns vital to heart attack prediction are selected. For the effectual prediction of heart attack, the neural network was trained with the preferred significant patterns. With Back-propagation as the training algorithm, they have employed the Multi-layer Perceptron Neural Network. The consequences thus attained have illustrated that the designed prediction system was skilled of predicting the heart attack efficiently.

A prototype Intelligent Heart Disease Prediction System (IHDPS) has been developed by Sellappan Palaniappan and Rafiah Awang [18] using Data Mining techniques, namely, Decision Trees, Naive Bayes and Neural Network. Results exposed that in realizing the objectives of the defined mining goals, each technique has its exclusive strength. IHDPS can respond complex “what if” queries whereas traditional decision support systems is unable to answer. It can foretell the possibility of patients getting a heart disease, using medical profiles such as age, sex, blood pressure and blood sugar. It facilitates significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be recognized. IHDPS is Web-based, user-friendly, scalable, reliable and expandable.

To utilize Artificial Intelligence tools as a clinical decision support in assessing cardiovascular risk in patients Beatrice Fidele et al. [19] have presented a research study. In the proposed artificial neural network, a two-layer neural network employing the Levenberg-Marquardt algorithm and the resilient back propagation have been utilized. It has been shown by exploiting the Long Beach dataset, how this network was efficient in predicting cardiovascular risk in individual patients. At an individual level the application of the network seems to better deal with the prediction of cardiovascular disease.

The ability of fuzzy neural network model to predict the likelihood of coronary heart disease has been evaluated by Basir Abidin et al. [20] for individuals based on knowledge of their biomarkers, risk habits and demographic profiles. The prediction performance of fuzzy neural network models were calculated in terms of percentage accuracies and compared with the prediction performance of logistic regression models. Provisionary consequences have illustrated that for the prediction of coronary heart disease in the sample studied, four markers namely body mass index, systolic blood pressure, total cholesterol level, and age are the appropriate markers. Fuzzy neural network models prediction performance were found to be sophisticated to the logistic regression performance in addition to other outcomes that are reported in related literature.

III. DISCUSSION OF HEART DISEASE DATASET

A. Heart Disease

Heart disease refers to numerous problems that distress the heart and the blood vessels in the heart. Coronary artery disease (CAD), Angina, Heart attack and Heart failure are some examples for the diverse types of heart diseases. Coronary heart disease (CHD) is a key reason of sickness and death in the modern society. The expense of handling of CHD is a major economic load and so prevention of coronary artery disease is an extremely essential step in the management. Some of the several methods that can be used for CHD prevention include health promotion activities, special protection schemes, programs to change way of living, identification in advance and excellent control of risk factors and continuous observation of arising risk factors [32]. Accumulation of plaques inside the walls of the coronary arteries that deliver blood to the myocardium cause CAD. Damage to myocardium may result due to the continued temporary oxygen deprivation that may be caused by CAD [15]. The
term ‘cardiovascular disease’ that represents a category of heart disease comprises a broad variety of conditions that upset the heart and the blood vessels and the way in which blood is pumped and circulated in the body [34]. CHD is caused by the decreased blood and oxygen supply to the heart due to the narrowing of the coronary arteries. CHD includes myocardial infarctions, commonly called as a heart attacks, and angina pectoris, or chest pain [17]. Heart attack results due to the abrupt blockage of a coronary artery, usually because of a blood clot. Insufficient blood flow to the heart muscles results in chest pains [35]. There are several types of cardiovascular disease such as high blood pressure, coronary artery disease, valvular heart disease, stroke, or rheumatic fever/rheumatic heart disease [36].

B. Dataset Description

The dataset is taken from Data Mining Repository of University of California, Irvine (UCI) [31]. Finally the system is validated using data sets from Cleveland, Hungarian and Switzerland.

(1) Cleveland data. Robert Detrano, M.D., Ph.D. collected these data at V.A. Medical centre. All published experiments relate to using a subset of 14 of the 76 attributes present in the processed Cleveland Heart Disease database. Specifically, ML researchers use only the Cleveland database till today. The existence of heart disease in the patient is indicated in the "goal" field by means of an integer that can take any value from 0 (no presence) to 4. Distinguishing disease existence (values 1, 2, 3, 4) from non-existence (value 0) has been the focus of the experiments conducted on the Cleveland database [30]. Six of the examples have been discarded because they had missing values. Class distributions are 54% heart disease absent, 46% heart disease present.

(2) Hungarian data. Andras Janosi, M.D. collected these data at the Hungarian Institute of Cardiology, Budapest. Due to a huge percentage of missing values three of the attributes have been discarded but the format of the data is exactly the same as that of the Cleveland data. 34 examples of the database were discarded on account of missing values and 261 examples were present. Class distributions are 62.5% heart disease absent and 37.5% heart disease present [33].

(3) Switzerland data. William Steinbrunn, M.D collected these data at the University Hospital, Zurich, Switzerland. Switzerland data has more number of missing values.

IV. AN EFFICIENT CLINICAL DECISION SUPPORT SYSTEM TO RISK LEVEL PREDICTION OF HEART DISEASE

Normally, directly support clinical decision making is the intention behind the design of a clinical decision support system and it presents patient-specific assessments or recommendations produced using characteristics of individual patients to clinicians for consideration [38]. In recent years, clinical decision support system based on computer-aided diagnosis methodologies have been proposed in the literature which by evaluating the data obtained by some of the methods or other sources (i.e., laboratory examinations, demographic and/or history data, etc.) from a computer-based application leads to a computer-aided diagnosis. The data analysis methods used in most of the proposed methods cannot provide clear and direct explanation for the decisions made to examine the risk factors for cardiovascular diseases as they are based on neural networks. Hence, a method based on easily obtained features capable of calculating the risk level of computer-aided diagnosis and providing explanation for the decisions made would be of immense clinical value [14]. So, the soft computing technique in particular the fuzzy logic technique could be used for assessing the risk level of heart patients in developing the clinical decision support system of heart disease diagnosis.

In the proposed system, the biomarkers for cardiovascular diseases described in the literature are age, sex, total cholesterol level, HDL, LDL, age, smoking status, hypertension, and pre-eclampsia that are mainly used to predict the risk level of heart patients. For better prediction of risk level, we make use of fuzzy logic, where the decision to be taken for the heart disease of patients is based on the weighted fuzzy rules. By considering the CDSS based on fuzzy logic, the efficiency is mainly depends on the fuzzy rules employed in the system. In general, the domain experts or professionals in the corresponding domain provide the fuzzy rules for prediction problem. But, here, we automatically generated the fuzzy rules to provide the better learning of fuzzy system using historic data. In addition to, the fuzzy rules are weighted in accordance with their importance using the attribute weightage. These weighted fuzzy rules are applied on the rule base of the fuzzy system and then the prediction can be carried out on the designed fuzzy-based CDSS. The detailed steps involved in the proposed clinical decision support system are explained in the following figure 1 and figure 2.

A. Data preprocessing

The purpose of data preprocessing is to extract useful data from raw heart disease datasets and then these data should be converted into the format necessary for prediction of risk level. Due to the irrelevant information in the heart disease datasets, the original raw data cannot be directly used in the prediction procedure, hence in data preprocessing phase, raw data need to be cleaned, analyzed and transformed for further step. So, the irrelevant attributes and information in the raw data are identified and it is converted into the row-column format after removing the irrelevant one. Here, each row represents the patient information and the column indicates a list of attributes (biomarkers). The last column gives the class label that corresponds to the risk level of the heart patients.

B. Classification of training dataset based on risk level

After data processing, the input training dataset used for prediction is classified into two subsets of data based on the class label described in the data. The input training dataset consists of two class labels, in which ‘0’ indicates that the disease status is less than 50% and ‘1’ represent that the risk level for heart disease is more than 50%. Using these two class
label, the dataset \( D_H \) is divided into two subsets of data, \( D_H = \{ D_{Hj} \mid 1 \leq j \leq 2 \} \), where, \( j \) denotes the class label that describes the risk level of patients. In addition to, each class contains \( m \) number of attributes and each attributes \( (\beta_i) \) presented in the dataset consists of attribute category that is the continuous symbolic value specified for every patients. The two subsets of data \( D_{Hj} \) obtained are then employed for generating a better set of weighted fuzzy rules automatically so that the fuzzy system can learn the rules effectively.

C. Automated approach to generate decision tree rules

This section describes the automated approach to generate the weighted fuzzy rules from the classified dataset in order to effectively learn the fuzzy system. By considering the heart disease datasets, a large number of attributes is presented but, the extraction significant attributes exactly suitable for prediction is very important. In order to choose the most relevant and important attributes, we have used the frequent attribute category that is mined from the input datasets. Then, based on the frequency of attribute category and the weightage of attributes, the fuzzy rules are generated automatically. The steps to be used for automatic generation of fuzzy rules are discussed in this sub-section.

**Fig. 1. Clinical decision support system**

**Fig. 2. Proposed fuzzy-based clinical decision support system**

(1) Mining of attribute category

In this step, the frequent attribute category corresponds to every attributes \( \beta_i \) presented in the datasets \( D_{Hj} \) should be mined so that the frequency of every attribute category within the class \( C_j \) is obtained by scanning the database. Here, the well known algorithms, such as Apriori [40] and FP-growth [37] are not suitable for mining of the frequent attribute category because these data format is different from the data that are suitable for traditional algorithms. Here, we simply mine the one length attribute category by finding the frequency in the database and then, the attribute category of the attributes \( \beta_i \) within the class \( C_j \) are arranged in accordance with their frequency. Then, for every attributes, a set of attribute category are selected from the sorted list based minimum support. The selected attribute category is then stored in a two vector, \( V_{MIN}^j \) and \( V_{MAX}^j \) for each class, in which one vector contains the minimum value corresponds to the attribute category of every attribute and second vector contains the maximum value corresponds to the attribute category of every attribute. It is denoted as, 

\[
V_{MIN}^j = \{\beta_{min1}^j, \beta_{min2}^j, \ldots, \beta_{minm}^j\}
\]

\[
V_{MAX}^j = \{\beta_{max1}^j, \beta_{max2}^j, \ldots, \beta_{maxm}^j\}
\]

(2) Selection of suitable attributes

Here, the suitable attributes is then identified employing the two vectors, \( V_{MIN}^j \) and \( V_{MAX}^j \) obtained from the previous step. The reason behind this step is that the input data contains a large number attributes, in which all the attributes are not so effective in predicting risk level of heart patient. So, the identification of suitable attributes should ensure the better accuracy in risk level prediction. For identifying the suitable attributes, we have used deviation method, where mined 1-length attribute category is used. The deviation range for the entire element presented in the minimum vector of two classes, \( V_{MIN}^1 \) and \( V_{MIN}^2 \) is identified by performing the one-to-one comparison of respective location. Similar way, the deviation range can be identified for the two maximum vectors of two classes, \( V_{MAX}^1 \) and \( V_{MAX}^2 \). The minimum and maximum deviation vector thus obtained is represented as, \( D_{MIN} \) and \( D_{MAX} \). Then, the suitable attributes are chosen if the deviation is found out, otherwise it is eliminated. The effective attributes selected for rule generation process is represented as, 

\[
A = \left[ \beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(k)} \right]
\]

Where, \( k \leq m \).
(3) Generation of decision rules and rule weighting
Rule generation and rule weighting is an important step for developing fuzzy-based clinical decision support system. The deviation vector, \( D_{\text{MAX}} \) and \( D_{\text{MIN}} \) obtained from the previous step is employed here to generate the decision rules that specified the risk level of heart patients in terms of numerical variables. The rules are generated automatically from the two deviation vectors that contain the deviation of each attributes comparing two classes. From the equal size deviation vector, three decision rules are generated from every element by comparing the corresponding elements of both vectors. Suppose, assume that the first element of the vector, \( D_{\text{MIN}} \) and \( D_{\text{MAX}} \) is ‘3’ and ‘8’, the corresponding generated decision rules are, “IF \( \beta^{(1)} \) is less than 3, THEN the risk is less than 50” and “IF \( \beta^{(1)} \) is greater than 8, THEN the risk is greater than 50” and “IF \( \beta^{(1)} \) is in between the 3 and 8, THEN the risk is either less than 50 or greater than 50”. These rules are then weighted based on the importance level of it in the database \( D_H \). For each rule generated, we find the number of patients (\( M \)) satisfy these rules by scanning the database \( D_H \). Once we find the value of \( M \) for the rule \( (R_1 \rightarrow R_2) \), the weightage of the rule is found by using the following formulae,

\[
W(R_1 \rightarrow R_2) = \frac{M_{(R_1 \rightarrow R_2)}}{N}
\]

Where, \( M_{(R_1 \rightarrow R_2)} \rightarrow \) Number patients who satisfy the rule, \( R_1 \rightarrow R_2 \)
\( N \rightarrow \) Total number of patients in the database

(4) Finding weighted fuzzy rules and decision tree rules
The extremely important task of generating fuzzy rules from the data described using numeric-symbolic values appears to be extremely difficult. Handling this type of values is extremely important because it is very close to human knowledge and rules with such values are normally more comprehensible and accountable when compared to rules with numerical values. Handling such values is permitted by the introduction of fuzzy set theory which by the construction of fuzzy leads to the generation of a set of fuzzy rules. The automatic method proposed here is based on the construction of fuzzy modalities that enables to generate fuzzy values from a set of rules with numerical values. The decision rules obtained from the previous contain IF and THEN part, in which IF part specifies the numerical variable and THEN part specifies the class label. At first, the numerical variable specified in the IF part of the decision rules is converted into the linguistic variable according to the fuzzy membership function and THEN part of the fuzzy rules is similar to that of decision rules. For example, “IF \( \beta^{(1)} \) is LOW, THEN the risk is less than 50 (class ‘0’) and “IF \( \beta^{(1)} \) is MEDIUM, THEN the risk is either less than 50 or greater than 50 (class ‘0’ or ‘1’) and “IF \( \beta^{(1)} \) is HIGH, THEN the risk is greater than 50 (class ‘1’)”. Similar way, we process entire decision rules with numeric variable and they are converted into the fuzzy rules using membership function. A group of fuzzy IF-THEN rules obtained is belonging to one of the most popular, most effective, and user friendliest knowledge representations so as to provide the effective learning for fuzzy system.

D. Developing clinical decision support system using fuzzy logic
This section describes the developing of clinical decision support system using fuzzy logic system for assessing the risk level of heart patient. Fuzzy logic introduced by Zadeh in the late 1960s [39] is the rediscovery of multi-valued logic designed by Lukasiewicz. A fuzzifier, fuzzy rules, fuzzy inference engine and defuzzifier exists in a fuzzy logic model.

**Fuzzifier:** Firstly, fuzzy linguistic variables, fuzzy linguistic terms and membership functions are used to convert a gathered crisp set of input data into a fuzzy set. This step is known as fuzzification. Enabling interpretation of fuzzy condition in a rule is the purpose of the fuzzification process.

**Fuzzy rule base:** The fuzzy rules that are important for any fuzzy system are defined after the inputs are fuzzified. A fuzzy rule contains a condition and a conclusion and its structure is similar to the IF-THEN rule. An entire fuzzy rule that is created to control the output variable exists in the rule base.

**Inference engine:** Subsequently, a set of rules defined in the fuzzy rule base is used as the basis for interpreting and by employing reasoning fuzzy outputs are generated. **Defuzzifier:** A fuzzy set (the aggregate output fuzzy set) is used as the input for the defuzzification process and membership functions based mapping of fuzzy sets to a crisp output is used to obtain a single number as the output. The general structure of the fuzzy logic system is shown in figure 4.

The designed clinical decision support system shown in figure 3 contains ‘m’ inputs and one output, where inputs are related to the ‘m’ attributes and output is related to the class label (risk level). Here, m-input, single-output of Mamdani fuzzy inference system with centroid of area defuzzification strategy was used for this purpose. Here, each input fuzzy set defined in the fuzzy system includes four membership functions (VL, L, M and H) and an output fuzzy set contains two membership functions (L and H). Each membership function used triangular function for fuzzification strategy. The fuzzy rule base contains the set of weighted fuzzy rules obtained from the proposed procedure to learn the system.

**Fig. 3.** Designed fuzzy inference system based on weighted fuzzy rules
V. RESULTS AND DISCUSSION

The proposed fuzzy logic-based clinical decision support system has been implemented using MATLAB (7.10). The experimental results of the clinical decision support system for risk prediction are explained in this section. Here, the performance of the proposed system is compared with the neural network-based system to evaluate the sensitivity, specificity and accuracy. 

A. Experimental environment and datasets

For experimentation, we have taken Cleveland, Hungarian and Switzerland heart disease dataset [41] which is a widely-accepted database obtained from UCI machine learning repository. The heart disease dataset are divided into two sets such as, (1) Training dataset (2) Testing dataset. The training dataset is used to generate the weighted fuzzy rules and the testing dataset is used to analyze the performance of the proposed system. Table 1 provides the description of these datasets.

Table 1. Details of datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Total instance</th>
<th>Training Data</th>
<th>Testing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleveland</td>
<td>303</td>
<td>202</td>
<td>101</td>
</tr>
<tr>
<td>Hungarian</td>
<td>294</td>
<td>196</td>
<td>98</td>
</tr>
<tr>
<td>Switzerland</td>
<td>123</td>
<td>82</td>
<td>41</td>
</tr>
</tbody>
</table>

B. Selected attributes for risk level prediction

At first, the suitable attributes are chosen from the training dataset using the proposed approach and then, the fuzzy rules are generated based on the chosen attribute. The attributes chosen by the proposed system for risk prediction is given in table 2.

Table 2. Selected attributes for fuzzy rule generation

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Selected attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleveland</td>
<td>Age, Trestbps (resting blood pressure), Chol (serum cholesteral in mg/dl), Thalach (maximum heart rate achieved), Oldpeak (ST depression induced by exercise relative to rest), Thal.</td>
</tr>
<tr>
<td>Hungarian</td>
<td>Age, Trestbps, Chol, Restecg (resting electrocardiographic results), Thalach, Exang (exercise induced angina), Oldpeak, Slope (slope of the peak exercise ST segment).</td>
</tr>
<tr>
<td>Switzerland</td>
<td>Age, Sex, Cp (chest pain type), Trestbps, Fbs (fasting blood sugar), Restecg, Thalach, Oldpeak, Slope, Ca (number of major vessels), Thal.</td>
</tr>
</tbody>
</table>

C. Performance analysis

In testing phase, the testing dataset is given to the proposed system to find the risk prediction of heart patients and obtained results is evaluated with the evaluation metrics namely, sensitivity, specificity and accuracy [42]. In order to find these metrics, we first compute some of the terms like, True positive, True negative, False negative and False positive based on the definitions given in table 3.

\[
Sensitivity = \frac{TP}{TP + FN} \\
Specificity = \frac{TN}{TN + FP} \\
Accuracy = \frac{TN + TP}{TN + TP + FN + FP}
\]

Where, \( TP \rightarrow \) True positive \\
\( TN \rightarrow \) True negative \\
\( FN \rightarrow \) False negative \\
\( FP \rightarrow \) False positive

Table 3. Terms used to define sensitivity, specificity and accuracy

<table>
<thead>
<tr>
<th>Outcome of the diagnostic</th>
<th>Condition (e.g. Disease)</th>
</tr>
</thead>
<tbody>
<tr>
<td>As determined by the Standard of Truth</td>
<td></td>
</tr>
</tbody>
</table>
The evaluation metrics are computed for both training and testing dataset in the testing phase and the obtained result for Cleveland, Hungarian and Switzerland datasets are given in the following tables. Table 4 shows the overall performance of the proposed system in risk prediction, in which, class 0 indicates that the risk level is below the 50% and class 1 indicates the risk level is above 50%.

The performance of the proposed clinical decision support system was analyzed with three different heart disease datasets. The results obtained from these datasets are given in the sub-section 5.3. Here, we compare the performance of the proposed clinical decision support system in risk prediction with the neural network-based system using these datasets. The results obtained for these three datasets are given in the following figures. By analyzing the plotted graphs, the performance of the proposed clinical support system has significantly improved the risk prediction compared with the neural network-based clinical support system.

### Table 4. The performance of the proposed clinical decision support system in risk prediction

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Class</th>
<th>Metric</th>
<th>Proposed System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Training Testing</td>
</tr>
<tr>
<td></td>
<td>&lt;50%</td>
<td>Accuracy</td>
<td>0.509901 0.623529</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>0.724771 0.765957</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.258065 0.447368</td>
</tr>
<tr>
<td>Cleveland</td>
<td></td>
<td>Accuracy</td>
<td>0.509901 0.623529</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>0.258065 0.447368</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.724771 0.765957</td>
</tr>
<tr>
<td></td>
<td>&gt;50%</td>
<td>Accuracy</td>
<td>0.715054 0.469388</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>0.80 0.31746</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.540984 0.742857</td>
</tr>
<tr>
<td>Hungarian</td>
<td>&lt;50%</td>
<td>Accuracy</td>
<td>0.509901 0.623529</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>0.724771 0.765957</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.258065 0.447368</td>
</tr>
<tr>
<td></td>
<td>&gt;50%</td>
<td>Accuracy</td>
<td>0.715054 0.469388</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>0.80 0.31746</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.540984 0.742857</td>
</tr>
</tbody>
</table>

### D. Comparative study

The results obtained from these datasets are given in the following figures. By analyzing the plotted graphs, the performance of the proposed clinical support system has significantly improved the risk prediction compared with the neural network-based clinical support system.

**Fig. 5.** Risk prediction for the patients who comes under below 50% (a) Training dataset (b) Testing dataset

**Fig. 6.** Risk prediction for the patients who comes under above 50% a) Training dataset (b) Testing dataset
VI. CONCLUSION

We have presented a weighted fuzzy rule-based clinical decision support system (CDSS) for computer-aided diagnosis of heart disease. The automatic procedure to generate the fuzzy rules is an advantage of the proposed system and the weighted procedure introduced in the proposed work is additional advantage for effective learning of the fuzzy system. The proposed clinical decision support system for risk prediction of heart patients contains two steps such as, (1) generation of weighted fuzzy rules, and (2) developing of a fuzzy rule-based decision support system. Here, the suitable attribute were generated after applying the mining procedure and these attributes were used to generate the fuzzy rule that are then weighted based on the frequency in the learning datasets. These weighted fuzzy rules were used to build the clinical decision support system using Mamdani fuzzy inference system. Finally, the experimentation was carried out on the UCI machine learning repository and the results in risk prediction ensured that the proposed clinical decision support system was improved significantly compared with the network-based system in terms of accuracy, sensitivity and specificity.

REFERENCES

Dr. Anooj P.K. received the B.Sc degree in Physics from Mahatma Gandhi University, Kerala and his MCA from Indira Gandhi University, New Delhi. He has also done hardware specialization post graduate diploma from Lal Bahadur Shastri Centre for Science and Technology, Kerala. He got his PhD in Computer Applications from Bharath University, Chennai, India.

September 2012  ATC-60203031©Asian-Transactions 10
He has been working as senior teacher and Lecturer in Computer Science in various institutions from year 2000. Since 2008, he is working as Registrar and Lecturer in Department of Information Technology, Al Musanna College of Technology, Directorate General of Technological Education, Ministry of Manpower, Oman.

He is now the reviewer of a couple of well renowned international journals. He is also National Coordinator and associative member of International Geochange committee and member of few international organizations.