Journal of King Saud University – Computer and Information Sciences (2012) **24**, 27–40

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules

P.K. Anooj *

Department of Information Technology, Al Musanna College of Technology, Directorate General of Technological Education, Ministry of Manpower, Oman

Received 23 June 2011; accepted 6 September 2011
Available online 22 November 2011

KEYWORDS

Clinical decision support system (CDSS);
Heart disease;
Fuzzy logic;
Weighted fuzzy rules;
Attribute selection;
Risk prediction;
UCI repository;
Accuracy;
Sensitivity and specificity

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 the medical domain application is the 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. Earlier, the use of computer was to build a 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 experts' opinions 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 knowledge from the patient's clinical data. The proposed clinical decision support system for the risk prediction of heart patients consists of two phases: (1) automated approach for the 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.

* Mobile: +968 95710984.
E-mail addresses: anoojpk@gmail.com, anoojphd@gmail.com

1319-1578 © 2011 King Saud University. Production and hosting by Elsevier B.V. All rights reserved.

Peer review under responsibility of King Saud University.
doi:10.1016/j.jksuci.2011.09.002

Production and hosting by Elsevier

1. Introduction

To extract hidden patterns and relationships from large databases, data mining merges statistical analysis, machine learning and database technology (Thuraisingham, 2000). In several areas of medical services, including prediction of the effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data, data mining techniques have been applied (Tang et al., 2005). 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 health care 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 (Subbalakshmi et al., 2011). Clinical decisions are often taken on the basis of doctors' perception and experience rather than on the knowledge rich data masked in the database (Parthiban and Subramanian, 2008). However unfortunately every doctor's expertise is not even in every sub-specialty and is in several places as a scarce resource (Parthiban and Subramanian, 2008). 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 the same category. The physicians may not be capable to diagnose it accurately (Shanthi et al., 2008). Unfortunately, due to complex interdependence on a variety of factors, accurate diagnosis of disease at a premature stage is quite a challenging task (Yan et al., 2003).

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 (Merijohn et al., 2008). 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 (Garg et al., 2005). Nevertheless, CDS can also be 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 (Miller, 1990; Miller and Masarie, 1990), whereas the CDS may relieve the clinician of the burden of reconstructing orders for each encounter for other tasks (such as the presentation of a predefined order set) (Osheroff, 2009). The CDS possibly will provide 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 (Abbasi and Kashiyarndi, 2006). The knowledge based clinical decision support system comprises rules mostly in the form of IF–Then statements. Generally the data are 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 CDSS 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 cases, such as that of chest pain management (Ali et al., 1999). 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 (Straszecka, 2007; De et al., 2001).

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, laboratory reports' 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 (Szolovit, 1995; Kong et al., 2008). Decision support systems that are implemented with the support of artificial intelligence have the capability to espouse in a new environment and to learn with instance (Warren et al., 2000; Anderson, 1997). 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 (Abbasi and Kashiyarndi, 2006). Since the idea of computer-based CDSSs emerged at first, significant research has been made in both theoretical and practical areas. 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 (Lin et al., 2006; Musen et al., 2001).

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 are then used for mining the frequent attribute category individually. Subsequently, the deviation range is computed using these frequent attribute categories so as to compute the relevant attributes. Based on the deviation range, the attributes are selected whether any deviation exists 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 rule 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.

2. 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 the 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, Khatibi and Montazer (2010) 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. Further, 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 a 91.58% 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), Tsipouras et al. (2008) 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 are 62% and 54%, respectively, while the average sensitivity and specificity increases 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 noninvasively acquired features, and was able to afford interpretation for the decisions made.

For the diagnosis of coronary artery disease based on evidence Setiawan et al. (2009) have developed a fuzzy decision support system. The coronary artery disease data sets obtained from the University of 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 the discretization of numerical attributes, the rules then were selected and fuzzified. Using the information from the support of extracted rules, fuzzy rules weights were proposed. To verify the proposed system, UCI heart disease data sets collected from US, Switzerland and Hungary, data from Ipoh Specialist Hospital Malaysia are 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 Jilani et al. (2009) have utilized data mining techniques. They have presupposed that the dependent variable was diagnosed – with dichotomous values showing the presence or the 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 a total of 16 variables out of which one was presupposed dependent and other 15 are independent variables. Data reduction techniques like principle component analysis were applied for better performance of the regression model in predicting acute coronary syndrome. They have considered only 14 out of 16 factors on the basis of data reduction results.

Using neural network, Patil and Kumaraswamy (2009) 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 to the mining process, the data warehouse was pre-processed. Once the preprocessing 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 weightage, 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 in predicting heart attack efficiently.

A prototype intelligent heart disease prediction system (IHDPS) has been developed by Palaniappan and Awang (2008) 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 to complex "what if" queries whereas the traditional decision support system 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 clinical decision support in assessing cardiovascular risk in patients Fidele et al. (2009) 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 Abidin et al. (2009) 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 performances were found to be sophisticated to the logistic regression performance in addition to other outcomes that are reported in the related literature.

3. Discussion of heart disease dataset

3.1. 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 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 the continuous observation of arising risk factors (Knopp, 2002). Amassment of plaques inside the walls of the coronary arteries that deliver blood to the myocardium causes CAD. Damage to the myocardium may result due to the continued temporary oxygen deprivation that may be caused by CAD (Setiawan et al., 2009). 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. 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 heart attacks, and angina pectoris, or chest pain (Patil and Kumaraswamy, 2009). 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 (Chen and Greiner, 1999). 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 (http://chinese-school.netfirms.com/heart-disease-causes.html).

3.2. Dataset description

The data set is taken from Data Mining Repository of the University of California, Irvine (UCI) (Newman et al., 1998). Finally the system is validated using data sets from Cleveland, Hungary and Switzerland. In those datasets, totally, 14 attributes such as Age, sex, chest pain type, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression, slope of the peak exercise ST segment, number of major vessels, thal and diagnosis of heart disease are presented.

3.2.1. Cleveland data

Robert Detrano, M.D., Ph.D., collected these data at V.A. Medical Centre. All published experiments related 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 1–4) from non-existence (value 0) has been the focus of the experiments conducted in the Cleveland database (Ephzibah, 2010). Six of the examples have been discarded because they had missing values. Class distributions are 54% heart disease absent, 46% heart disease present.

3.2.2. 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. Thirty-four 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 (Bradley, 1997).

3.2.3. Switzerland data

William Steinbrunn, M.D., collected these data at the University Hospital, Zurich, Switzerland. Switzerland data has more number of missing values. It contains 123 data instances and 14 attributes. Class distributions are 6.5% heart disease absent and 93.5% heart disease present.

4. An Efficient clinical decision support system to risk level prediction of heart disease

Normally, direct 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 the characteristics of individual patients to clinicians for consideration (Kawamoto et al., 2005). In recent years, clinical decision support system based on computer-aided diagnosis methodologies have been proposed in the literature by which 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 (Tsipouras et al., 2008). 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 on heart disease of patients is based on the weighted fuzzy rules. By considering the CDSS based on fuzzy logic, the efficiency 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 which, 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 Figs. 1 and 2.

4.1. 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 the 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

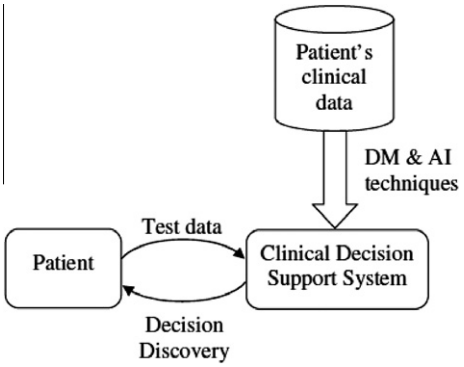

Figure 1 Clinical decision support system.

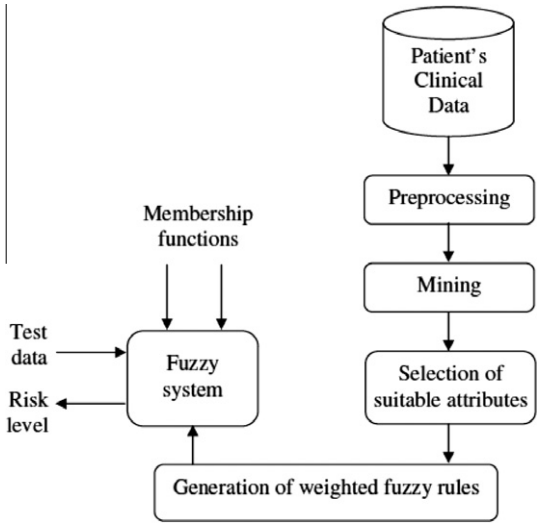

Figure 2 Proposed fuzzy-based clinical decision support system.

data need to be cleaned, analyzed and transformed for further steps. So, the irrelevant attributes are identified using the procedure discussed in Section 4.3.2 and are converted into the row–column format after removing the irrelevant ones. 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.

4.2. 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. Actually, the heart disorder is usually a blood vessel problem, for example, blockage of the blood flow or narrowing of the blood vessel. There are two output classes for the diagnosis of heart disease, less than 50% diameter narrowing (no heart disease) or greater than 50% diameter narrowing (has heart disease). Here, value 0 represents *no presence* of heart disease which means less than 50% diameter narrowing and values 1–4 represent the presence of heart disease that means greater the 50% diameter narrowing. According to, we modified the chosen UCI data into two class labels. In all three datasets, class 0 specifies the *no presence* of heart disease (less than 50% diameter narrowing) and class 1 specifies the *presence* of heart disease (greater the 50% diameter narrowing). Using these two class labels, the dataset (D_H) is divided into two subsets of data, $D_H = \{D_{Hj}; 1 \leqslant j \leqslant 2\}$, where, j denotes the class label that describes the risk level of patients. In addition to which, each class contains m number of attributes and each attribute (β_i) presented in the dataset consists of an attribute category that is the continuous value specified for every patient. 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.

4.3. Automated approach to generate weighted fuzzy 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 significance attributes exactly suitable for prediction are 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 the automatic generation of fuzzy rules are discussed in this sub-section.

4.3.1. Mining of attribute category

In this step, the frequent attribute category corresponding to every attribute β_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. The frequency of attribute category is computed by finding the occurrence of the attribute category in the whole dataset. For continuous attributes, we discretize them in equi-width to find the frequency. Here, the well known algorithms, such as Apriori (Agrawal et al., 1993) and FP-growth (Han et al.,

2004) are not suitable for mining of the frequent attribute category because these data formats are 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 β_i within the class C_j are arranged in accordance with their frequency. Then, for every attribute, a set of attribute categories are selected from the sorted list based minimum support. The selected attribute category is then stored in a two vector, V^j_{MIN} and V^j_{MAX} for each class, in which one vector contains the minimum value corresponds to the attribute category of every attribute and second vector that contains the maximum value corresponds to the attribute category of every attribute. It is denoted as, $V^j_{MIN} = \{\beta_{\min 1}, \beta_{\min 2}, ..., \beta_{\min m}\}$ and $V^j_{MAX} = \{\beta_{\max 1}, \beta_{\max 2}, ..., \beta_{\max m}\}$.

4.3.2. Selection of suitable attributes

Here, the suitable attributes are then identified employing the two vectors, V^j_{MIN} and V^j_{MAX} obtained from the previous step. The reason behind this step is that the input data contain a large number of attributes, in which all the attributes are not so effective in the predicting risk level of the 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 the 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^1_{MIN} and V^2_{MIN} is identified by performing the one-to-one comparison of the respective location. In a similar way, the deviation range can be identified for the two maximum vectors of two classes, $V^1_{MAX}$and V^2_{MAX}. The minimum and maximum deviation vectors thus obtained are represented as, D_{MAX} and D_{MIN}. 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, $A = [\beta^{(1)}, \beta^{(2)},...,\beta^{(k)}]$, where, $k \leqslant m$.

4.3.3. Generation of decision rules and rule weighting

Rule generation and rule weighting are an important step for developing fuzzy-based clinical decision support system. The deviation vectors, D_{MAX} and D_{MIN} obtained from the previous step are 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 attribute 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 vectors, D_{MIN} and D_{MAX} are '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 3 and 8, THEN the risk is either less than 50 or greater than 50*". These rules are then weighted based on its importance level 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 $M^{(R_1 \rightarrow R_2)}$ is the number of patients who satisfy the rule, $R_1 \rightarrow R_2$ and N is the total number of patients in the database.

4.3.4. Finding weighted fuzzy rules

The extremely important task of generating fuzzy rules from the data described using numeric–symbolic values appears to be extremely difficult. Handling these types 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 the generation of fuzzy values from a set of rules with numerical values. The decision rules obtained from the previous contain IF and THEN parts, 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 the 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')*". In a 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-friendly knowledge representations so as to provide the effective learning for the fuzzy system.

4.4. 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 the heart patient. Fuzzy logic introduced by Zadeh in the late 1960s (Zadeh, 1965) is the rediscovery of multi-valued logic designed by Lukasiewicz. A fuzzifier, fuzzy rules, fuzzy inference engine and defuzzifier exist 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 Fig. 3.

The designed clinical decision support system shown in Fig. 4 contains 'm' inputs and one output, where inputs are related to the 'm' attributes and output is related to the class

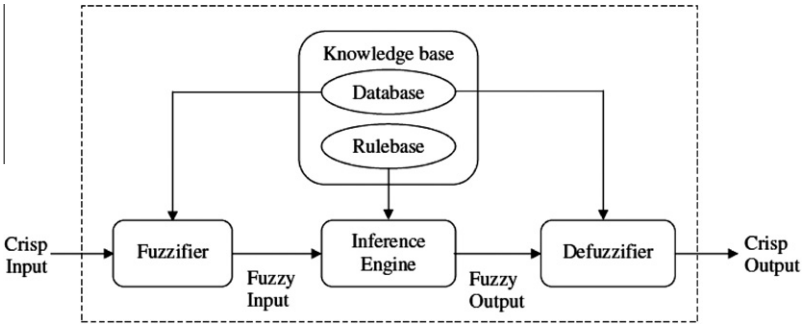

Figure 3 Fuzzy inference system.

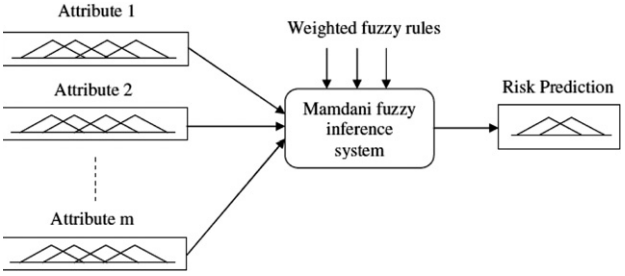

Figure 4 Designed fuzzy inference system based on weighted fuzzy rules.

label (risk level). Here, m-input, single-output of Mamdani fuzzy inference system with the 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 the fuzzification strategy. The fuzzy rule base contains a set of weighted fuzzy rules obtained from the proposed procedure discussed in Section 4.3.4 to learn the system. Then, for testing, the test data are given to the designed fuzzy system that predicts whether the patient has the heart disease or not.

5. Results and discussion

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.

5.1. Experimental environment and evaluation metrics

The proposed fuzzy logic-based clinical decision support system has been implemented using MATLAB (7.10). For experimentation, we have taken Cleveland, Hungarian and Switzerland heart disease datasets (http://archive.ics.uci.edu/ml/datasets.html) which are widely accepted databases obtained from the UCI machine learning repository. In the testing phase, the testing dataset is given to the proposed system to find the risk prediction of heart patients and obtained results are evaluated with the evaluation metrics namely, sensitivity, specificity and accuracy (Zhu et al., 2010).

Sensitivity, specificity and accuracy are the commonly used statistical measures to illustrate the medical diagnostic test and especially, used to enumerate how the test was good and consistent. Sensitivity evaluates the diagnostic test correctly at detecting a positive disease. Specificity measures how the proportion of patients without disease can be correctly ruled out. The association between both the sensitivity and specificity measures is defined by the graphical representation of the ROC curve and this helps to make a decision to find the optimal model to determine the best threshold for the diagnostic test. Accuracy can be concluded with the aid of the sensitivity and specificity measures in the presence of prevalence. Accuracy measures correctly figured out the diagnostic test by eliminating a given condition. 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 1.

$$\text{Sensitivity} = TP/(TP + FN)$$

Table 1 Terms used to define sensitivity, specificity and accuracy.

Outcome of the diagnostic test	Condition (e.g. disease) as determined by the Standard of Truth		
	Positive	Negative	Row total
Positive	TP	FP	TP + FP (total number of subjects with positive test)
Negative	FN	TN	FN + TN (total number of subjects with negative test)
Column total	TP + FN (total number of subjects with given condition)	FP + TN (total number of subjects without given condition)	N = TP + TN + FP + FN (Total number of subjects in study)

Table 2 Details of datasets.

	Total instance	Training data	Testing data
Cleveland	303	202	101
Hungarian	294	196	98
Switzerland	123	82	41

$$\text{Specificity} = TN/(TN + FP)$$

$$\text{Accuracy} = (TN + TP)/(TN + TP + FN + FP)$$

where TP is the True positive, TN is the True negative, FN is the False negative and FP is the False positive.

5.2. Experimentation

For experimentation, the heart disease datasets are divided into two sets such as: (1) training dataset and (2) testing dataset. Here, the missing values of a particular attribute are replaced by taking the average of the whole values in the same attribute. Along with this, we modified the chosen UCI data into two class labels. In all three datasets, *value 0* specifies the *no presence* of heart disease (less than 50% diameter

Table 3 Selected attributes for fuzzy rule generation.

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

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

Datasets	Class	Metric	Proposed system	
			Training	Testing
Cleveland	< 50%	Accuracy	0.509901	0.623529
		Sensitivity	0.724771	0.765957
		Specificity	0.258065	0.447368
	> 50%	Accuracy	0.509901	0.623529
		Sensitivity	0.258065	0.447368
		Specificity	0.724771	0.765957
Hungarian	< 50%	Accuracy	0.715054	0.469388
		Sensitivity	0.8	0.31746
		Specificity	0.540984	0.742857
	> 50%	Accuracy	0.715054	0.469388
		Sensitivity	0.540984	0.742857
		Specificity	0.8	0.31746
Switzerland	< 50%	Accuracy	0.364706	0.512195
		Sensitivity	0.625	0.333333
		Specificity	0.337662	0.526316
	> 50%	Accuracy	0.364706	0.512195
		Sensitivity	0.337662	0.526316
		Specificity	0.625	0.333333

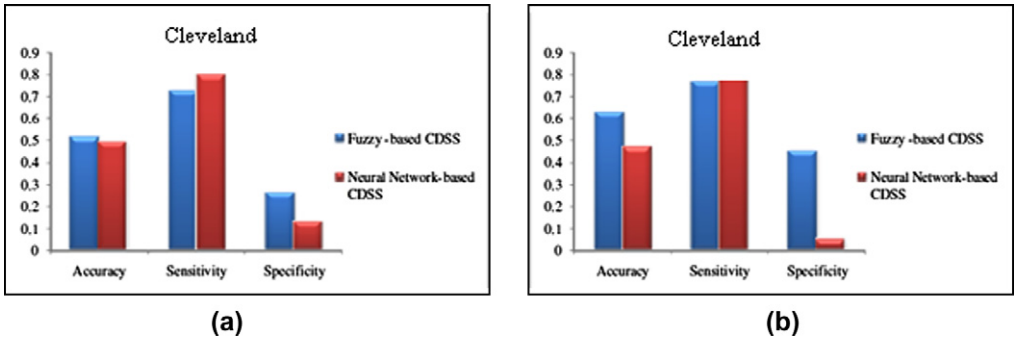

Figure 5 Risk prediction for the patients who come under below 50%: (a) training dataset and (b) testing dataset.

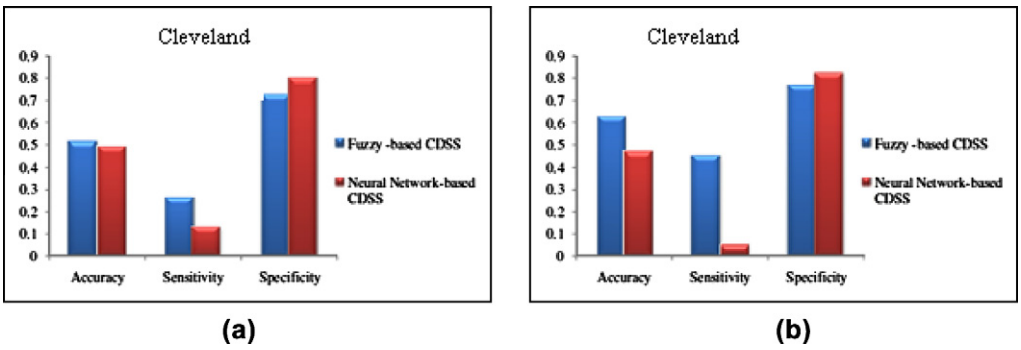

Figure 6 Risk prediction for the patients who come under above 50%: (a) training dataset and (b) testing dataset.

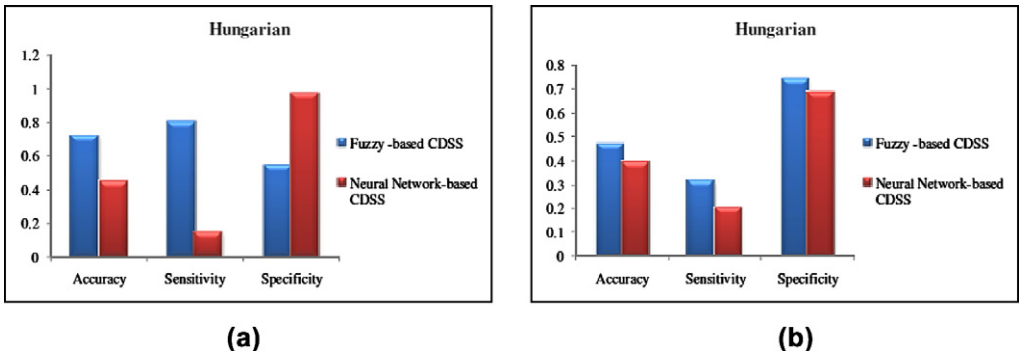

Figure 7 Risk prediction for the patients who come under below 50%: (a) training dataset and (b) testing dataset.

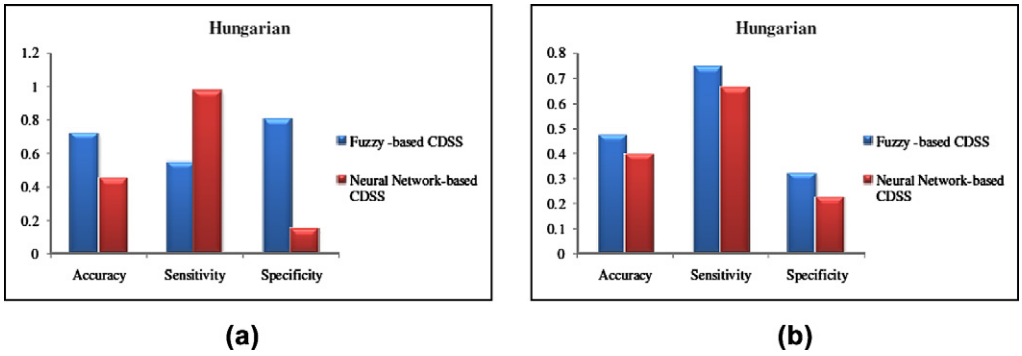

Figure 8 Risk prediction for the patients who come under above 50%: (a) training dataset and (b) testing dataset.

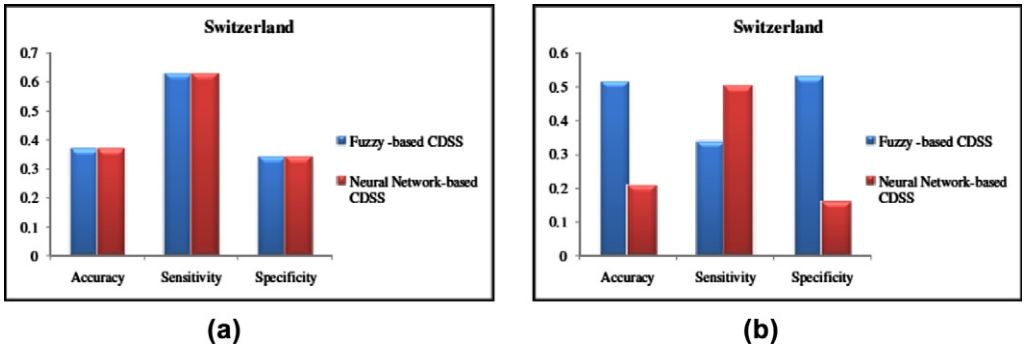

Figure 9 Risk prediction for the patients who come under below 50%: (a) training dataset and (b) testing dataset.

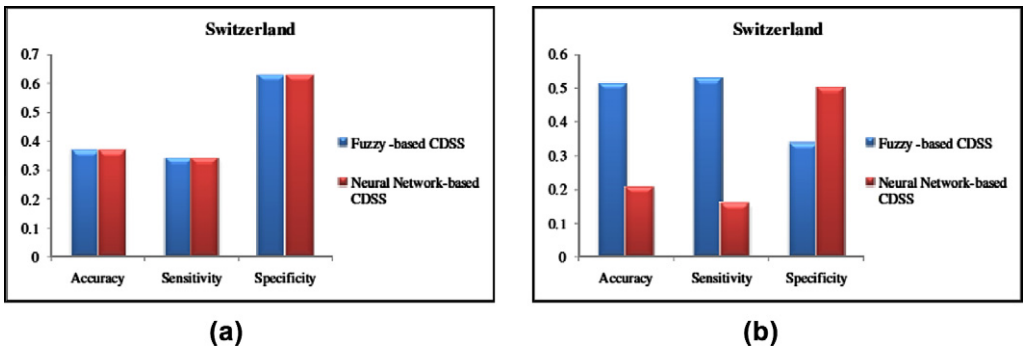

Figure 10 Risk prediction for the patients who come under above 50%: (a) training dataset and (b) testing dataset.

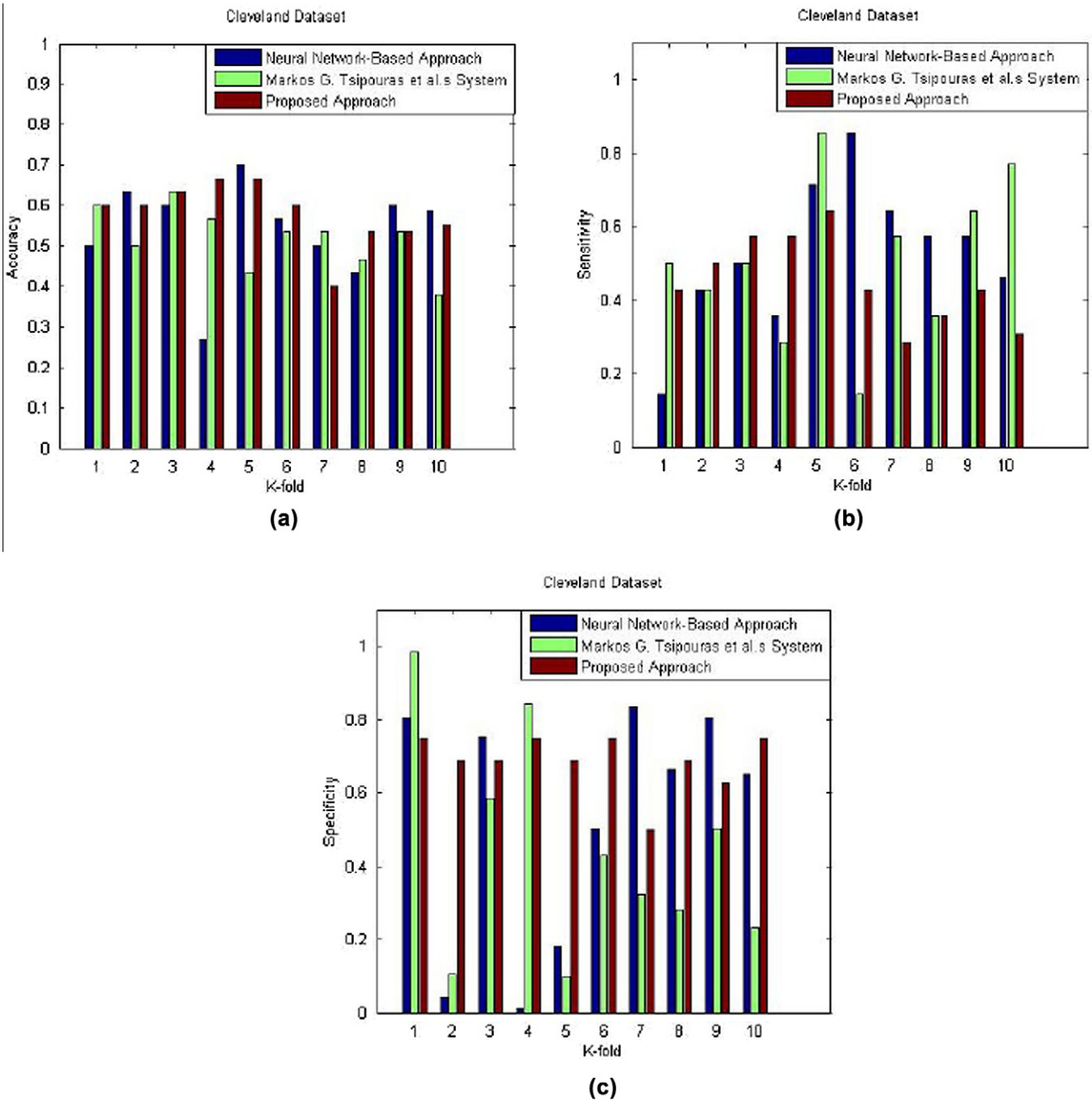

Figure 11 Analysis of Cleveland dataset: (a) accuracy graph, (b) sensitivity graph and (c) specificity graph.

narrowing) and*values 1–4* specifies the presence of heart disease (greater the 50% diameter narrowing). According to this, we have transformed it into two class data, where class 0 specifies the *no presence* of heart disease and class1 specifies the presence of heart disease. 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 2 provides the description of these datasets.

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 3.

Then, the fuzzy rules are generated from the selected attributes using the proposed system. Some of the fuzzy rules obtained are, "IF (resting blood pressure is H) then (Class is Class1) (0.20388)", "IF (Age is VL) then (Class is Class0) (0.12621)", "IF (fasting blood sugar is M) then (Class is Class1) (0.64078)", and "IF (exercise induced angina is VL) then (Class is Class1) (0.15534)". In the above rules, floating values specify the weight of the fuzzy rules obtained. Subsequently, the testing procedure is carried out to predict the risk level of the patient. 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 Table 4. Table 4 shows the overall

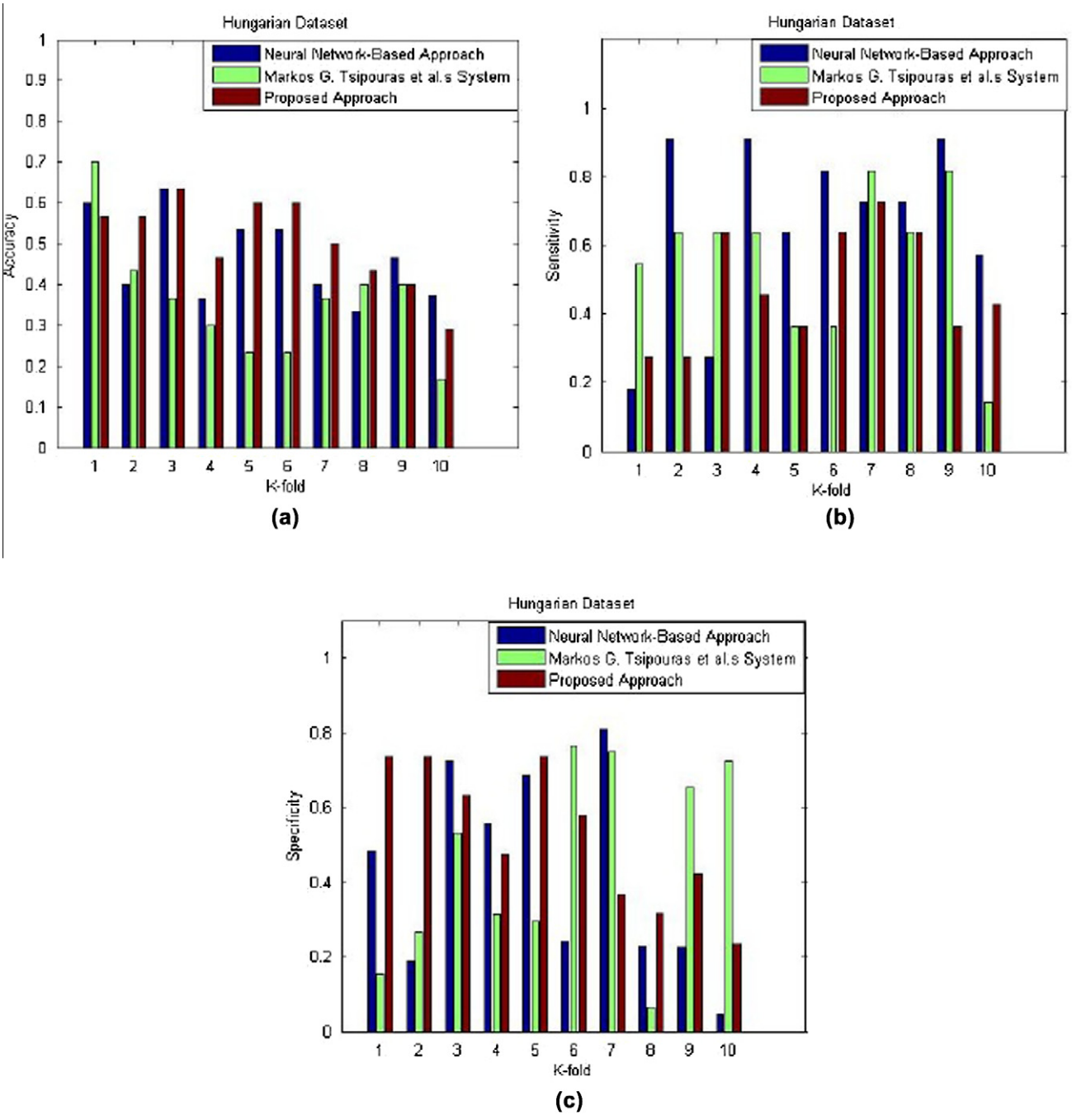

Figure 12 Analysis of Hungarian dataset: (a) accuracy graph, (b) sensitivity graph and (c) specificity graph.

performance of the proposed system in risk prediction, in which, class 0 indicates that the risk level is below 50% and class 1 indicates the risk level is above 50%.

5.3. Performance analysis

The performance of the proposed clinical decision support system was analyzed with three different heart disease datasets. Here, we compare the performance of the proposed clinical decision support system in risk prediction with the neural network-based system using these datasets. According to this, we have used a standard feed-forward neural network that was trained by the back-propagation algorithm. The results obtained for these three datasets are shown in Figs. 5–10. 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.

5.4. Comparative analysis of the proposed system

This section presents the performance analysis of the proposed system using k-fold cross-validation method. According to this, the original data are divided into k subsets of data and

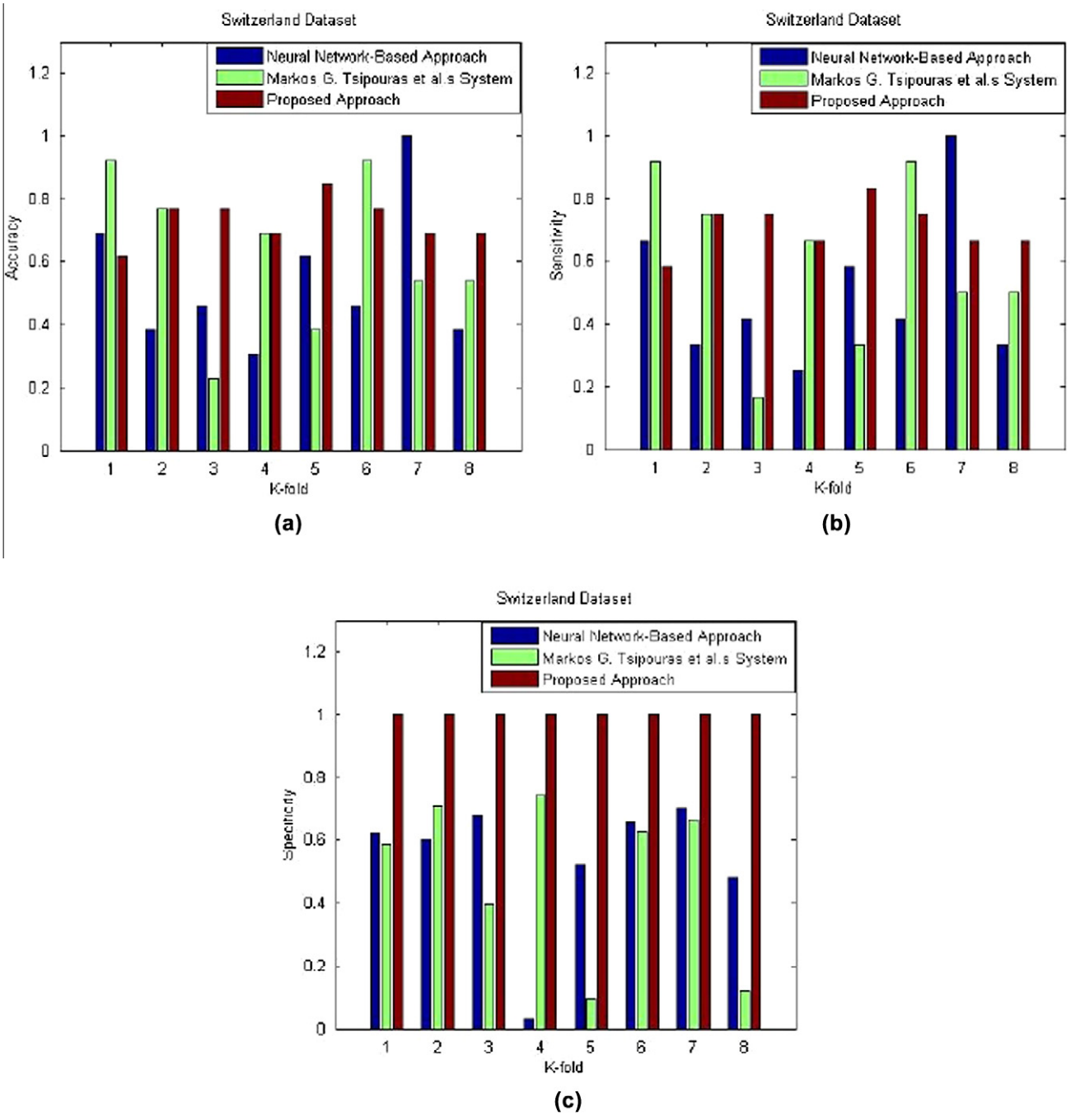

Figure 13 Analysis of Switzerland dataset: (a) accuracy graph, (b) sensitivity graph and (c) specificity graph.

for every validation, a single subset is used as the testing data and the remaining subsets are utilized as training data. This procedure is repeated until all the subsets of data are utilized as testing data. Here, we have chosen $k = 10$ (for Cleveland and Hungarian datasets) and $k = 8$ (for Switzerland dataset) so that, the input data is divided into 'k' sub-samples to extensively analyze the proposed system. Along with this, the comparative analysis is also performed to find the efficiency of the proposed system. For comparison, we make use of the system proposed by Tsipouras et al. (2008), who developed a fuzzy rule-based decision support system (DSS) for the diagnosis of coronary artery disease. In the previous system (Tsipouras et al., 2008), the rules are extracted from the decision tree after inputting the input data. Then, the extracted crisp rules are transformed into fuzzy rules that are given to the fuzzy model after optimizing the parameters. In addition to this, the results are also analyzed with the help of standard feed-forward neural network that was trained using the back-propagation algorithm. Hence, an extensive analysis is carried out using the 10-fold cross validation method to compare the efficiency of the three CDSSs used.

Then, three datasets are given to the proposed system, Tsipouras et al.'s (2008) system and neural-network-based system for predicting the risk level of the heart disease. Here, the prediction results are analyzed with the k-fold cross validation method and their corresponding sensitivity, specificity and accuracy values are computed. The values are then plotted as graphs, in which Fig. 11 is obtained for the Cleveland data. When analyzing these graphs, the proposed system outperformed other two methods in most of the sub-samples. For the sensitivity computation, the neural network-based system achieved 52.473% compared with the sensitivity of our system (45.22%). But, the overall specificity of the proposed system is 68.75% compared with the other systems' specificity (52.461% and 43.896%). In overall accuracy, the proposed system achieved 57.851%, while the accuracy of the network-based system and the previous method are 53.862% and 51.793% respectively.

Similarly, in Hungarian datasets, the proposed system produced higher accuracy (50.583%) with regard to the graph shown in Fig. 12. Here, neural-network-based approach achieved 46.417% and the fuzzy-based system achieved only 36%. When analyzing the Switzerland data (graph shown in Fig. 13), the proposed system outperformed in sensitivity, specificity and accuracy computation compared with the existing systems. The overall sensitivity of the proposed system is improved by 10% higher than the fuzzy-based system and 20% higher than the neural network-based system. In addition, the specificity of our proposed system is increased nearly by 50% higher than the previous systems. Hence, the overall accuracy of the proposed system is improved by 10% than the fuzzy-based system and 20% than the neural network-based approach.

6. Conclusion

We have presented a weighted fuzzy rule-based clinical decision support system (CDSS) for computer-aided diagnosis of the 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 the 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 attributes were generated after applying the mining procedure and these attributes were used to generate the fuzzy rules 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 improved significantly compared with the network-based system in terms of accuracy, sensitivity and specificity.

References

Abbasi, M.M., Kashiyarndi, S., 2006. Clinical decision support systems: a discussion on different methodologies used in health care.

Abidin, B., Dom, R.M., Rahman, A.R.A., Bakar, R.A., 2009. Use of fuzzy neural network to predict coronary heart disease in a Malaysian sample. In: Proceedings of the 8th WSEAS International Conference on Telecommunications and Informatics, Istanbul, Turkey, May 30–June 1.

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, pp. 207–216.

Ali, S., Chia, P., Ong, K., 1999. Graphical knowledge-based protocols for chest pain management. In: Proceedings of the Conference on Computers in Cardiology, Hannover, Germany, pp. 309–312.

Anderson, J., 1997. Clearing the way for physicians' use of clinical information systems. Communication of the ACM, 83–90.

Bradley, A.P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition 30 (7), 1145–1159.

Chen, J., Greiner, R., 1999. Comparing Bayesian network classifiers. In: Proceedings of UAI-99, Stockholm, Sweden, pp. 101–108.

De, S.K., Biswas, A., Roy, R., 2001. An application of intuitionistic fuzzy sets in medical diagnosis. Fuzzy Sets and Systems 117, 209–213.

Ephzibah, E.P., 2010. Cost effective approach on feature selection using genetic algorithms and LS-SVM classifier. IJ CA Special Issue on Evolutionary Computation for Optimization Techniques, EC OT.55, pp. 55–56.

Fidele, B., Cheeneebash, J., Gopaul, A., Goorah, S.S.D., 2009. Artificial neural network as a clinical decision-supporting tool to predict cardiovascular disease. Trends in Applied Sciences Research 4 (1), 36–46.

Garg, A.X., Adhikari, N.K., McDonald, H., Rosas-Arellano, M.P., Devereaux, P.J., Beyene, J., Sam, J., Haynes, R.B., 2005. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. Journal of the American Medical Association, PubMed 293 (10), 1223–1238.

Han, J., Pei, J., Yin, Y., Mao, R., 2004. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Mining and Knowledge Discovery 8 (1), 53–87.

Jilani, T.A., Yasin, H., Yasin, M., Ardil, C., 2009. Acute coronary syndrome prediction using data mining techniques – an application. World Academy of Science, Engineering and Technology, p. 59.

Kawamoto, K., Houlihan, C.A., Balas, E.A., Lobach, D.F., 2005. Improving clinical practice using clinical decision support systems:

a systematic review of trials to identify features critical to success. BMJ 330, 765.
Khatibi, V., Montazer, G.A., 2010. A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. Expert Systems with Applications 37 (12), 8536–8542.
Knopp, R.H., 2002. Risk factors for coronary artery disease in women. The American Journal of Cardiology 89 (12A), 28–34.
Kong, G., Xu, D.L., Yang, J.B., 2008. Clinical decision support systems: a review on knowledge representation and inference under uncertainties. International Journal of Computational Intelligence Systems 1 (2), 159–167.
Lin, L., Hu, P.J.H., Sheng, O.R.L., 2006. Decision Support Systems 42.
Merijohn, G.K., Bader, J.D., Frantsve-Hawley, J., Aravamudhan, K., 2008. Clinical decision support chairside tools for evidence-based dental practice. The Journal of Evidence-Based Dental Practice 8 (3), 119–132.
Miller, R.A., 1990. Why the standard view is standard: people, not machines, understand patients' problems. The Journal of Medicine and Philosophy 15 (6), 581–591.
Miller, R.A., Masarie, F.E., 1990. The demise of the ''Greek Oracle'' model for medical diagnostic systems. Methods of Information in Medicine 29 (1), 1–2.
Musen, M., Shahar, Y., Shortliffe, E.H., 2001. Clinical decision support systems. In: Shortliffe, E.H., Perrault, L., Wiederhold, G., et al. (Eds.), Medical Informatics: Computer Applications in Health Care and Biomedicine, Springer-Verlag, New York, pp. 573–609.
Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J., 1998. UCI repository of machine learning databases. Department of Information and Computer Science, University California Irvine.
Osheroff, J.A., 2009. Improving Medication Use and Outcomes with Clinical Decision Support: A Step-by-step Guide. The Healthcare Information and Management Systems Society, Chicago, IL.
Palaniappan, S., Awang, R., 2008. Intelligent heart disease prediction system using data mining techniques. International Journal of Computer Science and Network Security 8 (8), 108–115.
Parthiban, L., Subramanian, R., 2008. Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences 3, 3.
Patil, S.B., Kumaraswamy, Y.S., 2009. Intelligent and effective heart attack prediction system using data mining and artificial neural network. European Journal of Scientific Research 31 (4), 642–656.
Setiawan, N.A., Venkatachalam, P.A., Hani, A.F.M., 2009. Diagnosis of coronary artery disease using artificial intelligence based decision support system. In: Proceedings of the International Conference on Man-Machine Systems, Batu Ferringhi, Penang, 11–13 October, 2009.
Shanthi, D., Sahoo, G., Saravanan, N., 2008. Input feature selection using hybrid neuro-genetic approach in the diagnosis of stroke disease. International Journal of Computer Science and Network Security 8, 12.
Straszecka, E., 2007. Combining uncertainty and imprecision in models of medical diagnosis. Information Sciences 176, 3026–3059.
Subbalakshmi, G., Ramesh, K., Chinna Rao, M., 2011. Decision support in heart disease prediction system using naive Bayes. Indian Journal of Computer Science and Engineering (IJCSE) 2 (2), 170–176.
Szolovit, P., 1995. Methods of Information in Medicine 34, 111.
Tang, T.I., Zheng, G., Huang, Y., Shu, G., Wang, P., 2005. A comparative study of medical data classification methods based on decision tree and system reconstruction analysis. IEMS 4 (1), 102–108.
Thuraisingham, B., 2000. A primer for understanding and applying data mining, IT professional. IEEE Computer Society, pp. 28–31.
Tsipouras, M.G., Exarchos, T.P., Fotiadis, D.I., Kotsia, A.P., Vakalis, K.V., Naka, K.K., Michalis, L.K., 2008. Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Transactions on Information Technology in Biomedicine 12 (4), 447–458.
Warren, J., Beliakov, G., Zwaag, B., 2000. Fuzzy logic in clinical practice decision support system. In: Proceedings of the 33rd Hawaii International Conference on System Sciences, Maui, Hawaii, 4–7 January 2000.
Yan, H., Zheng, J., Jiang, Y., Peng, C., Li, Q., 2003. Development of a decision support system for heart disease diagnosis using multilayer perceptron. In: Proceedings of the 2003 International Symposium on Circuits and Systems, May 25–28, pp. 5, 709–712.
Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 338–353.
Zhu, W., Zeng, N., Wang, N., 2010. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical SAS® implementations. In: NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland.

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 has been working as a senior teacher and lecturer in computer science in various institutions from year 2000. Since 2009, he is working as the registrar and lecturer in the Department of Information Technology, Al Musanna College of Technology, Directorate General of Technological Education, Ministry of Manpower, Oman. He is currently pursuing his Ph.D. degree working closely with Dr. V. Khanaa, IT Dean, Bharath University, Chennai, and Dr. T. Jebarajan, Department of Computer Science Engineering, Karpaga Vinayaga College of Engineering, Anna University, Chennai.