An Enhanced Risk Prediction System for Cardiovascular Disease in India using Fuzzy Classification

Maddula Krishnam Raju
Department of Computer Sciene Engineering,
Vizag Institute of Technology,
Visakhapatnam, Andhra Pradesh, India

S.K. Shafiulilah
Department of Computer Sciene Engineering,
Vizag Institute of Technology,
Visakhapatnam, Andhra Pradesh, India

ABSTRACT
In recent times most of the people are very cautious about their health. Information and Communication Technology (ICT) is playing an extensive role in the development of medical domain application, which has been one of the most active research areas for the scientists and practitioners. Heart disease rates among the major cause of mortality in developing countries like India. Fuzzy System is a one of the promising method to provide insights for better evaluation of cardiovascular diseases. Over the years many developers built Knowledgebase Decision Support System (KDSS) on diseases, based on the medical expert's information. This process is time consuming and KDSS depends on medical expert's opinion which may be subjective. To overcome this problem, machine learning techniques have been developed to gain knowledge automatically from past information and raw data. This work propose a novel Fuzzy rule based Disease Decision Support System (FDDSS) for the diagnosis of heart disease, based on the score obtained, the knowledge automatically from the patient's clinical data. The FDDSS predicts the Cardiovascular Diseases which was considered as the most important cause of mortality in India by the year 2015. The experimentation result proves that, the proposed FDDSS is better than existing manual and automation methods. A few FDDSS experimental outcomes will be helpful to the physicians to take accurate decisions.

General Terms
Text Mining, Machine Learing, Clustering and Classification.

Keywords
Fuzzy logic, Cardiovascular Disease, Coronary Heart Disease, Knowledgebase Decision Support System (KDSS), Fuzzy rule based Disease Decision Support System (FDDSS).

1. INTRODUCTION
To retrieve hidden patterns and relationships from large sets, Data mining combines statistical analysis, artificial intelligence, machine learning and database technologies [1][4]. From last decade of clinical services, including prediction of effectiveness of surgical process, clinical testing, and the discovery of relationship between clinical and diagnosis data, Mining techniques have been applied [5]. Now a day's clinical diagnosis is a very critical process, entailing precise patient data, a philosophical understanding of the clinical literature and many years of medical experience [2]. The biggest challenge of healthcare data is mining form hidden information for exact decision makings are collected in a huge amount by the health care industry [2]. A Clustering Analysis for Heart Failure Alert System using RFID and GPS is a good technique for saving of heart patients through the mechanism of Global Positioning System [6]. Medical

decisions are taken perception and seniority rather than on the, knowledge data hidden in the database [3]. Unfortunately every doctor is not evenly expert in every sub specialty and they are in several places a scarce resource. Clinical care and reducing costs can be enhanced by a medical diagnosis system [3]. Medical Decision Support Systems are widely classified into two major parts namely one is Knowledge based and another Non-knowledge based [2]. The Knowledge based system consists rules, maximum in the form of "if "and "then" statements. The adaptive rules from a knowledge base server provide efficiency than others in certain case, i.e., chest pain management [7].

Cardiovascular Disease is one of the leading diseases in India and worldwide. It affects people at younger ages in low and middle income countries, compared with high income countries, thereby having a greater economic impact on low and middle income countries. India is the second most populous country in the world with an estimated population of over 1 billion [8]. An automated advanced clustering algorithm for text classification is a good mechanism for how to classify the data into different clusters [15]. Concurrently; India has also seen an exponential rise in prevalence of Heart disease [3]. WHO has declared India the CAD capital of the world [4] . Data from the registrar general of India shows that heart disease are major cause of death in India, studies to determine the precise cause of death in A.P have revealed that CVD cause about 30% of deaths in rural areas [5]. Hence there is a need for decision support system for predicting heart disease. The objective of this work is to build a model for heart disease prediction which helps the physicians to make accurate diagnosis.

The paper is organized as follows: In the next section, some related concepts about fuzzy rules, principle component analysis and heart disease are presented. In section 3 describes overview of work related to fuzzy classification and heart disease datasets. In section 4 proposed algorithm FDDSS is presented. Performance results are presented in section 5 and conclusions are given in section6.

2. RELATED WORKS
About 25 % of deaths in the age group of 25-69 years occur because of heart diseases. In urban areas, 32.8% deaths occur because of heart ailments, while this percentage in rural areas is 22.9 [9]. If all age groups are included, heart diseases account for about 19% of all deaths. It is the leading cause of death among males as well as females. It is also the leading cause of death in all regions though the numbers vary. The proportion of deaths caused by heart disease is the highest in south India (25%) and lowest 12% in the central region. What the result is quite provocative, if you look at rural areas of poorer states like Uttar Pradesh and Bihar, the leading cause

of death among middle aged males is cardiovascular disease. The data forms part of the first set of results from the study, dubbed the 'Million Deaths Study' because it aims to investigate one million deaths by 2014. The study is based on the existing system of tracking the health status of 6 million people across 1.1 million households in all the states and union territories through the units of Sample Registration System (SRS) which the RGI uses to track birth and mortality data in the country [10]. India is accounted for by non communicable diseases, while communicable diseases and those associated with child birth and nutritional disorders account for 38 % of deaths. Other causes of deaths in the 25-69 years age group - urban and rural areas taken together - are respiratory diseases such as asthma(10.2%), tuberculosis (10.1 %), malignant tumours (9%), digestive diseases (5.1%) and diarrheal disease (5%).Malaria, which had been a leading cause of death, now accounts for only 2.8% deaths.

About 9.5 million deaths, which is about one in six deaths worldwide, occur in the country every year. Over three-quarters of these deaths take place in the home and more than half of these do not have a certified cause. The law on 'Medically Certified Causes of Death' has not been effective in providing any useful inputs as just about 0.4 million deaths are registered under this and that too in some cities [11]. Earlier, only smaller studies conducted in villages in coastal Andhra Pradesh had indicated that cardiovascular diseases had replaced communicable diseases as the top killer in rural areas [12]. But, no large- scale studies have been done so far. Unlike deaths from communicable diseases or injuries having known causes, most non- communicable diseases can have multiple causes. For example, heart attack could be caused by smoking, high blood pressure, high cholesterol and other factors. For cardiovascular diseases in rural areas are nothing so far. Regarding acute heart attacks - not chronic heart disease - that are a major cause of death in urban India, it is mainly killing young people in their productive years. Large percentage of these heart attacks are attributed to smoking [13].

A new study has indicated that older people at age 60 can lower the risk for heart attack and stroke within the first five years if they quit smoking in Washington. The study conducted at German Cancer Research Centre analyzed 25 individual studies, compiling data from over half a million individuals age 60 and older. According to the study, twice as many smokers die from cardiovascular disease than life-long non-smokers do [15][17]. The increase in risk depends on the number of cigarettes that a person has smoked in his or her lifetime. After one quits smoking, this risk continues to decrease. On average, the risk for former smokers is only 1.3 times that of people who have never smoked in their lives [14].

The literature presents in Medical Decision Support Systems describes the number of researchers that have made use of Machine Learning, Neural Networks and Data Mining techniques [16]. Among those, to support decision makers in risk prediction of heart disease, the helpful of research have been presented.

From Dempster-Shafer theory of evidence and fuzzy sets theory, Vahid Khatibi and Gholam Ali Montazer [1] have been proposed an inference engine named fuzzy-evidential hybrid inference engine mechanism. This mechanism has two phases. In the first phase, through fuzzy sets, it applies the fuzzy inference rules on the acquired fuzzy sets to generate the first phase results. In the second phase is assumption of propositions and functions. Finally it has yielded 92%

accuracy rate for its accurate prediction by applying the inference engine on the Coronary Heart Disease (CHD) risk assessment [17].

3. DISCUSSION OF HEART DISEASE DATASET

The heart is one of the most important and sensitive organs in human body. Heart disease refers to various problems that distress the heart and the blood vessels in the heart. Angina, Cardiovascular Disease, Heart attack and Heart failure are some examples for the different types of heart diseases Cardiovascular Disease is an important reason of sickness and death in now a day. The term 'cardiovascular disease' that represents a category of heart disease has a broad variety of conditions that struggles the heart, the blood vessels and which way the blood is pumped and circulated in the body [4]. Coronary heart disease is caused by the decreased blood levels and oxygen supply to the heart due to the narrowing of the coronary arteries. And it includes myocardial infarctions, commonly called as a heart attacks, and angina pectoris, or chest pain [7]. Heart attack results due to the abrupt blockage of a coronary artery, usually because of blood clots. Insufficient blood flow to the heart muscles results in chest pains [3]. There are several types of cardiovascular disease such as high blood pressure, coronary artery disease, stroke, or rheumatic fever/rheumatic heart disease [18].

According to world health statistics report, compiled by WHO, mortality due to cardiac causes has overtaken, mortality due to all cancers put together. In India alone 4,280 sudden cardiac deaths per lakh deaths annually. By the year 2030, India will rank among the highest risk for heart disease [19]. Cardiovascular Disease is a class of diseases that involve the heart or blood vessels. It includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack) [20]. Other CVDs are stroke, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, atrial fibrillation, congenital heart disease, endocarditic, aortic aneurysms, and peripheral artery disease [12].

3.1 Risk factors of Cardiovascular Disease

It is estimated that 90% of CVD is preventable, the prevention of CVD in women less than 55 and men less than 45 years old. However, in those who are older it is recommends in some individuals.

3.2 Prevention of Cardiovascular Disease by using FDDSS

Currently practiced measures to prevent cardiovascular disease include:

1. A low-fat, high-fiber diet including whole grains and fruit and vegetables.

2. Tobacco cessation and avoidance of second-hand smoke.

3. Limit alcohol consumption to the recommended daily limits, consumption of 1–2 standard alcoholic drinks per day may reduce risk by 30%. However, excessive alcohol intake increases the risk of cardiovascular disease.

4. Lower blood pressures, if elevated.

5. Decrease body fat if overweight or obese.

6. Increase daily activity to 30 minutes of vigorous exercise per day at least five times per week.

7. Reduce sugar consumptions, decrease psychosocial stress.

3.3 Heart disease in Andhra Pradesh

Several studies show a high prevalence of diabetes and other risk factors for heart disease in Andhra Pradesh. The prevalence of risk factors was shown in Table 2.sudden cardiac death contributed to 10% of overall mortality in Andhra Pradesh. Study in rural Andhra Pradesh showed the CVD was the leading cause of mortality accounting for 32% of all deaths [21].

3.4 Heart disease in India

CHD prevalence appears to be worsening in India. Prevalence of CHD will rise in India compared to china and established market economies from the year 1990-2020[4]. Leeder et al [5] estimated total years of life lost due to CVD among Indian men and women aged 35-64 to be higher than the countries china and Brazil as shown in figure 6.

4. AN EFFICIENT FDDSS TO RISK LEVEL PREDECTION OF HEART DISEASE

Disease decision making is the intention behind the design of disease decision support system and it specifies patient-specific preventions or recommendations produced using characteristics of individual patients [20]. In early days, clinical decision support system based on computer diagnosis methodologies has been proposed in the survey. That was evaluated by specific methods or other lab sources like laboratory examinations, demographic and history data.

4.1 Pre-processing

The aim of data pre-processing is to extract useful information from raw heart disease datasets and such type of data should be converted into specific format i.e., prediction of risk level of disease. In disease datasets, abundant information or irrelevant information is there, so retrieving of exact information is biggest task in the data mining. Hence in data pre-processing phase, raw data need to be cleaned, analyzed and transformed for further usage.

4.2 Classification of training dataset based on risk level

After completion of pre processing, take the training dataset as input used for prediction of disease that is classified into two subsets of data based on the class labels. The input training dataset consists of two class labels, where '0' indicates that the status of disease is less than 50% and '1' represent that the risk level for heart disease is more than 50%. By using these two class labels, the dataset D is divided into two subsets of data, D= {D_p; 0≤ p≤ 1}, where, p denotes the class label that describes the risk level of patients p. In addition to, each class contains n number of attributes and each attribute presented in the dataset consists of attribute category that is the continuous symbolic value specified for every patients. The two subsets of data D 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 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, it is using the frequent attribute category that is mined from the input datasets. Then, based on the frequency of attribute category and the weight age of attributes, the fuzzy rules are generated automatically [22][23]. The steps to be used for automatic generation of fuzzy rules are discussed in this sub-section.
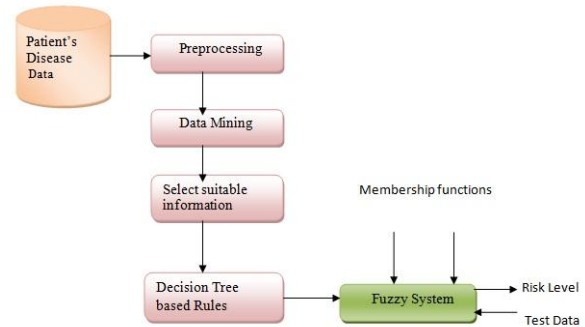

Figure 1: Fuzzy-based Disease Decision Support System

4.4 Developing disease decision support system using fuzzy logic

Every human in the Universe of discourse and assign a degree of membership in the fuzzy subset RISK with the help of a membership function. This membership function is depending on several risk factors. An example can be described as: With the correct Systolic Blood Pressure =120 and diastolic blood pressure =80, cholesterol HDL > 45 and LDL < 100, Sugar Fasting <=100, Sugar PP <=120, HBA1C<= 6, Age < 38 = Young, Age >= 90 as old, chest pain < Typical Angina, gender, heartbeat, ECG, exercise. However to illustrate it will take membership function for age [21].

Age(x) is defined and membership function can be calculated as

Age(x) = {0 if Age(x) <= 39,

1 if Age(x) >=90}

Membership function may be estimated by Blood pressure, Blood Sugar, weight, cholesterol, etc. Each input consist a membership function and fuzzy variable. The strength of the fuzzy rule can be calculated based on the membership function of fuzzy variable. The condition of the heart disease is calculated by taking maximum number of selected outputs and concludes the final result. Here fuzzy reasoning is an approximate reasoning and it is the process of drawing conclusion from fuzzy rules and fuzzy sets. Fuzzy inference is an inference based on fuzzy rules, fuzzy sets and fuzzy reasoning. The components of FDDSS have rule, inference, fuzzifier and defuzzifier. Fuzzifier creates fuzzy sets; inference engine applies fuzzy rules to fuzzy sets to find the resultant fuzzy output and defuzzification is a process to converts output value into threshold value. Suppose to find picture of a person, Fuzzy rules are invoked. These rules have been generated and with incorporation of these rules computers can understands and respond to normal human Language.

4.5 The FDDSS algorithm has following steps

Step1. Medical data set is assumed to be available before the clustering analysis is carried out.

Step2. Signifying Ijk which is the influence of point 'j' on cluster 'k', the cluster centre and covariance matrix are calculated.

Step3. Let 'f' is a parameter which has been defined by the user, this parameter is called fuzzifier.

Step4. To find out all weighted combinations on cluster 'k' i.e., Wk

$$W_k = \sum_{j=1}^{N} I_{jk}^{f} Y_j$$

Where 'Wk' refers cluster weights, 'I' refers clustered data sets, 'f' is a fuzzier and 'Yj' denotes remaining un clustered data.

Step5. The fourth step is updating, if the comparison between cluster centers and data then the cluster parameter is alternated with the update of the coefficients, which are weighing coefficients till a convergence criterion is met. It is calculated by the following formula.

$$Djk = (Yj - Wk) (Yj - Wk) T$$

$$S_k = \sum_{j=1}^{n} I_{jk}^{f} \left(Y_j - W_k \right) \left(Y_j - W_k \right)^{T}$$

$$S_k = \sum_{j=1}^{n} I_{jk}^{f} D_{jk}$$

Where Djk distance between two clusters, Sk is prediction score of the disease[25].

Step6. Find the accuracy of the data set. Accuracy= no. of correctly predicted test data/ Total number of test data.

The algorithm has improved10.8% accuracy against J4.8 and 19.8%improvement over naïve bayes for non medical data sets. This approach reached10.26% improvement over J4.8 and8.6% improvement against naïve bayes respectively for heart disease data set.

5. RESULTS AND DISCUSSION

The proposed FDDSS algorithm was tested on different data sets from Effects of computerized clinical decision support systems on practitioner performance and patient outcomes UCI repository [6] and one real life data set Andhra Pradesh heart disease data set. A brief description about the data set is presented in figure2; it shows total cause of deaths in India from age between 25 to 39. Figure3 shows risk prediction for the patient's all ages in India. Figure4 shows Prevalence of total risk factors taking from Andhra Pradesh. Figure5 purely describes about Cardiovascular Disease and total years of life lost due CVD in 2000 and 2030 from India. And finally, figure6 shows comparison between three countries with total years of life lost due CVD in 2000 and 2030.

Figure2 describes comparison between different diseases, but Cardiovascular Disease is major cause of death in India. When compared to female candidates the male persons are very high in CVD and below age of 25 years those are getting a chance to 69% of heart disease. So, CVD plays a crucial role in heart diseases.

Figure3 shows risk prediction of heart disease in all ages, but not in particular age it compares all ages. Finally all ages consisting of Cardiovascular Disease but in particular age. Here also all age percentage of male candidates is very high compared to females due to the reason of heart disease in India.
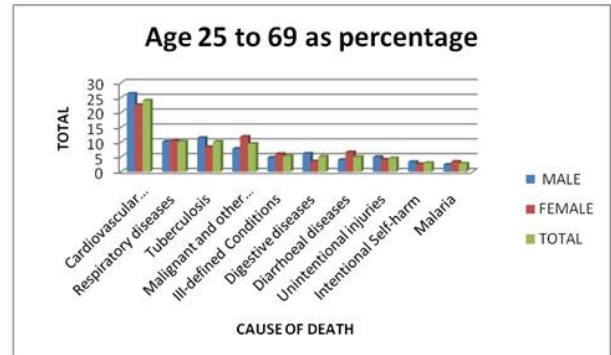

Figure 2: Risk prediction for the patient's age between 25-39 in India
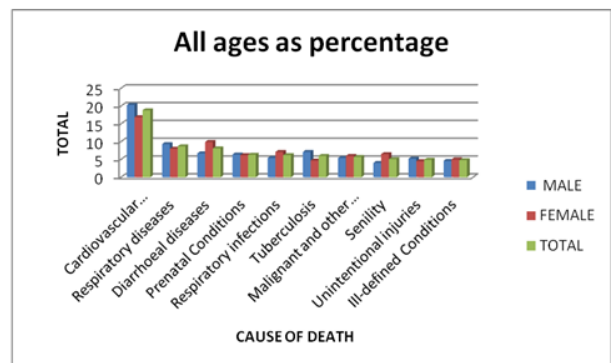

Figure 3: Risk prediction for the patient's all ages in India

Figure4 refers prevalence of risk factors in Andhra Pradesh; the biggest issue is Cholesterol problem in Andhra Pradesh. It is one of the reasons for Cardiovascular Disease. Majority of 60% peoples are struggled with Cholesterol in AP.
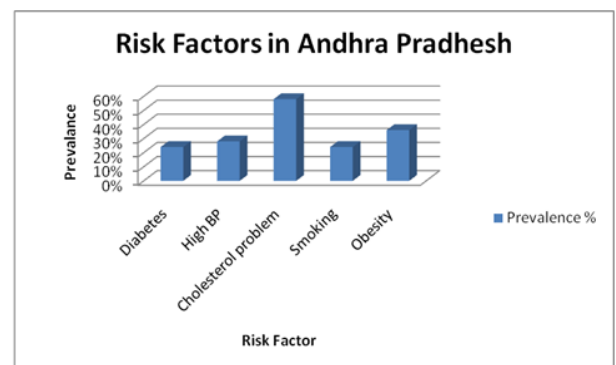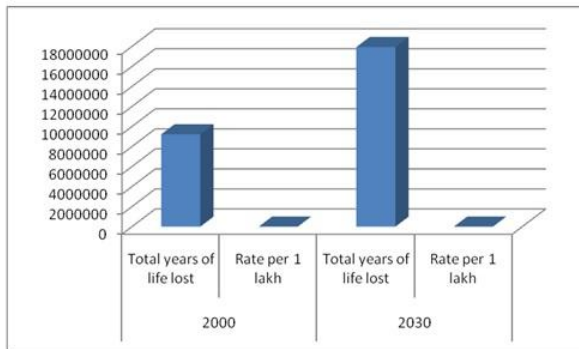

Figure 4: Prevalence of risk factors in Andhra Pradesh

The following result describes total years of life lost due CVD in 2000 and 2030 from India. In the year of 2000 eight chores of the peoples are suffered with CVD approximately that will be increase in the year of 2030 eighteen chores.

Figure 5: Total years of life lost due CVD in 2000 and 2030 from India

India, with 1,284,213,589 (1.28 billion) people is the second most populous country in the world, while China is on the top with over 1,371,453,688 (1.37 billion) people. But the following figure shows India is high effected cause of heart disease when compared to Chaina and Brazil. India represents almost 17.31% of the world's population as well as 14% of heart diseases will be there in the year of 2030[25].
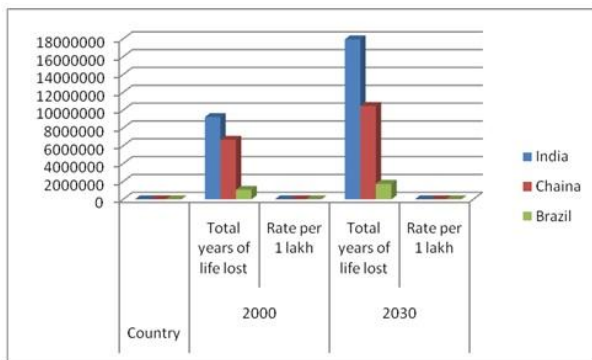
Figure 6: Comparison of Total years of life lost due CVD in 2000 and 2030

6. CONCLUSIONS

It is a weighted fuzzy rule based disease decision support system for computer-aided diagnosis of heart disease. The automatic procedure to generate the fuzzy rules by using proposed system FDDSS and the weighted procedure introduced in the proposed work is additional advantage for effective learning of the fuzzy system. FDDSS is as risk prediction of heart patients. Selection of suitable attributes were generated after applying the data 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 disease decision support system. Finally, the experimentation was carried out on the machine learning repository and the results in risk prediction ensured that the proposed disease decision support system was improved significantly compared with the network-based system in terms of accuracy. In this research paper we presented a fuzzy data mining approach for cardiovascular disease classification, it is one of the major heart diseases. The system is designed for Andhra Pradesh population. Andhra Pradesh is in risk of more death due to heart disease. Heart disease can be handled successfully if more research is encouraged to develop prediction system in this area.

The future enhancement is prevention at the community level is one such avenue to reduce the projected burden of CVD.

Community prevention efforts may include greater tobacco control, elimination of artificial trans fat, reducing dietary sodium intake, reducing air pollution, reducing obesity, and increasing physical activity with a focus on children and the design of new communities.

7. REFERENCES

[1] George K . Merijohn, James D. Bader, Julie Frantsve-Hawley, Krishna Aravamudhan "Clinical Decision Support Chairside Tools for Evidence- Based Dental Practice", The Journal of Evidence- Based Dental Practice , Elsevier, Vol:8, No:3, pp:119-132, 2008.

[2] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol 8, No: 8, pp: 108-115, 2008.

[3] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol: 3, No: 3, 2008.

[4] Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, pp: 28-31, 2000.

[5] Tzung-I Tang, Gang Zheng, Yalou Huang, Guangfu Shu, Pengtao Wang,"A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis", IEMS, Vol. 4, No. 1, pp. 102-108, 2005.

[6] Garg AX, Adhikari NKJ, McDonald H, et al., "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes", Journal of the American Medical Association, Vol: 293, No: 10, pp: 1223-38, 2005.

[7] Rao, Gudikandhula Narasimha, and P. Jagdeeswar Rao. "A Clustering Analysis for Heart Failure Alert System Using RFID and GPS." ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol I. Springer International Publishing, 2014.

[8] Kasabov N., Song Q. and Denfis, "Dynamic Evolving Neuro Fuzzy Inference system and its application for Time- Series Prediction" IEEE transactions on Fuzzy systems, Vol. 10, No. 2, 144-154, 2002.

[9] Jian Yu,Min-Shen Yang, "A Generalized Fuzzy Clustering Regularization Model with optimally tests and model complexity analysis" IEEE transactions on Fuzzy Systems, Vol 15,Issue 5, 904-915, 2007.

[10] Sato, M,Sato, Y, "Fuzzy clustering model for fuzzy data, Fuzzy systems", International Joint conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium Proceedings, Vol. 4, 2123-2128, 1995.

[11] Bezdek J C, "Fuzzy Mathematics in Pattern classification",Ph.D. thesis Centre for Applied Mathematics Cornell University Ithica N.Y. 1973.

[12] Kamber and Han, "Data Mining concepts and techniques", Elsevier, 2003.

[13] Fogler and Klir, "Fuzzy sets, Uncertainty and information", Prentice Hall,1988.

[14] de Carvalho Francisco de A.T., Camilo P. Tenorio, "Fuzzy K-means lustering algorithms for interval valued data based on adaptive quadratic distances" Fuzzy Sets and Systems Vol. 161, Issue 23, 2010.

[15] G. Narasimha Rao, et al. "An Automated Advanced Clustering Algorithm For Text Classification". In International Journal of Computer Science and Technology, vol 3,issue 2-4, June, 2012, eISSN : 0976 - 8491,pISSN : 2229 – 4333.

[16] Keith, Wai, Choi, "Mining Fuzzy states in a Donor Database for Direct Mining by a charitable organization" First IEEE international Conference on Cognitive Informatics Volume Issue9, 239-246, 2002.

[17] Zarandi M.H. Fazel, Razaee Zahara S., "A Fuzzy Clustering Model for Fuzzy Data with Outliers" International Journal of Fuzzy System Applications, Vol. 1, Issue 2, 2011.

[18] Xiang Li,Hau- Wong San and Wu Si, "A fuzzy minimax clustering model and its applications" Information Sciences: an International Journal, Vol. 186, Issue 1, 2012.

[19] D'Urso Pierpaaaolo, Paolo Giordani, "A weighted fuzzy c-means clustering model for fuzzy data" Computational Statistics & Data Analysis, Vol. 50, Issue 6, 2006.

[20] Pal K and Mitra P, "Data Mining in Soft Computing Framework: A Survey" IEEE transactions on neural Networks, vol 13,no 1 (Jan 2002).

[21] Cruse R. and Borgelt C. "Fuzzy Data Analysis Challenges and Perspective"

[22] Berkhin, P. "Survey of Clustering Data Mining Techniques".

[23] Halkidi M., "Quality assessment and uncertainty handling in data mining process".

[24] Chau M.,Cheng, R., and Kao. B.,"Uncertain Data Mining : A new Rsearch Direction".

[25] G. Narasimha Rao, et al. "Distributional features for text categorization based on score" In Proc. of ICCCSA-2012, MRCET, Hyderabad, pp:235-240, ISBN:978-81-921580-8-2

[26] http://www.indiaonlinepages.com/population/india-current-population.html