Knowledge Based MDSS using Data Mining

Taranath N.L., Shanthakumar B. Patil, Premajyothi Patil, C. K. Subbaraya

Abstract: Medical Decision Support System (MDSS) links the patient information to promising diagnostic and treatment paths. It can be built either as Knowledge-based system or Learning-based system. Knowledge-based systems are human-engineered maps from best medical practices and patient data will be recommended. Learning-based systems derive the mapping techniques from data mining, statistical approaches and machine learning techniques. An Integrated decision support system integrates both Knowledge-based and Learning-based systems to provide a robust solution to the information challenge in the presence of partial information. In this work, we design a framework and concrete implementation of Integrated Medical Decision Support System to assist the Doctors in clinical decisions regarding the prescription of drugs. It uses the Knowledge base for prescribing the drugs to the patients, however if the available data is partial it employs the machine learning techniques to answer the query. It is suitable for many different healthcare settings and many different users. The framework is query-based and it can be adapted for use with many different end-user interfaces.

Index Terms: Artificial Intelligence, Data Mining, Learning based Systems, Knowledge based Systems.

I. INTRODUCTION

Under normal circumstances reliable information exchange between distributed heterogeneous parties can be easily accomplished using existing techniques [4]. These techniques fail to be of practical use under adverse situations like time and information constrained setting where available patient detail is partial. In this work, a framework is designed for reliable information exchange between distributed heterogeneous parties, using ontological concepts, inference rules and machine learning techniques.

The main task is to design and implement an Integrated Medical Decision Support System, which helps the medical professionals to assist in clinical decisions regarding the prescription of medicine. An Integrated MDSS algorithm takes user queries for prescribing the medicine and uses its Knowledge base to give result to the query. If the available data is partial, it employs machine learning technique to impute missing data and gives the better result.

II. REVIEW OF LITERATURE

2.1. Knowledge based approach for Medical Decision Support Systems

Revised Version Manuscript Received on July 03, 2017.

Prof. Taranath N L, Research Scolar, Dept. of CS & E, VTU, NCET, Bengaluru, Karnataka, India, E-mail: taranath.taras@gmail.com

Dr. Shantakumar B Patil, Professor and Head, Dept. of CS & E, NCET, Bengaluru, Karnataka, India, E-mail: shantakumar.p@gmail.com

Dr. Premajyothi Patil, Professor, Dept. of CS & E, NCET, Bengaluru, Karnataka, India, E-mail: shantakumar.p@gmail.com

Dr. C.K. Subbaraya, Principal & Professor, Dept. of CS & E, AIT, Chikkamagaluru, Karnataka, India, E-mail: subrayack@gmail.com

Knowledge based [5] can be Fuzzy Logic Rule Based or Rule Based Systems & Evidence Based Systems [25]. Fuzzy Logic Rule Based [3] is a form of knowledge base and has achieved several important techniques and mechanisms to diagnose the disease and pain in patient. Ranking Vector Machine Learning Technique is used for pain management in patient who cannot communicate verbally. The technique of pattern recognition can assist medical professionals in measuring the pain [15] which is an extension of algorithm of Vector machine. The effectiveness of fuzzy set theory [19], Rough set theory can be improved by proposing complement fuzzy set and to discuss vagueness and uncertainty. The main scope is that it does not need data such as probability distribution in statistics, basic probability assignment and grade of membership of value of possibility.

Rule Based Systems & Evidence Based Systems [18] used to acquire the knowledge of domain experts into expressions that can be considered as rules. When a large number of rules have been compiled into a rule base, the working knowledge will be considered against rule base by integrating rules until a final conclusion is obtained. It is very helpful for storing a large amount of data and information. The gap between the physicians and CDSSs, can be closed by evidence based technique. It is a very powerful tool for improvement of clinical care and also patient outcomes. It has the tendency to improve quality and safety as well as reducing the cost [20]. Knowledge base contains the some rules, inference engine integrates rules with the patient data and to display the result to the users as well as to provide input to the system through communication mechanism.

M. Frize and R. Walker [4] to provide case based reasoning, they proposed Knowledge-based expert system. The Knowledge-based clinical decision support system contains rules in the form of IF-Then statements. The data is associated with these statements. For example if the pain intensity is up to a certain level then generate warning etc., They transformed raw patient data into patient cases and then provided inference rules to perform near-match search queries. This approach was not successful and suffered a significant loss of performance when patient data is incomplete (e.g. patients omit details, or access restrictions prevent viewing of remote medical records). Also it is difficult for an expert to transfer their knowledge into distinct rules.

2.2 Clinical Decision Support Systems using Neural Network and Genetic Algorithm

CDSS without a Knowledge base are called as Non-knowledge based CDSS. These systems use a form of artificial intelligence called as machine learning [6]. Non-knowledge based CDSSs are then further divided into two main categories: Neural Network and Genetic Algorithm.

Neural Network [16] derives relationship between the symptoms and diagnosis. It uses the nodes and connections with weights. This fulfills the need not to write rules for input. However, the system fails to explain the reason for using the data in a particular way. So its reliability and accountability can be a reason. It has been observed that the self organizing process of training the neural network in which it isn't given any priory information about the categories it is required to identify, is capable of extracting relevant information from input data in order to generate clusters correspond to class.

Furthermore it requires only a small proportion of available data to train the network. For example, in identifying the pain in infant child, neural networks extract the two features MFCC and LPCC from infant cry and are fed them into recognition module. The neural networks have been widely applied to non-linear statistical modeling problem and for modeling large and complex databases of medical information. Goal of training is to optimize performance of network in estimating output for particular input space.

Back propagation training algorithm, a popular approach used with medical databases adjusts weight of an ANN to minimize a cost function. The ANN maintains correct classification rates and allows a large reduction in complexity of the systems. The use of the weight-elimination cost function is well enough to overcome the network memorization problems. The neural networks are also very important in complex especially multi-variable systems to avoid costly medical treatment and for diagnosis of pain. It has the advantage that it does not need any input from experts. Eliminating the need of expert helps the system to eliminate the need of large databases to store input and output. It can work on incomplete data by guessing the data based on the successive data trend. But it has disadvantage that sometime the training process needs too much time.

Genetic Algorithms [7] are based on evolutionary process. Selection algorithm evaluates components of solutions to a problem. Solution that comes on top are recombined and the process runs again until a proper solution is observed. The generic system goes through an iterative procedure to produce the purpose the best solution of a problem.

2.3 Machine learning techniques to address Medical Information challenge

Oana Frunza, Diana Inkpen and Thomas Tran [8] proposed a machine learning techniques (ML) against the existing decision making process. The empirical domain of automatic learning is used in tasks such as medical decision support, medical imaging, protein to protein interaction, extraction of medical knowledge and for overall patient management care. It is also very important in complex especially multi-variable systems to avoid costly medical treatment and for diagnosis of pain [24]. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better, more efficient medical care.

2.4 Statistical approach for Clinical Decision Support Systems

Development of Clinical Decision Support Systems using statistical method [10] as an integral part is very common. For example to focus the economics of post operative pain with focus on the local regional anesthetic [22], a bibliographic database survey can be a good option. Data can be collected as a questionnaire mentioning the status of patient how he looks like, its way of talking, what he feels and many more. It can be a better way of quantitative and qualitative assessment of post operative pain [23].

Artificial Intelligence is an integral part of Decision Support Systems [21]. These systems provide a good aid to medical professionals [14]. Preliminary results confirm that for practical medical scenarios, where patient data is huge, a Knowledge-based system helps in decision making.

Decision Support Systems [11] implemented with the aid of Artificial Intelligence have the capability to adopt in new environment and to learn with time. Different approaches are used to gather information used for the process of Decision making Support Systems/ Expert Systems [15]. It does not need any input from experts. Need of expert helps the system to eliminate the need of large databases to store input and output. It can work on incomplete data by guessing the data based on the successive data trend.

Statistical methods are based on evolutionary process. Selection algorithm evaluates components of solutions to a problem. Solution that comes on top are recombined and the process runs again until a proper solution is observed. It is one of most simple and useful method used for data collection. It can be in the form of a survey, experiment result or questionnaire.

2.5 Flexible and Accurate Motif Detector algorithm to address Medical Information challenge

Avrilia Floratou, Sandeep Tata, and Jignesh M. Patel [12] were proposed a new algorithm called Flexible and Accurate Motif Detector (FLAME). In this paper, the main focus is on subsequence of Existing sequence mining algorithms. For a large class of applications, such as biological DNA and protein motif efficient mining of contiguous approximate patterns are required. It is desirable for a motif mining algorithm to be able to deal with a variety of notions of similarity. They presented a powerful new model for approximate motif mining that fits several applications with varying notions of approximate similarity. FLAME is a motif mining algorithm [13] which can efficiently find motifs that satisfy our model.

Developing such models poses an interesting fact: On the hand, one requirement is to build a model that is highly reactive enough to observe the occurrence of a pattern even in the presence of disturbance, which will be called as noise and on the other hand, it can be so generalistic as it matches unrelated subsequences. Since different applications may have different criteria for how to strike this balance, a natural approach is to develop a flexible model with a few intuitive parameters can be set by the user based on the application characteristics.

FLAME is helpful in finding frequent patterns in DNA sequences, which has profound importance in life sciences; the computational biology community has developed numerous algorithms for detecting frequent motifs using the Hamming distance notion of similarity.

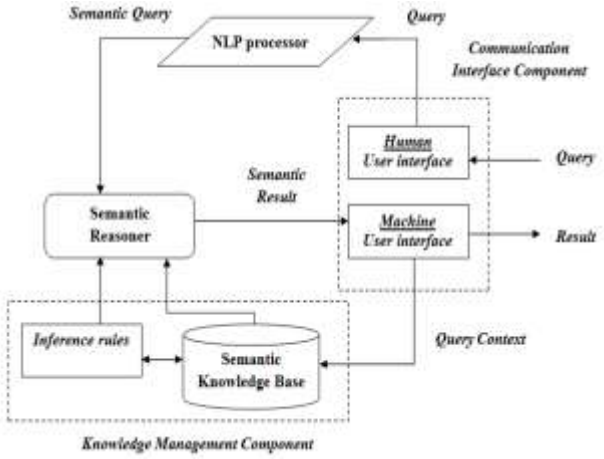
YMF, Weeder, MITRA and Random Projections are examples of algorithms in this category. Compared to this class of algorithms, FLAME is more flexible and can use more powerful match models. It is more scalable than these existing methods and can be an order of magnitude faster for mining large databases. There are several applications of motif mining in addition to those mentioned above.

It is often the first step in discovering association rules in sequence data (basic shapes and frequent patterns). It can also be used to find good seeds for clustering sequence data sets. Records of medical signals, like ECG or respiratory data from patients can also be mined to find signals that can indicate a potentially critical condition. An important part of gene regulation is mediated by specific proteins called transcription factors that influence the transcription of a particular gene by binding to specific sites on DNA sequences called transcription factor binding sites.

It has been observed that binding sites often emerge as a combination of two or more simple motifs separated by variable length spaces, especially in eukaryotic organisms. Mining these combinations of patterns called structured motifs is a challenging situation. The user has to specify the minima and maxima number of gaps between the simple motifs. However, since such gaps may not be known upfront, it is desirable to develop a more general structured motif mining algorithm. They assumed that the input sequence is composed of symbols from a discrete alphabet set. However, this method can also be applied to continuous time series data sets by converting such data sets into a symbolic sequence data set by simply discretizing the numeric data. In fact, such a transformation is frequently carried out for mining continuous time series data sets. They noted that their results produce more false positives when the available information is partial.

III. SYSTEM ARCHITECTURE:

3.1. Architecture for Knowledge based Medical Decision Support System:



Here, we describe the high level components of our proposed system framework. The knowledge management component has two major functions. It provides an abstraction for aggregating query specific information, using a plug in design, which allows real-time querying of distributed information repositories like relational databases, semantic knowledge stores and semi-structured document repositories.

The semantic query is broken down into atomic sub-queries and launched against the corresponding knowledge stores. Plugin translate the sub-queries into formats specific to each repository, and then translate the result back into a common semantic format to facilitate direct processing by a semantic reasoner. In addition to managing knowledge, it is responsible for identifying relevant inference rules for the decision making process. An inference rule describes a relationship between various facts in a knowledge base, permitting logical deduction of additional information from basic facts. All inference rules used by our proposed framework are based on ontological concepts, and can be processed by semantic reasoners.

The query execution component generates the response for the user query. It produces an answer with two tokens: the first is the query result, and the second is a proof showing the logical derivation of the result. This function of the Query execution component forms the core of the knowledge based part of our system. The main engine uses a semantic reasoner with the following inputs: a semantic query, aggregated information in N3 triple format, and inference rules. In the event that the semantic reasoner cannot answer a given query because information is missing, it returns the answer `unknown result'. This label is contrasted with the `negative result' and `positive result' labels which provide a negative answer to the query when all necessary data is present.

When the semantic reasoner produces an `unknown result' response to a query, the query execution component uses machine learning to impute the missing data and assigns a confidence value to the result. The confidence value is measured on a scale from 0 to 1, where 1 represents total confidence in the assigned value and zero represents complete uncertainty. This function of the query execution component forms the core of the learning-based (i.e. data-driven) part of our system. The knowledge base is then added with the missing data, and the query is reevaluated against the newly complete knowledge base, as indicated in the figure.

IV. IMPLEMENTATION:

4. 1 Module Specifications:

After careful analysis, the Integrated Medical Decision Support System has been identified to have the following modules:

- Communication interface module.
- Semantic reasoner module.
- Machine learning classifier module.

1. Communication Interface Module:

This module accepts user questions as composite queries and initiates the decision support workflow. Optional query constraints can be submitted as a problem context. Both human and computational agents can submit queries in either a constrained language or in a machine processable format. Each user query is translated into a semantic query that can be processed by a semantic reasoning engine using N3 triple notation. Finally the answer to the query is displayed to the user through this interface.

2. Semantic Reasoner Module:

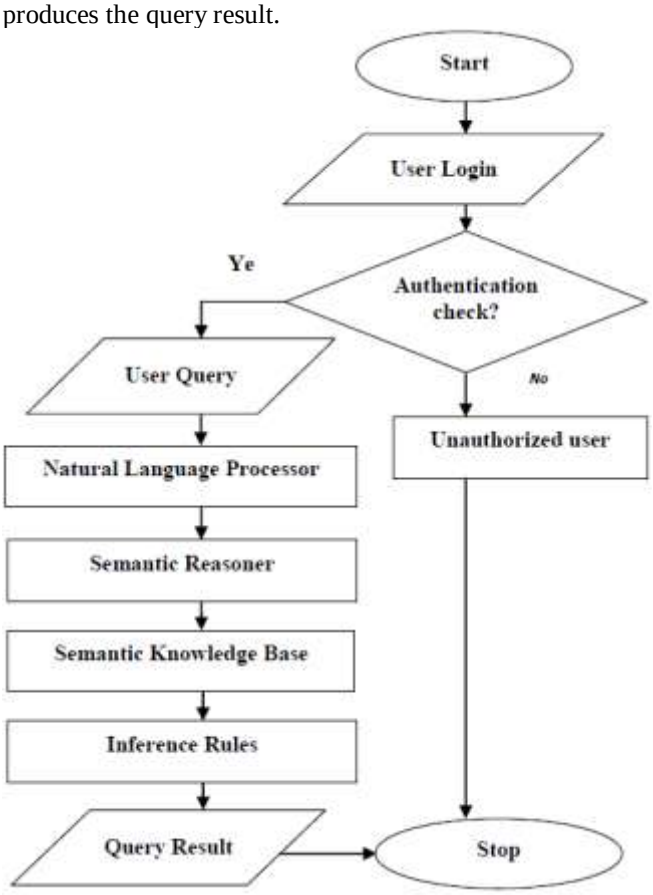
The semantic reasoner has the following inputs: a semantic query, aggregated information in N3 triple format and inference rules. It applies inference rules to the Knowledge base to answer the query. It produces the query result. In the event that the semantic reasoner cannot answer a given query because information is missing, it returns the answer unknown result.

3. Machine Learning Classifier Module:

The Machine learning Classifier uses the Classification algorithms (J48, JRip and Bagging) on the training set initially. Next the input test set is supplied by the user to predict the values to the missing field. Once the test set is supplied, this module predicts the value to the missing field based on the trained data sets.

4.2 Flow chart diagram for Knowledge based Systems:

The Knowledge base system has two major functions. It provides an abstraction for aggregating query specific information. It is also responsible for identifying relevant inference rules for the decision making process. It applies the semantic rules to the Knowledge base to answer the query. It produces the query result.



4. 2. Necessary requirements for constructing the Knowledge base

The Knowledge base is constructed using the Protégé tool that has inbuilt Euler Sharp inference engine. Protégé is a plug-in architecture tool adapted to build both simple and complex ontology-based applications.

4.2.1 Code snippet for designing the attributes related to Patient only:

The attributes that relate to only Patient class was identified such as bloodgroup etc. and the end user interface is designed by the following code snippet.

```
<?xml version="1.0" encoding="UTF-8"?>
<tells uri="" xmlns="http://dl.kr.org/dig/2003/02/lang">
<defattribute name="hasBloodGroup"/>
<rangestring>
 <attribute name="hasBloodGroup"/>
</rangestring>
<functional>
<attribute name="hasBloodGroup"/>
</functional>
</tells>
```

4.2.3 Code snippet to design the object properties between Drug and Patient:

The object properties between the classes Drug and Patient were identified such as is Consuming, has Consumed By etc. and the end user interface is designed by the following code.

```
<?xml version="1.0" encoding="UTF-8"?>
<tells uri="" xmlns="http://dl.kr.org/dig/2003/02/lang">
<defrole
name="hasToConsumeDrugForCurrentCondition"/>
<defrole name="hasConsumedBy"/>
<equalr>
<ratom name="hasConsumedBy"/>
<inverse>
<ratom name="isConsuming"/>
</inverse>
</equalr>
<defrole name="isConsuming"/>
<equalr>
<ratom name="isConsuming"/>
<inverse>
<ratom name="hasConsumedBy"/>
</inverse>
</equalr>
</tells>
```

4.3 Creating the Patient Instances

Nearly 500 Patient Instances were collected from the Government Hospitals of the rural places. These instance includes various attributes of the patient like Patient Id, Phone Number, Sex, Age, Email Id, Address, Mobile No, Bloodgroup, TreatedBy, Current_consuming_drug, Past_condition, Current_condition, Drug_for_current_condition etc.

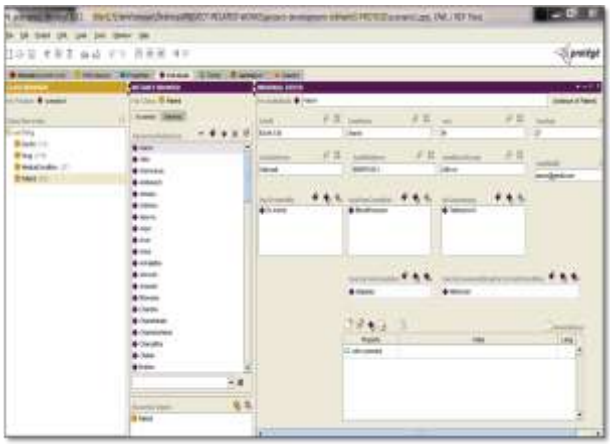


Fig 1: Creating the Patient Instances

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

4.4 Creating the Medical Condition Instances

Nearly 44 Medical Condition instances related to General Physician treatment were considered. These instances includes various medical conditions like Pregnancy, Burns and Scalds, UTI, Gastritis, Chickenpox, Asthma, Migrane, Throat pain, Diabetes,

Haematoma (Blood Clot), Blood Pressure etc.
Along with these medical conditions associated symptoms with each of these medical conditions, its causes and related tests were also created.

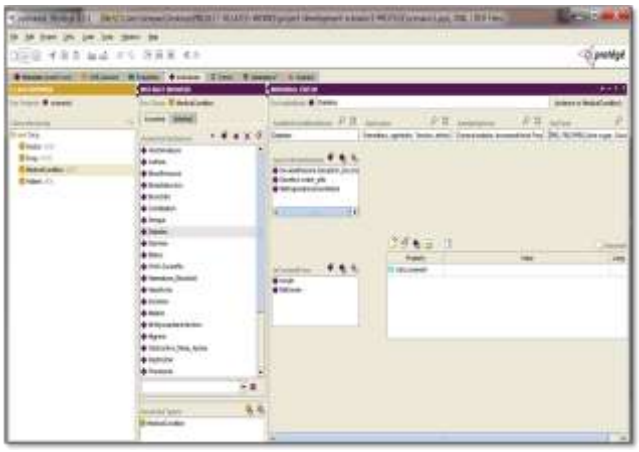


Fig 2: Creating the Doctor Instances

V. CONCLUSION:

In this work, a knowledge based Medical Decision Support System is developed to assist medical professionals, who are working at the remote places for prescribing the drugs. In this approach, we have considered knowledge base with machine learning classification to predict missing values by considering the symptoms of the patients.

Our approach of considering knowledge base makes use of inherent advantages of both approaches in order to offer the required accuracy for this domain. While we have sketched our framework in operation by collection of specific real world data sets and rule bases, we have outlined its usage in any medical decision context.

FUTURE ENHANCEMENT

We have considered artificial intelligence and machine learning techniques yield great advantage in imputing the missing data, which assist the general physicians who are working at the remote areas in prescribing the drugs. The machine learning activity in our work is carried out by considering diabetic dataset. It may be extended to other diseases such as cancer, cardiac, kidney, tumors etc. for prescription of the drugs. The prediction accuracy of the developed Integrated MDSS may be increased by considering more number of training sets and other more severe models of missingness including MAR and NMAR.

REFERENCES

1. M.M.Abbasi and S. Kashiyarndi, “Clinical Decision Support Systems: A discussion on different methodologies used in Health Care”, International Journal of Computer Science and Information Security, Vol. 8, No. 4, pages 249-256, 2010.
2. M. Frize and R. Walker, “Clinical decision-support systems for intensive care units using case-based reasoning”, Medical engineering & physics, Vol. 22, No.9, pages 671-677, 2006.
3. Y. Ye and S. J. Tong, “A Knowledge-Based Variance Management System for Supporting the Implementation of Clinical Pathways”, Management and Service Science, IEEE-2009, pages 1-4, 2009.
4. M. Goadrich, L. Oliphant and J. Shavlik, “Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction”, Proc. 14th Int’l Conf. on Inductive Logic Programming, pages 211-214, 2004.
5. T. Mitsumori, M. Murata, Y. Fukuda, K. Doi and H. Doi, “Extracting Protein-Protein Interaction Information from Biomedical Text with SVM ”, IEICE Trans. Information and Systems, Vol. E89D, No. 8, pages 2464-2466, 2006.
6. Oana Frunza, Diana Inkpen and Thomas Tran, “A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts”, IEEE Vol. 23, No. 6, pages 246-246, June 2011.
7. M. Zhu, Z. Zhang, J. Hirdes and P. Stolee, “Using machine learning algorithms to guide rehabilitation planning for home care clients”, BMC medical informatics and decision making, Vol. 7, Issue 1, pages 41-43, 2007.
8. D. Rossille, J. Lauren and A. Burgun, “Modeling a decision-support system for oncology using rule-based and case-based reasoning methodologies”, International Journal of Medical Information, pages 299-306, 2005.
9. Y. Ye and Z. Jiang, “A Semantics Based Clinical Pathway Workflow and Variance Management Framework”, Service Operation and Logistica and Informatics, IEEE, pages 758-763, 2008.
10. Avrilia Floratou, Sandeep Tata, and Jignesh M. Patel, Member, IEEE, “Efficient and Accurate Discovery of Patterns in Sequence Data Sets”, IEEE Transactions On Knowledge and Data Engineering, Vol. 23, No. 8, pages 30-37 August 2011.
11. P. Patel, E. Keogh, J. Lin and S. Lonardi, “Mining Motifs in Massive Time Series Databases”, Proc. of IEEE Int’l Conf. Data Mining (ICDM), pages 370-377, 2002.
12. X. Garg, N. K. J. Adhikari and H. McDonald, “Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review”, JAMA, pages 1223-1238, 2005.
13. L. Lin, P. Hu, O. R. Liu Sheng, “A decision support system for lower back pain diagnosis: Uncertainly management and clinical evaluation”, Decision Support Systems, pages 1152 -1169, 2006.
14. M. Frize, C. M. Ennett, M. Stevenson and H. Trigg, “Clinical decision support system for intensive care units: using artificial neural networks”, Medical Engineering & Physics, pages 217-225, 2001.
15. D. J. Spiegel halter and R. P. Knill-jones, “Statistical and Knowledge-based Approaches to Clinical Decision-support Systems with an Application in Gastroenterology”, J. R. Statist. Soc., pages 55-77, 2004.
16. E. Sivasankar and R. S. Rajesh, “Knowledge Discovery in Medical Datasets Using a Fuzzy Logic rule based Classifier”, IEEE International Conference on Electronic Computer Technology, pages 208-213, 2010.

AUTHOR PROFILE

Prof. Taranath N L Research Scholar, NCET, Bengaluru has completed his M.Tech and currently pursuing his Ph.D degree in Visweswaraiah Technological University, Belagavi. He has published more than 5 journals and 15 Papers in National/International conferences held at various cities throughout India. His research interests are in the field of Data Mining, Artificial Intelligence, Machine Learning.

Dr. Shantakumar B Patil, Professor and Head, Dept. of CS & E, NCET, Bengaluru has completed his Ph.D degree in Computer Science and Engineering from Dr. MGR University, Chennai, Tamilnadu. He has published more than 15 journals and 15 papers in National and International Conferences. His research interests are in the field of Data Mining, Artificial Intelligence, Machine Learning. He is guiding 5 research students under VTU, Belagavi, Karnataka.

Dr. C.K.Subbaraya, Principal & Professor , Dept. of CS & E, AIT, Chikkamagaluru has completed his Ph.D degree in Fluid Mechanics from Bengaluru University, Karnataka. He has published more than 50 journals and 100 papers in National and International Conferences. His research interests are in the field of Data Mining, Artificial Intelligence, Machine Learning.