BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **13**, No 3

Sofia • 2013 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2013-0030

# Design of a Diabetic Diagnosis System Using Rough Sets

*Margret Anouncia S., Clara Madonna L. J., Jeevitha P., Nandhini R. T.*

*School of Computing Science and Engineering, VIT University, Vellore-14, Tamilnadu, India*
*Email: smargretanouncia@vit.ac.in*

**Abstract:** *Traditionally the diagnosis of a disease is done by medical experts with experience, clinical data of the patients and adequate knowledge in identifying the disease. Such diagnosis is found to be approximate and time-consuming since it purely depends on the availability and the experience of the medical experts dealing with imprecise and uncertain clinical data of the patients. Hence, to improve decision making with uncertain data and to reduce the time consumption in diagnosing a disease, several simulated diagnosis systems have been developed. Most of these diagnosis systems are designed to possess the clinical data and symptoms associated with a specific disease as knowledge base. The quality of the knowledge base has an impact not only on the consequences, but also on the diagnostic precision. Most of the existing systems have been developed as an expert system that contains all the diagnosis facts as rules. Notably, applying the concept of a fuzzy set has shown better knowledge representation to improve the decision making process. Therefore an attempt is made in this paper to design and develop such diagnosis system, using a rough set. The system developed is evaluated using a simple set of symptoms that is added to clinical data in determining diabetes and its severity.*

**Keywords:** *Disease diagnosis, knowledge based system, a rough set, diabetes.*

## 1. Introduction

Without a systematic diagnosis in medical science, no correct treatment can be accomplished. The advancement in computer technology and in the field of artificial intelligence has encouraged the researchers to develop software assisting doctors in decision making. Also, the focus of all medical systems that are developed is to build better health care facility in order to reduce time, cost and medical error, say Hussain et al. [10]. Some of the traditionally followed knowledge representation techniques describe knowledge in terms of rules and cases. In addition to these techniques, several other representations, such as frames, scripts, semantic networks and conceptual graphs are widely adopted in literature. Though the techniques appear to be easy and flexible, they have proved to be ambiguous and difficult to scale.

Recently, the concept of logic pre-dominates knowledge representation and the rough set is such knowledge representation technique.

## 2. Related works

Currently for medical planning, diagnosis and treatment, knowledge-based systems and intelligent computing systems are used. In general, these systems are widely used in areas that require heuristic and logic in reasoning where knowledge is predominant over data. Mostly these system approaches involve models of either fuzzy logic, or artificial neural networks, or a genetic algorithm, or combination of all these techniques with an appropriate reasoning mechanism. Szolovits and Pauker [19] say that medical decision making involves categorical reasoning and probabilistic reasoning that helps the system to realize expert-like behaviour. It is suggested that a program with an expertise in the area of medical consultation should use a combination of categorical reasoning to establish different contexts, that are not wide enough after categorization, and probabilistic reasoning, in order to make comparisons and eventually to recommend therapy.

According to Pandey and Mishra [14] these knowledge-based systems use reasoning techniques like Rule-Based Reasoning (RBR), Case-Based Reasoning (CBR) and Model-Based Reasoning (MBR) to provide significant performance in the area of diagnosis.

Earlier, Golobardes et al. [6] stated that CBR is one of the most popular prediction techniques in medical diagnosis as it is easy to apply and provide good explanation of the output. Their system demonstrated the power of integrating the CBR with a genetic model to improve the performance in diagnosing breast cancer.

Especially for medical diagnosis, CBR can be combined with other techniques to promote improvements in diagnosis, Lin [13]. In his work, for the diagnosis of a liver disease, an intelligent model has been structured using CBR and Regression Trees (CART) to increase the accuracy of the diagnosis. The rules are extracted from CART and the model acted as a supporting system in the diagnosis of a liver disease. The system identified the presence of the disease using regression trees and proceeded to diagnose the disease further using the CBR to provide treatment suggestions for the disease.

In a similar way, to improve classification and prediction accuracy, Pham and Triantaphyllou [16] suggested a new meta-heuristic approach called the Homogeneity Based-Algorithm (HBA) that is combined with traditional classification methods. The system considered some cases, such as false-positive, false-negative, unclassifiable case and the fatal results due to incorrect diagnosis in these cases. In such cases the HBA defines the misclassification cost of models extracted from classification algorithms and organizes the extracted models as mutually exclusive decision regions expressing them by homogenous sets. On the other hand, applying fuzzy logic into medical diagnosis became popular. Some of the remarkable applications included "A Fuzzy Expert System for Heart Disease Diagnosis" by Adeli and Neshat [2]. The system provided better results in about 94 % compared to that of an expert. Similarly, Baydaa S. Bhnam [4] also appreciates the effectiveness of fuzzy logic in the designed expert system that helped in decision making for diagnosing liver and pancreas diseases. The system provided a better result due to the application of fuzzy logic to infer the conclusion. Hasan et al. [8] developed a knowledge-based online diagnosis system to determine his/her probable diseases very quickly with the aid of a knowledge-based expert system. The method adopted a fuzzy based methodology for the diagnosis. Priti Srinivas, Sajja and Shah [17] developed a prototype of knowledge-oriented decision support system for advisory, diagnosis and awareness in abdomen pain. The model incorporated very few sample rules for diagnosis. Chung and Chen [5] presented a knowledge-based decision system using a rule base for healthcare. It not only performs intelligent diagnoses but also produces inferential advices for the interrelated diseases involving overweight or obese, diabetes, high blood pressure and high cholesterol conditions. A five-layer fuzzy ontology to model the domain knowledge with uncertainty and extend the fuzzy ontology to the diabetes domain was designed by Lee and Wang [12]. Experimental results indicate that the method proposed can analyze data and further transfer the acquired information into knowledge to simulate the thinking process of humans. The work had limitations for having addressed the issues in the context of only one data set. The system modelled only domain knowledge of diabetes and the scope for improvement is left with the fuzzification. Adeyemo and Akinwonmi [1] designed a neural model for rapid diagnosis of diabetes mellitus. The focus of the work was to help the physicians in the diagnosis process through measurable symptoms. This network achieved a Mean Square Error (MSE) of 0.57 on the diagnosis test data set and 49.0 on the treatment test data set. The network gave acceptable results when predicting the diagnosis, but not the associated treatment. On the other hand, in early 2000, Grzymala-Busse et al. [7] proposed the prediction of melanoma using data mining system LERS (Learning from Examples based on Rough Sets). The system proved to have significant performance by reducing the error rate of classification. Thus the applicability of rough set theory for better disease diagnosis became evident. Also, for an incomplete information system, a method called relative discernible measure is proposed to measure the relative discern ability of the attributes. The advantage of this method is that the attributes in an incomplete information system can be reduced without pre-processing continuous data by discrimination, say Tsang et al. [21]. Followed by Eric et al., a remarkable usage of rough set in medical diagnosis was suggested by Tettey et al. [20]. They suggested that the rough sets analysis is a better and quick way of data analysis and rule extraction over neuro-fuzzy model. Rule extractions using a rough set model and neuro-fuzzy models were compared and a rough sets model was found to be more advantageous. Rough sets have features, such as sensitivity and specificity analysis for characterizing medical data and prior setting. Similarly, Herbert and Yao [9] say that the advantage of using rough sets is mainly the use of upper and lower approximations which form the basic tools and the rough sets are directly computed from the data input. The traditional approach in medical science, such as the employment of analysts to use statistical methods to deliver reports, is no longer efficient due to the increase of uncertainty in data. According to them, a Decision Maker (DM) faces difficulty in choosing a suitable rough set model for data analysis. There are two rough set models: the Pawlak model and the probabilistic model. These two approaches use either user-defined parameters which can be given by the domain experts or derive the probability thresholds from the cost associated, making a classification. Pattaraintakorn and Cerkone [15] proposed that since rough sets have the ability to handle colour images, signals and graphs, they are also used in bioinformatics. Rough sets are useful in analyzing and extracting essential attributes from the given data. Rough sets are used in the knowledge discovery process in data with missing values. Association rules are used to derive results from the data for the problem domain under concern. Yang and Wu [22] say that rough sets prove to be efficient in representing medical data or knowledge that is usually incomplete with a high degree of uncertainty and for analyzing such data and synthesize approximations. Rough sets can be applied to identify a set of significant symptoms that cause the disease and to induce decision rules using the data provided by the clinic or human experts. Recently, a Rule-Based Expert System (RBES) for Neurological Disorders, i.e., Alzheimer, Parkinson, Huntington's disease, Cerebral Palsy, Meningitis, Epilepsy, Multiple Sclerosis, Stroke, Cluster headache, Migraine, Meningitis in children (Al-Hajji [3]. Though this expert system is intended to be used as a consultation system for differential diagnosis of 10 types of neurological disorders cases, it leaves a scope on improving the reasoning mechanism. Very recently Srivastava et al. [18] have proposed a Soft Computing Diagnostic System for Diabetes. The system defined limited fuzzy sets and the performance is focused on guiding the patients to evolve proper strategies in maintaining their blood sugar level. Therefore, from literature it is clear that the rapid and unclear changes in diseases symptoms and the equally rapid growth of technology have induced the development of diagnosis systems. Even though the existing systems have proved to be efficient, there are certain aspects that lower the efficiency, such as accuracy, reliability and time consumption. Hence, to improve the mentioned qualities in disease diagnosis, an expert system with proper knowledge representation through a rough set is proposed. The diagnosis system using a rough set is intended to minimize the time consumption and to provide accurate diagnosis for a disease with incomplete or changing symptoms. The rough set generated makes it easy to consider all possible chances of the occurrence of a disease or its specific type, making the system more reliable.

## 3. Developed framework

A centralized knowledge base consisting of facts of diabetes in terms of symptoms is formalized. Four different modules processing the data related to the disease diagnosis, such as patient data, pathological data, doctor's input and inference engine are designed to have interaction with this knowledge base.

### 3.1. Description of the system

As mentioned in the previous section, the disease diagnosis system comprises of:

I. Patient Description
II. Doctor's Input/ Description
III. Pathological Details
IV. Inference engine

A detailed illustration of the modules defined in the architecture, is given in Fig. 1.

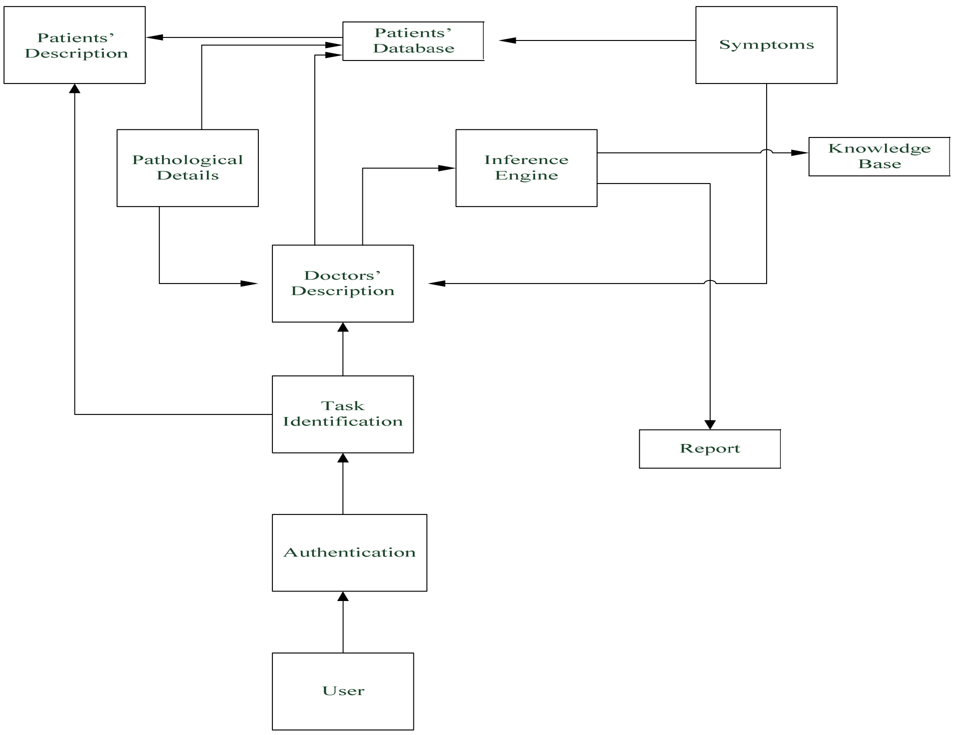


Fig. 1. Disease diagnosis system

#### 3.1.1. Patient description

Initially, the details of a new patient are obtained by the receptionist and recorded. The details, such as patients' first name, last name, address, city, state, pin code, gender, date of birth, contact number and consultant name are obtained and stored. After consultation with a doctor, if any pathological tests are recommended, the results of the pathological tests are updated along with the patient record. Once the doctor finalizes the diagnosis, the result is also updated in the patient's record. Thus these descriptions add all the factual data of the patients visiting the physician.

### 3.1.2. Doctor's input/description

The description includes a questionnaire to be answered by the user. The answers given by the patient are recorded and stored for decision making.

The questionnaire includes facts listed below:

- Excess thirst
- Excess hunger
- Frequent urination
- Fruity breath odour
- Bedwetting
- Weight loss
- Being overweight
- Poor wound healing
- Weight fluctuation
- Frequent infections
- Blurred vision
- Irritability
- Increased fatigue
- Itchy skin
- Family history of pre diabetes
- Depression and stress
- Tingling sensations in feet and fingers
- Family history of diabetes during pregnancy
- Diabetes during previous pregnancy
- Previous pregnancy that resulted in a child with birth weight of 9 pounds or more.

### 3.1.3. Pathological details

The pathological details include details of haematological, biochemical, clinical pathological and bio-chemical urine tests. When needed and as per the instructions given by the doctor, the patient might need to take a few of these pathological tests. The reports of the tests taken are stored in the patient's record for future reference.

### 3.1.4. The inference engine

The inference engine is the main processing module. The input to the engine is the set of answers to the questionnaire from the doctor's input. The inference engine interacts with the knowledge base which is constructed using rough sets for the process of diagnosis. It processes the indiscernible sets of the universal set representing the concepts associated with diabetes and its symptoms. The interpretations are carried using the user input and the available knowledge base.

## 4. Implementation of the disease diagnosis system

### 4.1. Tools

With all the analyzed requirements in mind and after considering the limitations of the existing systems, a new diagnosis system is designed. The system is implemented using Java and JSP. The rough set generation is implemented efficiently using Java and the interactions are implemented using JSP. The connection between the database, the front end and the Java programs is established using the JDBC connection.

### 4.2. Methodology

#### 4.2.1. Rough set – a general idea

The information system consists of a universe with a finite number of objects, a finite set of attributes, a set of values for the attributes and functions. The input to the system is called a target set which consists of objects which are the subset of the universe. This set is used to generate the lower and upper approximations which are then used for diagnosis.

A rough set is one of the most commonly used analysis methods for diseases with inconsistent and incomplete data. Medical databases are usually incomplete due to the rise of new and unidentified symptoms and types. Such incomplete databases cannot be efficiently used for accurate analysis of the disease. For such systems the use of rough sets for analysis is favourable.

A rough set is derived from an information system. Generally, a finite set of an information system is defined as $S = \langle U, Q, V, f \rangle$ where $U$ is the universe, $Q$ is the finite set of attributes, $V$ is the set of values for the attributes and $f$ is the function. From this finite set the indiscernible objects are derived to form the rough set. The disease is analyzed using this set by calculating the closest approximation. This method minimizes the time consumed and also the accuracy is enhanced.

#### 4.2.2. Rough set approach

Adopting the concept of a rough set in knowledge representation of a disease diagnosis, an input target set is created by the user. The indiscernible objects are separated out from the given input. The objects that are indiscernible with each other and do not have any indiscernible objects in the KB are considered as a lower approximation set. The objects that are different, are compared with the knowledge base and the objects indiscernible with these objects along with the lower approximation set form the upper approximation set.

#### 4.2.3. The Pawlak rough set model

Considering all the variants of the rough sets, Pawlak rough set model appears as a pioneer and better. In this model the approximation sets are formed around the equivalence classes which are created through an equivalence relation of the existing objects in the universe set. Thus a partition is generated from the universe set. The region, derived from the approximations, is used as a guiding principle for decision making.

According to Pawlak model (K o m o r o w s k i et al. [11]), the definitions of lower and upper approximations:

$$\underline{\mathrm{Apr}}(A) = \{x \in U | [x] \subseteq A\}$$

$$\overline{\mathrm{Apr}}(A) = \{x \in U | [x] \cap A \neq \varnothing\}.$$

The lower approximation of A, $\underline{\mathrm{Apr}}(A)$ is the union of all elementary sets that are included in A. The upper approximation A, $\overline{\mathrm{Apr}}(A)$ is the union of all elementary sets that have a non-empty intersection with A.

The positive, negative, and boundary regions of A can be defined as

$$\mathrm{POS}\ (A) = \underline{\mathrm{Apr}}(A),$$

$$\mathrm{NEG}\ (A) = U - \overline{\mathrm{Apr}}(A)$$

$$\mathrm{BND}\ (A) = \overline{\mathrm{Apr}}(A) - \underline{\mathrm{Apr}}(A)$$

The positive region, POS($A$), consists of all objects that are definitely contained in the set $A$. The negative region, NEG($A$), consists of all objects that are definitely not contained in the set $A$. The boundary region, BND($A$) consists of all objects that may be contained in $A$. Since the approximations are formed from equivalence classes, inclusion into the boundary region reflects uncertainty about the classification of objects.

This approach poses a great challenge in data analysis. The major challenge in data analysis using rough sets is to reduce the size of the region to minimum. This is solved by relaxing the definitions of POS and NEG regions. The probabilistic model explains how the size of the regions can be reduced, say Peter and Stephen.

### 4.3. The rough sets – as implemented in the system

The main goal of the rough set analysis is to synthesize the approximation of concepts from the acquired data. The data can be acquired from measurements or from human experts.

#### 4.3.1. Information system

A data set is represented as a table, where each row represents a case, an event, a patient, or simply an object. Every column represents an attribute (a variable, an observation, a property, etc.) that can be measured for each object; the attribute may also be supplied by a human expert or user. This table is called an information system.

Let IS be the information system. It is a pair IS = ($m$, $n$) where $m$ is the set of objects called the Universe and $n$ is a non-empty finite set of attributes such that $a : m \rightarrow V_a$ for every $a \in n$. The set $V_a$ is called the value set of a. The knowledge base for the system is shown in Table 1.

Table 1. Knowledge base of the disease diagnosis system

| No | Symptoms | Type 1 | Type 2 | Pre diabetes | Gestational |
|---|---|---|---|---|---|
| 1 | Excess thirst | 1 | 1 | 1 | 0 |
| 2 | Frequent urination | 1 | 1 | 1 | 0 |
| 3 | Excess hunger | 1 | 1 | 1 | 0 |
| 4 | Weight loss | 1 | 0 | 0 | 0 |
| 5 | Over weight | 0 | 1 | 1 | 1 |
| 6 | Weight fluctuation | 0 | 1 | 0 | 0 |
| 7 | Blurred vision | 0 | 1 | 1 | 0 |
| 8 | Increased fatigue | 1 | 1 | 0 | 0 |
| 9 | Irritability | 0 | 1 | 1 | 0 |
| 10 | Frequent infection | 0 | 1 | 1 | 0 |
| 11 | Itchy skin | 0 | 0 | 1 | 0 |
| 12 | Family history | 0 | 0 | 1 | 0 |
| 13 | Depression & stress | 0 | 0 | 1 | 0 |
| 14 | Tingling sensation | 0 | 0 | 1 | 0 |
| 15 | Fruity breath odor | 1 | 0 | 0 | 0 |
| 16 | Bed wetting | 1 | 0 | 0 | 0 |
| 17 | Poor wound healing | 0 | 1 | 0 | 0 |
| 18 | Family history of diabetes during pregnancy | 0 | 0 | 0 | 1 |
| 19 | Previous pregnancy | 0 | 0 | 0 | 1 |
| 20 | Baby over 9 pounds during previous pregnancy | 0 | 0 | 0 | 1 |

### 4.3.2. Indiscernibility

A decision system (i.e., a decision table) expresses all the knowledge about the model. This table may be unnecessarily large in part because it is redundant in at least two ways. The same or indiscernible objects may be represented several times.

*A binary relation $R \subseteq X \times X$ which is reflexive (i.e., an object is in relation with itself xRx), symmetric (if xRy then yRx) and transitive (if xRy and yRz then xRz), is called an equivalence relation.*

The equivalence class of an element $x \in X$ consists of all objects $y \in X$ such that $xRy$. Let IS = ($m$, $n$) be an information system, then with any $B \subseteq n$, an equivalence relation $\mathrm{IND}_A(B)$:is associated:

$$\mathrm{IND}_A(B) = \{(x; x') \in U^2 \mid \forall a \in B \; a(x) = a(x')\},$$

$\mathrm{IND}_A(B)$ is called the $B$-indiscernibility relation.

If $(x; x') \in \mathrm{IND}_A(B)$, then objects $x$ and $x'$ are indiscernible from each other by attributes from $B$.

The indiscernibility sets for the given knowledge base are explained below by considering one attribute, type 1.

IND({type 1}) = {{1, 2, 3, 4, 8, 15, 16}, {5, 6, 7, 9, 10, ...}}.

The indiscernibility sets by considering attributes type 1 and 2,

IND({type 1, type 2}) = {{1, 2, 3, 8}, {5, 6, 7, 9, 10, 17}, {11, 12, 13, 14, 18, 19}, {4, 15, 16}.

4.3.3. Set approximations

The set approximation is one of the important tasks in analyzing the input set. The process requires the creation of equivalence relations among the objects present in the universe set. An equivalence relation induces partitioning of the universe (the set of cases in our example). These partitions can be used to build new subsets of the universe. The set of objects with a positive outcome cannot be defined crisply using the attributes available. It may happen that a set cannot be defined in a crisp manner. Here the rough set concept comes into use. Although it is not possible to define those objects crisply, it is possible to delineate them to have a positive outcome – objects that certainly do not have a positive outcome and objects that belong to the boundary between the certain cases. When this boundary is non-empty, the set is rough.

Considering the disease diagnosis system, the equivalence classes (formed by taking all the attributes into consideration) for the knowledge base of the system is identified as follows:

{1, 2, 3}
{4, 15, 16}
{5}
{6, 17}
{7, 9, 10}
{8}
{18, 19, 20}
{11, 12, 13, 14}

Further, in obtaining the target set from the user, the objects in it are compared with each other. If the objects are indiscernible and do not have any other indiscernible object when compared with the equivalence classes, then these objects form the lower approximation set. If the objects are discernible, the objects are compared with the equivalence classes and if any class has another object that is indiscernible, these objects along with the lower approximation set, form the upper approximation the set.

These approximation sets are used for comparison and further analysis with the knowledge base.

An example diagnosis is given in details in the following section:

Let the input set given to the inference engine be {1, 2, 3, 4, 8, 9}.

The objects in the set are then compared with each other and the indiscernible objects are compared with the equivalence classes. If there are no indiscernible objects, these objects form the lower approximation set. The other set, which includes the objects from the equivalence classes other than the input objects and along with the lower approximation set, forms the upper approximation set. Hence, for the given input the lower approximation set is $L_{\text{approx}} = \{1, 2, 3, 8\}$ and the other objects are compared with the equivalence classes resulting in the upper approximation set $U_{approx}$= {1, 2, 3, 4, 15, 16, 7, 8, 9, 10}. Considering the upper approximation set, the KB is compared with the attributes (for the objects in $U$ only) and the attribute with the closest values of objects in the upper approximation set is displayed as a result of the diagnosis. In this case the diagnosis for the given input is type 3. The count of the input is 6. The count of the attributes for the objects in $U$ is:

type 1: 7 type 2: 8 type 3: 6 type 4: 0.

The closest approximation is Type 3 and hence the result.

### 4.4. Procedure

The procedure of the diagnosis of the disease by the inference engine is depicted as a flow diagram in Fig. 2.

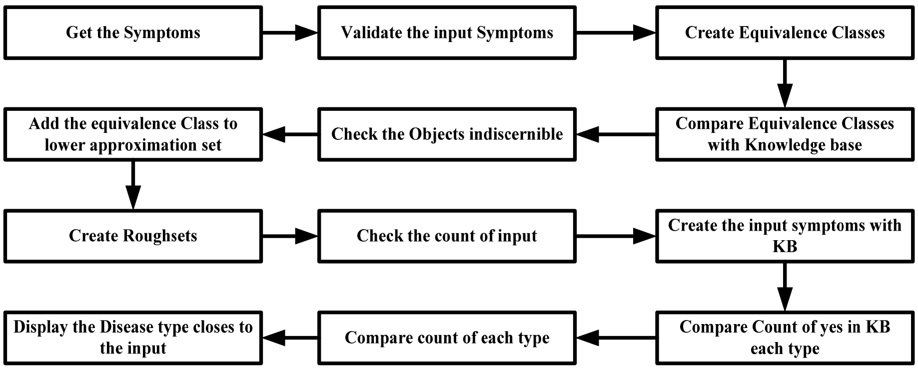


Fig. 2. Disease diagnosis procedure

Some of the screen shots of the developed system are provided in Figs 3-5.

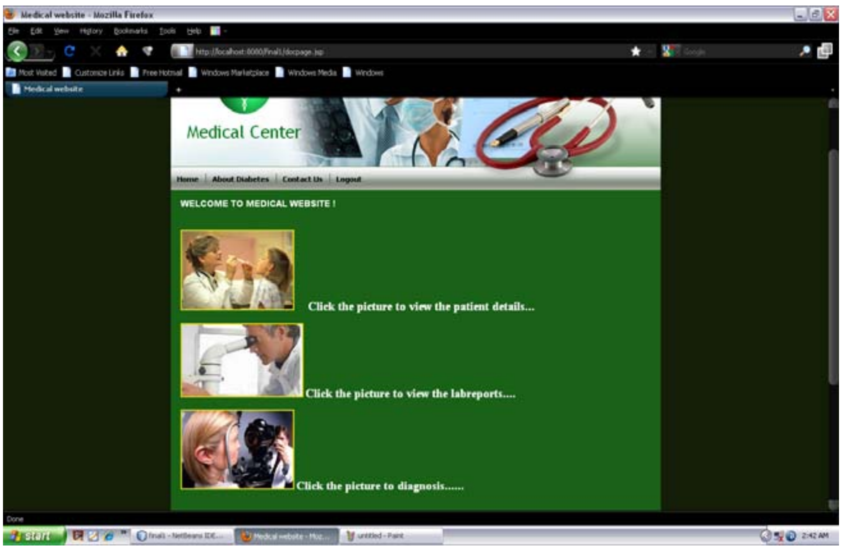


Fig. 3. Doctor's home page

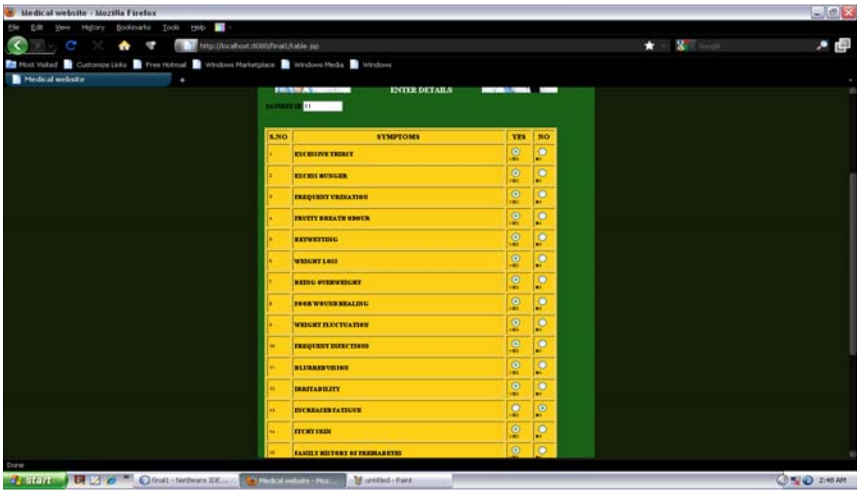


Fig. 4. Questionnaire


Fig. 5. Resultant diagnosis

## 5. Evaluation of the disease diagnosis system

Rough set methodology is used to develop a disease diagnosis system which diagnoses the type of diabetes an individual has. To test the system for its correctness and reliability with respect to the analyzed requirements, black box testing is adopted. In black box testing the tester inputs data and views the output from the test object. This testing involves thorough test cases to the user who then verifies them for a given input, the generated output is the same as the expected output in the test case.

Every unit in the system is considered and analyzed and a unit testing is made. Black box testing is favoured in all testing phases. Finally, performance analysis is made and depicted in Fig. 7. In order to demonstrate the efficiency of the system, a simple rule based system is developed. Sample rules created for the RBS are listed as:

i. **IF** ((Excess thirst=true) and (Frequent urination=true) and (excess hunger=true)) **then** type1 **or** type2 **or** pre diabetes.

ii. **IF** ((Weight loss= true) and (Fruity breath odor = true) and (Bed wetting = true) **then** type1.

iii. **IF** ((Itchy skin = true) and (Family history = true) and (Depression & stress = true) and (Tingling sensation = true)) **then** pre diabetes.

iv. **IF** ((Blurred vision = true )and ( Irritability and Frequent infection =true )) **then** type2 **or** pre diabetes

v. **IF** (Increased fatigue = true) **then** type2

vi. **IF** (Over weight = true) **then** type2 **or** pre diabetes **or** Gestational

vii. **IF** ((Weight fluctuation = true ) and (Poor wound healing = true)) **then** type2

viii. **IF** ((Family history of diabetes during pregnancy= true) and ( Previous pregnancy = true ) and (Baby over 9 pounds during previous pregnancy = true )) **then** Gestational

Likewise the 25 rules are framed to cover the possible symptoms and diagnosis.

5.1. Performance study

Once the integration of the system is completed, its performance study is done with respect to other existing systems. In this case the existing rule-based system is considered for comparison. The diagnosis is done using a set of rules that are generated using the data in the KB. According to the rule base the system matches the symptoms and the type to make the diagnosis. As the input to the system is a set of answers by the user, the system does not make comparisons with the other possible symptoms (other than the input) in the KB, that are related to the input, due to this the result might prove to be wrong at times. It is also observed that the time consumption is more in rule-based systems as the system has to check every rule generated.

In the system which uses rough set technique, the approximation sets are generated first and the diagnosis is done by taking only those objects into account. The diagnosis is made after taking all the possible conditions into consideration. The whole knowledge base needs not to be scanned for diagnosis as is the case of rule-based systems where the whole set of rules are to be checked for making the correct diagnosis.

Considering fifty patients' symptoms as an input for both rule-based and rough set based system, a comparison was made. The overall performance of the system is shown in Fig. 6**.**

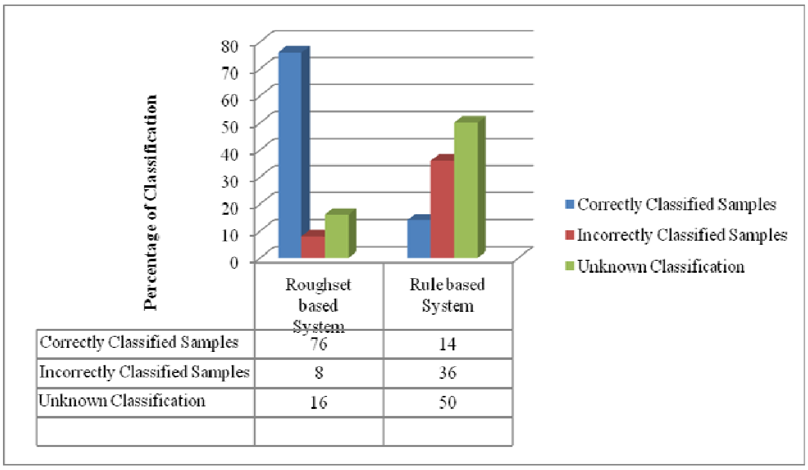


Fig. 6. Performance graph

It was observed that the accuracy of prediction is better in the case of the rough set based system. Due to the generation of the approximation sets the accuracy of prediction is enhanced. Ultimately the performance of prediction in the rough set based system is 76 %. The system has a misclassification of 8 and 16 % of unknown prediction where it suggested two types simultaneously. From the results obtained through the rule-based system with a limited number of rules in the rule base, it is noted that 50 % of the prediction is unknown, 36 % is misclassification and only 14 % – correct prediction. The prediction may be improved by increasing the rule base. As per the comparison with the limited number of symptoms, the performance of the rough set based system is observed to be better.

## 6. Conclusion and future enhancement

A diagnosis system for diabetes is proposed in this paper. The system is implemented to diagnose the type of diabetes with the input symptoms given by the user. The system proves to be advantageous in aspects, such as accuracy and time consumption due to the rough set based knowledge representation. The system is adaptable for any number of symptoms and is evaluated with respect to the rule-based system containing the symptoms in terms of rules. The results obtained through the rough set based system are comparatively better than the rule-based system. Though the system is developed with respect to some aspects, such as accuracy and time consumption in mind, there are still a few limitations existing. The few limitations of concern include that the present system knowledge base is designed only for one disease.

The symptoms accommodated for the disease are based only on the patients' input.

The system mainly focuses on diagnosing a disease with the user input and comparing the input with the knowledge base. This system can be further extended to large databases where the time consumption for data retrieval and analysis is more. This rough set technique has advantages over other AI techniques and can be efficiently used in real time and intelligent systems. As future enhancement the methodology can be extended for various diseases diagnoses that depend on pathological interpretations.

## References


1. Adeyemo, B., A. E. Akinwonmi. On the Diagnosis of Diabetes Mellitus Using Artificial Neural Network Models Artificial Neural Network Models. – African Journal of Computing & ICT, Vol. **4**, 2011, No 2, Issue 1, 1-9.
2. Adeli, A., M. Neshat. A Fuzzy Expert System for Heart Disease Diagnosis. – In: Proceedings of the International Multi Conference of Engineers and Computer Scientists, Vol. **1**, Hong Kong, 2010.
3. Al-Hajji, A. A. Rule-Based Expert System for Diagnosis and Symptom of Neurological Disorders. – In: Neurologist Expert System (NES), Proceedings of International Conference on Communications and Information Technology, 2012, 67-72.
4. Baydaa, S. Bhnam. Design a Fuzzy Expert System for Liver and Pancreas Diseases Diagnosis. – AL-Rafidain Journal of Computer Sciences and Mathematics, Vol. **7**, 2010, No 2, 129-141.
5. Chung, Ping-Tsai, Bing-Xing Chen. A Knowledge-Based Decision System for Healthcare Diagnosis and Advisory. – In: Proceedings of International Conference on Systems, Man, and Cybernetics (SMC), 2011, 2357-2362.
6. Golobardes, E., X. Llora, M. Salamo, J. Marti. Computer Aided Diagnosis with Case-Based Reasoning and Genetic Algorithms. – Knowledge-Based Systems, Vol. **15**, 2002, 45-52.
7. Grzymala-Busse, J. P., J. W. Grzymala-Busse, Z. S. Hippe. Melanoma Prediction Using Data Mining System LERS. – In: 25th Annual International Computer Software and Applications Conference, 2001, 615-620.
8. Hasan, M. A., A. R. Chowdhury, K. M. Sher-E-Alam. Human Disease Diagnosis Using a Fuzzy Expert System. – Journal of Computing, Vol **2**, 2010, Issue 6, 66-70.
9. Herbert, J. P., Jing Tao Yao. Criteria for Choosing a Rough Set Model. – Computers & Mathematics with Apllications, Vol. **57**, 2007, Issue.6, 908-918.
10. Hussain, W., W. Ishak, F. Siraj. Artificial Intelligence in Medical Applications: An Exploration. – In: National Conference on Management Science: New Paradigm for Knowledge Economy, University of Malaysia, Serdang, Selangor, 2001. Preprint: Artificial Intelligence in Medical Application: An Exploration. Health Informatics, Europe Journal, 2002.
11. Komorowski, J., L. Polkowski, A. Skowron. Rough Sets: A Tutorial. 1999, Rough Fuzzy Hybridization: A New Trend in Decision-Making, 3-98. **http://secs.ceas.uc.edu/~mazlack/dbm.w2011/Komorowski.RoughSets.tutor.pdf**
12. Lee**,** Chang-Shing, M. H. Wang. A Fuzzy Expert System for Diabetes Decision Support Application. – Systems, Man, and Cybernetics. Part B: Cybernetics, IEEE Transactions, Vol. **41**, 2011**,** Issue 1, 139-153.
13. Lin, Rong-Ho. An Intelligent Model for Liver Disease Diagnosis. – Artificial Intelligence in Medicine, Vol. **47**, 2009, Issue 1, 53-62.
14. Pandey, B., R. B. Mishra. Knowledge and Intelligent Computing System in Medicine. – Computers in Biology and Medicine, Vol. **39**, 2009, Issue 3, 215-230.
15. Pattaraintakorna, P., N. Cerconeb. Integrating Rough Set Theory and Medical Applications. – Applied Mathematics Letters, 2008, 400-403.
16. Pham, H. N. A., E. Triantaphyllou. An Application of a New Meta-Heuristic for Optimizing the Classification Accuracy when Analyzing Some Medical Datasets. – Expert Systems with Applications, Vol. **36**, 2009, 9240-9249.

17. S a j j a, P r i t i S r i n i v a s, D i p t i M. S h a h. Knowledge Based Diagnosis of Abdomen Pain Using Fuzzy Prolog Rules. – Journal of Emerging Trends in Computing and Information Sciences, Vol. **1**, 2010, Issue. 2, 55-60.
18. S r i v a s t a v a, P., N. S h a r m a, R. S i n g h. Soft Computing Diagnostic System for Diabetes. – International Journal of Computer Applications, Vol. **47**, 2012, 22-27.
19. S z o l o v i t s, P., S. G. P a u k e r. Categorical and Probabilistic Reasoning in Medical Diagnosis. – Artificial Intelligence, Vol. **11**, 1978, 115-144.
20. T e t t e y, T., F. V. N e l w a m o n d o, T. M a r w a l a. HIV Data Analysis via Rule Extraction Using Rough Sets. – In: Intelligent Engineering Systems, 11th International Conference, 2007, 105-110.
21. T s a n g, E. C.. S u-y u n Z h a o, D. S. Y e u n g, J. W. T. L e e. Continuous Attributes Reduction in Incomplete Information System Based on Rough Set Technique. – Machine Learning and Cybernetics, Vol. **5,** 2005, 3130-3134.
22. Y a n g, H s u-H a o, C h a n g-L u n W u. Rough Sets to Help Medical Diagnosis – Evidence from a Taiwan's Clinic. – Expert Systems with Applications, Vol. **36**, 2009, Issue. 5, 9293-9298.