Application Tool based on C4.5 Decision Tree for Diagnosing Diabetes Infection Symptoms

[1]Amal H. Khaleel, [2] Ghaida A. Al-Suhail and [1] Bushra M. Hussan

[1]Department of Computer Science, University of Basrah, [2]Department of Computer Engineering, University of Basrah, Basrah, Iraq
Email: amal_albahrany@yahoo.com, ghaida-alsuhail@yahoo.com, bushra5040@yahoo.com

***Abstract*—**Diabetes is a fatal disease which can lead to many other dangerous illnesses such as blindness, hypertension, kidney failure, heart attacks, and gangrene. Nowadays, the intelligent diabetic systems exploit important tools to both medical industry and diabetes patients due to their crucial role in improving the quality of healthcare in many ways. Such systems can be considered as a very helpful tool to allow the physicians and the doctors intervene to find the proper treatment for the patient with all kinds of conditions. This paper therefore presents an application of intelligent system for diabetes detection symptoms to support and give advice to clinical management and patients. It basically relies on how to detect and find the Probability of Infection Diabetes. Hence, if a person suffers from the symptoms of a group, a patient will be referred to the possibility of diabetes. System diagnostics are examined based on the algorithm of Szajnar and Setla. It starts whether there is infect or not infect/doubt in the possibility of injury. When the person is doubted with the probability of injury, the probability of injury with symptoms can pass through a set of resulting rules from the C4.5 decision tree algorithm. The results reveal the finding ratios of Incidence percentage by the sum of the values of the outputs of the rules derived. It was ascertained the validity of the results by comparing them with the Indian diabetes database whereas if the injury rate of less than 50 is not infected and greater or equal to 50 is infected. Consequently, the implementation of this expert application tool shows very good results.

Keywords—Diabetic Diagnosis; Data Mining, Intelligence Tools; C4.5 Classifiers; Healthcare; Medical Expert System

I. INTRODUCTION

Studies in the field of medical decision support systems have been established and due to the high success rate of these studies, interest in this field is increasing every day. These systems frequently use various artificial intelligence techniques. There is a greater interest in the study of diseases that are common throughout the world. Diabetes is one of them [1].

Diabetes is a serious, life-threatening and chronic disease. It is estimated that this figure will reach 366 million by 2030 [2], Diabetes is a disease in which levels of Blood Glucose, (blood sugar), are above normal. People with diabetes have problems converting food to energy because the pancreas does not make enough insulin or because the cells in the muscles, liver, and fat do not use insulin properly, or both. Some families have a history with diabetes, once one of the parents is diabetic, and then some of sons as well as the grandchildren are diabetic. Diabetes does not distinguish between adult and young. It is one of the major chronic illnesses prevailing today. This disease has many complications. It causes severe damage for the body organs. These serious degenerative complications are such as retinopathy, neuropathy and nephropathy. Furthermore, diabetes affects an estimated 2-4% of the world's population [3].

Most medical resources reported that 90 to 95% of diabetic is diagnosed as type-2. Simply, in these cases the pancreas is not able to produce enough insulin to keep the blood sugar level within normal ranges. In addition, the majority of this type diabetics do not know they are suffering from it [1].

The intelligent systems used in diabetic are important within the medical industry because they allow doctors and nurses to quickly gather information and process it in various ways in order to assist with making diagnosis and treatment decisions. These systems could help in diverse areas from the storing and retrieval of medical records, storing and retrieval of key substances in medicines, examination of real-time data gathered from monitors, analysis of patient history for the purposes of diagnosis, analysis of family history (for cardiac conditions for example), and in many other areas. An expert system is a computer program that provides expert advice as if a real person had been consulted where this advice can be decisions, recommendations or solutions [4].

In this paper we propose system to detection and find probability of infection diabetes. The rest of the contents are organized as follows: Section 2 illustrates the prior work related, while Section 3 bbackground and existing Method. Section 4 presents details related proposed algorithm and section 5 describes the Results and discussions. Conclusion and future work are given in section 6.

II. RELATED WORK

There is a variety of research work which has been carried out by many researchers based on the observed medical diabetes data.

P. M. Beulah et. al [5] introduced the ability to access diabetic expert system from any part of the world. They collect, organize, and distribute relevant knowledge and service information to the individuals. The project was designed and programmed via the dot net framework. The system allows the availability to detect and give early diagnosis of three types of diabetes namely type 1, 2, gestational diabetes for both adult and children.

Cindy. M et.al [6] presents a case-based decision support system prototype to assist patients with Type 1 diabetes on insulin pump therapy, detect common problems in blood glucose control, and retrieval metrics were developed to find the most relevant past cases for solving current problems to control blood glucose levels.. The system is developed in JAVA.

M.Wiley et.al [7] presented diabetes management tool that monitors and controls blood glucose (BG) levels in order to avoid serious diabetic complications. They mentioned the difficult task for physicians, to manual large volumes of blood glucose data to tailor therapy of each patient. Also they describe three emerging applications that employ AI to ease this task. Actually, their system enables: (a) automatic problems detection in BG control (b) offering solutions to the detected problems (c) remembering the effective and/or ineffective solutions for individual patients type1 diabetes (T1D).furthermore their system might be embedded in insulin pumps or smart phones to provide low-risk advice to patients in real time. Finally, they used support vector regression (SVR) model for building the system.

W.Szajnar and G.Setlak [8] proposed a concept of building an intelligence system of support diabetes diagnostics, where they implemented start-of-art method based on artificial intelligence for constructing a tool to model and analyze knowledge acquired from various sources. The initial target of their system was to function as a medical expert diagnosing diabetes and replacing the doctor in the first phase of illness. Diagnostics the sequence of dealing with their system were as flow: (1) getting patient information and symptoms (2) competing basic medical examination in details (3) based on previous information the system find out whether the patient has diabetes and decides whether it is type1 or type2. The systems used decision tree as a model for classification.

M. Kalpana and A.V Senthil Kumar [9] proposed a fuzzy expert system framework which constructs large scale knowledge based system effectively for diabetes. The knowledge is constructed by using the fuzzification to convert crisp values into fuzzy values. By applying the fuzzy verdict mechanism, diagnosis of diabetes becomes simple for medical practitioners. The proposed fuzzy expert system for diabetes application was implemented with the MATLAB using rules for knowledge representation.

S. Kumar and B. Bhimrao [10] developed a natural therapy system for healing diabetic; they aim to help people's health and wellness, which don't cost the earth. Their main goal was to integrate all the natural treatment information of diabetes in one place using ESTA (Expert System Shell for Text Animation) as knowledge based system. ESTA has all facilities to write the rules that will make up a knowledge base. Further, ESTA has an inference engine which can use the rules in the knowledge base to determine which advice is to be given to the user. Their system begins with Consultation asking the users to select the disease (Diabetes) for which they want different type of natural treatment solution then describes the diabetes diseases and their symptoms. After that describes the Natural Care (Herbal /Proper Nutrition) treatment solution of diabetes disease.

N. Nnamoko et.al [11] proposed a fuzzy expert system framework that combines case-based and rule based reasoning effectively to produce a usable tool for Type 2 Diabetes Mellitus (T2DM) management, to produce crisp outputs to patients in the form of low-risk advice. The extended framework features a combined reasoning approach for simplified output in the form of decision support for clinicians. With seven operational input variables and two additional preset variables for testing, the results of the proposed work compared with other methods using similarity to expert's decision as metrics.

Margret Anouncia S. et.al [12] proposed a diagnosis system for diabetes. 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. The inference engine interacts with the knowledge base which is constructed using rough sets for the process of diagnosis. The system is implement\ted using Java and JSP.

Abdelhak.et.al [13] proposed an approach based on using a multi-criteria decision guided by a case based reasoning (CBR) approach. The study is intended to experiment with a multiple criteria decision approach to medical care in the diagnosis and the proposed therapy for diabetic patients. The system is developed in JAVA with an interconnecting module to the JCOLIBRI system.

Cindy Marling et.al [14] Presented systems are for CARE-PARTNER, which supports the long-term follow-up care of stem-cell transplantation patients; diabetes Support System, which aids in managing patients with type 1 diabetes on insulin pump therapy; renal disease; diagnosis and treatment of stress-related disorders using case-based reasoning.

III. BACKGROUND AND EXISTING METHOD

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Medical and biological research studies have used various techniques of data analysis including, statistical machine learning and other analysis methods. A medical report always

gives useful information for diagnosis and also facilitates therapeutic improvement [15].

The data mining techniques have been applied on diabetes data base are: Supervised machine learning algorithm, like Naive Bayes, Decision tree, and Classification, Clustering, Association Rule Mining, Temporal Data Mining, More specifically, there are various data classification algorithms available in DM. Among these, Decision trees (DTs) used for this research is discussed hereafter.

A. Decision Tree Technique

Decision trees (DTs) are one of the fundamental techniques used in data mining. They are tree-like structures used for classification, clustering, feature selection, and prediction [16].

A decision tree (DT) is a flow chart-like tree structure that represents the knowledge for classification, where each internal node (no leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label and the topmost node in a tree is the root node [17].

For a given record, the classification process starts from the root node. The attribute in the node is tested, and the value determines which edge is to be taken. This process is repeated until a leaf is reached. The record is then classified as the class of the leaf. Decision tree is a simple knowledge representation for a classification model [18].

Decision trees accept several types of variables: nominal, ordinal, and interval. A variable can be of any type regardless of whether it serves as an input or as the target. As a result, decision trees are easily interpretable, amenable to graphical display, and intuitive for humans and fast and usually produce high-quality solutions.

The objective of all DT algorithms is to minimize the size of the tree while maximizing the accuracy of the classification. Decision trees (DTs) are widely used in data mining for classification purposes [16].

B. DT Construction [19]

The Some premises guide DT algorithm is given as the follows:

- If all cases are of the same class, the tree is a leaf and so the leaf is returned labeled with this class.
- Calculate the gain in information that would result from a test on the attribute (based on the probabilities of each case with a particular value for the attribute being of a particular class).
- Find the best attribute to branch on, depending on the current selection criterion.

C. The Requirements of DT Algorithms [20]

The There are several requirements that must be met before applying decision tree algorithms:

1) *Decision tree algorithms represent supervised learning, and as such require preclassified target variables.*
2) *The training dataset should be rich and varied.*
3) *The target attribute classes must be discrete. That is, one cannot apply decision tree analysis to a continuous target variable. Rather, the target variable must take on values that are clearly demarcated as either belonging to a particular class or not belonging.*

D. Decision Trees Algorithms [21]

A decision tree model consists of two parts: creating the tree and applying the tree to the database.

To achieve this, different algorithms are commonly used for mining knowledge represented in decision trees as follows:

1) *Classification And Regression Trees (CART) Algorithm*
Algorithm CART is published by L.briemen and associates (1984).

2) *CHi-square Automatic Interaction Detection (CHAID) Algorithm*
Algorithm CHAID is published by J.A.Hartigan (1975).

3) *Iterative Dichotomiser 3 (ID3) Algorithm*
Algorithm ID3 of R. Quinlan proposed in 1986.

4) *Change Iterative Dichotomiser (C4.5) Algorithm*
Algorithm C4.5 was worked out by Quinlan 1993, this algorithm is in fact only one improvement of ID3.

E. The C4.5 Algorithm

The C4.5 algorithm is J. Ross Quinlan's extension of his own ID3 algorithm for generating decision trees [8]. It visits each decision node recursively, selecting the optimal split, until no further splits are possible [15].

C4.5 algorithm improves ID3 algorithm by using a gain ratio as the criterion for selecting the branching attribute. The attribute with the maximum value on gain ratio(X) is selected as the branching attribute. Also C4.5 allows the tree to grow and prunes the unnecessary branches later by replacing a sub tree by a leaf or the most frequently used branch [18].

F. The Measures in C4.5 Algorithm

There are many measures that affect the learning algorithm of C4.5, these measures are as follows:

1) Entropy Measure

The concept of entropy is known from Shannon's information theory. It is a measure of the uncertainty concerning an event and from another view point is a measure of randomness of the messages [22].

The entropy measure can be used for attribute evaluation. Since the entropy-based attribute evaluation, also known as information gain, it is one of most popular supervised attribute

selection mechanisms. It also plays a major role in a very popular algorithm for decision tree learning [23].

Formally, the entropy or the expected information is needed to classify a tuple in D is defined as

$$\text{Info(D)}= -\sum_{i=1}^{m} pi \log 2(pi) \quad (1)$$

Where, Info(D) is also known as entropy(D), m is number of class, pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by
|Ci,D|/|D|. Info(D) is just the average amount of information needed to identify the class label of a tuple in D. Note that, at this point, the information we have is based solely on the proportions of tuples of each class [9].

2) Gain Measure

The gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is

$$\text{Gain(A)}= \inf\ o(D) - \inf\ oA(D) \quad (2)$$

Where Info(D) is information that we need (before the partitioning) and previously defined and $\text{Info}_A(D)$ is the information that we need (after the partitioning) in order to arrive at an exact classification is measured by

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{/ DJ /}{/ D /} \times Info(Dj) \quad (3)$$

The term $\frac{\upsilon_j}{D}$ acts as the weight of the jth partition. $\text{Info}_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information (still) required is, the greater the purity of the partitions will be.

In other words, Gain(A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest Gain(A), is chosen as the splitting attribute at node N [17].

3) Gain Ratio Measure

C4.5 uses an extension to information gain known as gain ratio, which attempts to overcome bias the information gain measure toward tests with many outcomes. It applies a kind of normalization to information gain using a "split information" value defined analogously with Info(D) as

$$\text{SplitInfo}_{A(D)}= -\sum_{j=1}^{v} \frac{/ Dj /}{/ D /} \times \log 2(\frac{/ Dj /}{/ D /}) \quad (4)$$

This value represents the potential information generated by splitting the training dataset D, into v partitions, corresponding to the v outcomes of a test on attribute A.
Note that, for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D. It differs from information gain, which measures the information with respect to classification that is acquired depending on the same partitioning. The gain ratio is defined as

$$\text{GainRatio(A)}= \frac{Gain(A)}{SplitInfo(A)} \quad (5)$$

The attribute with the maximum gain ratio is selected as the splitting attribute. Note, however, that as the split information approaches, the ratio becomes unstable. A constraint is added to avoid this, whereby the information gain of the test selected must be large at least as great as the average gain over all tests examined [17].

G. The C4.5 algorithm [21]

Entry: sample S;
Beginning
Initialize with the empty tree; the root is the current node;
Repeat
Calculate the entropies for each value of each attribute;
Calculate the profit for each attribute;
Choose the maximum profit;
Choose the test for the current node;
Decide if the node current is final (leaf);
If the node is final then to affect a class;
If not to select a test and create the under tree;
End if
Pass to the following node not explored if there are;
Until obtaining a decision tree;
End.

In C4.5 algorithm, the entropy and profit (Gain Ratio) are considered as mathematical measures in classification process.

H. Performance Measures

For the calculation of the predicted positive cases the below mentioned formulas are used [24]

- True positive (TP): Those Sick people who are correctly diagnosed as sick
- False positive (FP): The Healthy people who are incorrectly identified as sick
- True negative (TN): The Healthy people who are correctly identified as healthy
- False negative (FN): The Sick people who are incorrectly identified as healthy

Various performance measures like sensitivity, specificity, accuracy and F-Measure are calculated using this matrix as depicted in Table 1:

TABLE1 THE PERFORMANCE MEASURE FORMULAS [25]

Performance Measure	Formulas
Precision	TP / (TP+ FP)
Recall (Sensitivity)	TP / (TP+ FN)
Accuracy	(TP + TN) / (TP+ TN +FP +FN)
Specificity	TN / (TN+ FP)
F-Measure	(2* Recall* Precision) / (Precision + Recall)

I. Data Set and Attributes

The Indian Diabetes dataset, the dataset consists of 8 attributes plus class (Table 2). The dataset was collected from 768 females. The diagnosis can be carried out depending on personal data (age, number of times pregnant) and results of medical examination (blood pressure, body mass index, result of glucose tolerance test, triceps skin fold thickness, serum insulin, pedigree function). There are 500 samples of class 1 (diabetes) and 268 of class 2 (not diabetes). The original source of the data in Indian is the National Institute of Diabetes, and we have used in our work is taken from - http://mlearn.ics.uci.edu/MLRepository.html [26]

TABLE2 the Characteristics Used for Diabetes Type II Diagnose [27]

No. of Feature	Feature	Descriptions and Feature Values
1	Number of times Pregnant	Numerical values
2	Plasma Glucose Concentration	Numerical values
3	Diastolic Blood Pressure	Numerical values in (mm Hg)
4	Triceps Skin Fold Thickness	Numerical values in mm
5	2-Hour Serum Insulin	Numerical values in (mu U/ml)
6	Body Mass Index (BMI)	Numerical values in (weight in kg/(height in m)^2)
7	Diabetes Pedigree Function (DPF)	Numerical values\
8	Age	Numerical values
9	Diagnosis of type 2 diabetes disease	sick=1 , Normal=0

IV. THE PROPOSED ALGORITHM

The proposed algorithm consists of two stages, as follows:

A. Application Tool for Detection Diabetes

If the person suffers most of the clinical symptoms (such as increased thirst, hunger, going to the toilet more often (especially at night), unexplained weight loss, slow healing wounds, general weakness (feeling very tired), blurred vision, tingling in the fingers of the feet and hands, drought in the skin and mouth, the rapid rate of breathing); or if the person is more predisposed to Type2 diabetes (such as old age, unfavorable family interview, obesity, hyperlipidemia , hypertension), it may be infected from diabetes (Type1, Type2 ,and pre-diabetes).

In this application tool, we could implement the diagnostic scheme of diabetes with regard to factors increasing diabetes risk as proposed by W. Szajnar and G. Setlak in [28], Figure (1) illustrates the details of this scheme.

In Figure (2) Screen-1: symptoms, if the person's age is 20 years or less, he might be infected with (diabetes Type 1 or non-infected). On other hand, if the person's age is greater than 20 years, it is possible that he is infected with (diabetes Type 2 or with pre-diabetes or is not infected). Depending on the type of blood test (such as fasting or random or tolerance), unit of measurement and the level of blood sugar, the caregiver (doctor) can diagnose the patient status (as shown in Figure (2) Screen-2).

Note : The measurement blood sugar level uses the unit mg/dl or nmol/l. The unit nmol/l is converted to the mg/dl by using the formula [28] :

$$mg/dl = nmol/l \times 18 \quad (6)$$

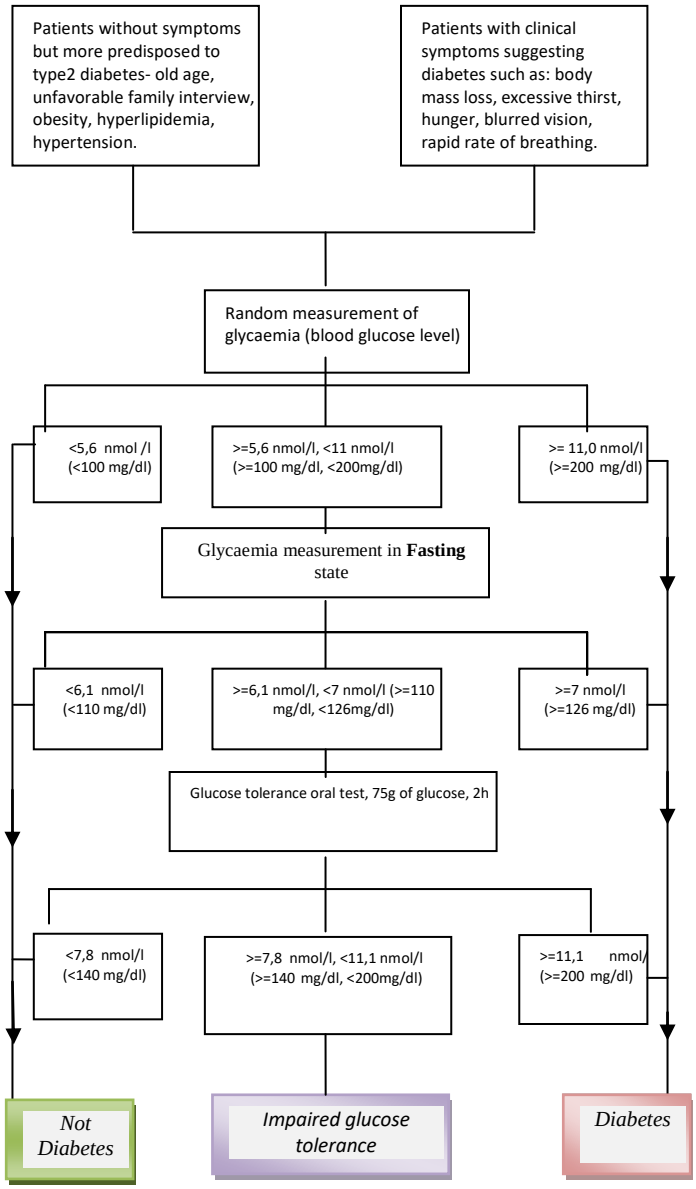
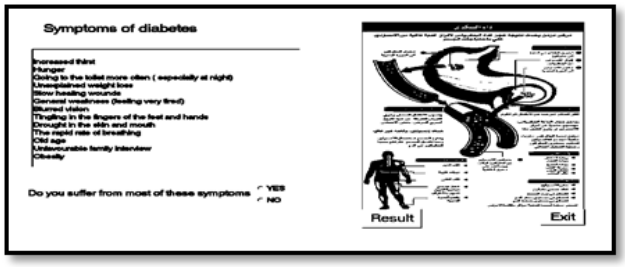
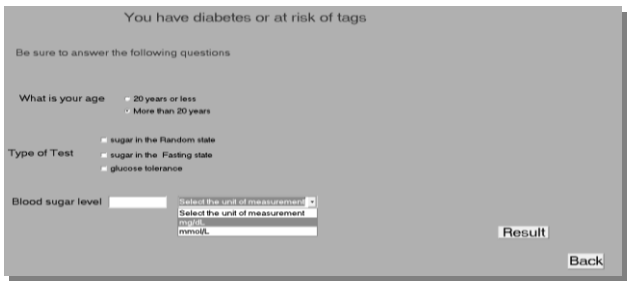


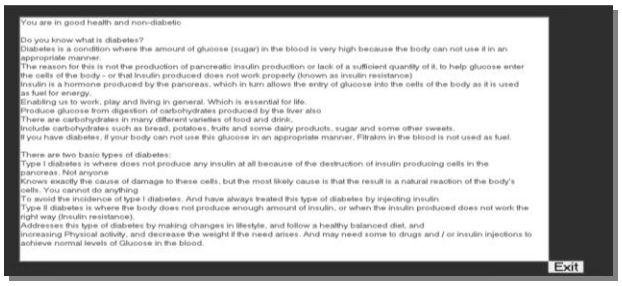
Fig. 1. Diagnostic scheme of diabetes with regard to factor increasing diabetes risk [28]



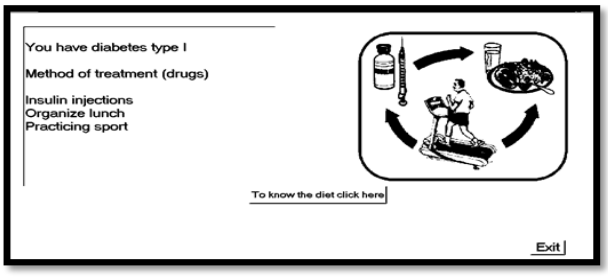
Screen-1: Symptoms



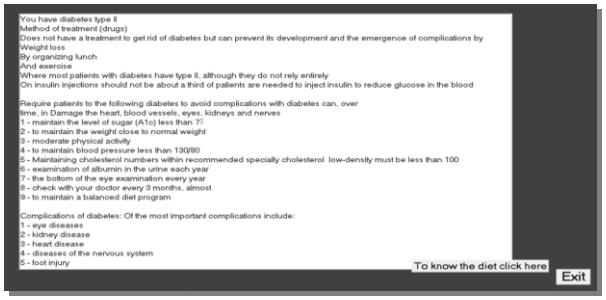
Screen-2: Age Diabetes Risk



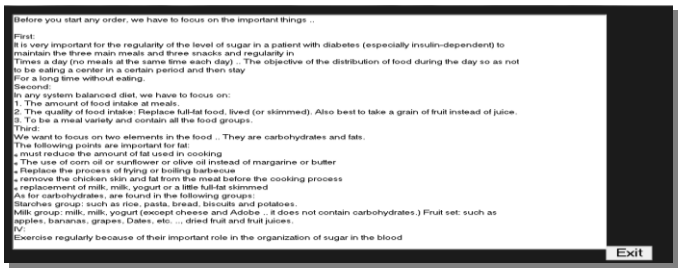
Screen-3: Good health (Non_diabetes)



Screen-4: Infected with Diabetes Type I



Screen-5: Infected with Diabetes Type II



Screen-6: To know the diet, Press Icon

Fig. 2. Application tool for detecting diabetes symptoms

B. Discovering the Probability of Infection Diabetes

If the type of test is glucose tolerance oral test and the result is impaired glucose tolerance (IGT) then the person is infected from pre-diabetes. To discover the probability of the infection diabetes, there are several factors required to expert or predict the risk of diabetes as follows (see Figure 3):

1. Sex (male, female)*
2. Age
3. Length
4.Weight
5.Blood_pressure
6.Blood sugar level (Glucose tolerance test)
7. Genetics

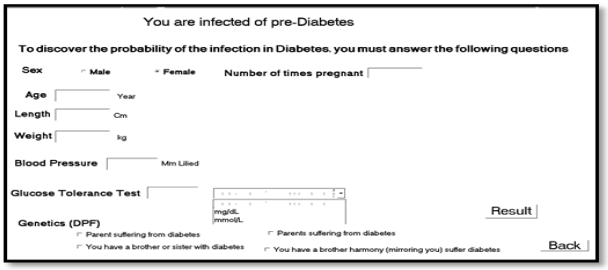


FIG. 3. THE INFORMATION SCREEN

* If sex is female then display the number of passed pregnancy factor in this screen.

Determine the probability of the risk of infection of diabetes Type 2 has relied on two factors as follows:

1. Indian Diabetes dataset after handling

In this study, the missing values have been handled by getting rid of the zero values (missing values) in the Indian Diabetes dataset. Hence, we cannot rely on replacing the zero values because the dataset will not be real therefore we used the remove way as follows:

- In the attribute Pregnant, the zero values are left assuming that they are the real values.
- In the attributes Plasma- Glucose, Diastolic BP, and BMI, the zero values were removed (because the number of missing values in these attributes is small).
- In the attributes Triceps SFT and Serum– Insulin, the zero values were not removed (because the number of missing values in these attributes is large), so these attributes were removed.
- Finally, the attributes diabetes pedigree function (DPF) and age patient, were not including the missing values so there is no need for handling them .

After handling the missing values, only 724 (475 Class1, 249 Class2) instances remain out of 768 with 6 attributes: Pregnant, plasma-glucose, diastolic BP, BMI, DPF and age.

2. Probability of Infection

The determination of the probability of infection includes diabetes for all factors using the C4.5 Rules -Based Diabetes Method

For Diabetes Classification, we propose the C4.5 rules algorithm to generate requited rules rather than tree. In this algorithm, we reduce the phases by generating rules which are required in classification rather than generating binary tree and then transforming the tree into rules to be used in classification.

TABLE 3. The symbols used in the formulas of C4.5 Rules -Based Diabetes Method

Symbol	Description
S	Attribute
C	Number of classes (i.e. i=1,2).
A	Number of split parts (it can be split based on value of threshold, where threshold represents the mean value of values attribute)
N	Total number of elements (the number of records of patients)
n1	Number of elements that fall under the if condition part (if value of attribute<=value of threshold)
n2	Number of elements that fall under the else condition part (else if value of attribute > value of threshold)
S1	Number of people which does have illness within the group n (The result of diagnosis of disease = 1)
S2	Number of people which doesn't have illness within the group n (The result of diagnosis of disease= 0)
S11	Number of people which does have illness within the group n1 (The result of diagnosis of disease = 1)
S12	Number of people which doesn't have illness within the group n1 (The result of diagnosis of disease = 0)
S21	Number of people which does have illness within the group n2 (The result of diagnosis of disease = 1)
S22	Number of people which doesn't have illness within the group n2 (The result of diagnosis of disease = 0)

The following steps describe the details of our proposed C4.5 rules algorithm in Diabetes Classification:

- C4.5 Algorithm Steps:

1. Input : The Diabetes Dataset (Indian)

2. Replace the missing data by the values calculated from hybrid k-nearest neighbour imputation approach.
3. Repeat

For each attribute (S) in the diabetes dataset do:

i. Calculate entropy :

$$\text{Entropy}(s) = -\sum_{i=1}^{c} pi \log 2\, pi$$

ii. Calculate the profit (Gain Ratio):

$$\text{Gain Ratio}(S) = \frac{Gain(s)}{SplitInfo(s)}$$

iii. Choose the test for the attribute that have maximum profit and generate if rule and print it.

iv. If the attribute is final then /*all values in attribute belong to same class/

Else

Select the following attribute test not explored if there are, and create the new rule.

End if

Until obtaining decision rules.

End.

The formulas of C4.5 Rules -Based Diabetes Method as follows:

$$P1 = \frac{number\ \ people\ with\ diabetes}{total\ number\ of\ \ people}$$

$$P2 = \frac{number\ \ people\ without\ \ diabetes}{total\ number\ of\ \ people}$$

$$\text{Gain (S)} = entropy\,total(s) - \sum_{v \in A} \frac{/SV/}{/S/} entropy(sv)$$

$$\text{SplitInfoA(s)} = -\sum_{v \in A}^{v} \frac{/sv/}{/s/} \times \log 2 (\frac{/sv/}{/s/})$$

Gain(S,A)= average entropy before splitting – average entropy after splitting

The entropy before splitting

$$= -(\frac{s1}{n}) \times \log 2 (\frac{s1}{n}) - (\frac{s2}{n}) \times \log 2 (\frac{s2}{n})$$

The entropy after splitting $= \sum_{v \in A} entropy(sv)$

$$= (\frac{n1}{n}) \times T1 + (\frac{n2}{n}) \times T2$$

$$\text{T1(entropy)} = -\frac{s11}{n1} \times \log 2 (\frac{s11}{n1}) - (\frac{s12}{n1}) \times \log 2 (\frac{s12}{n1})$$

$$\text{T2(entropy)} = -\frac{s21}{n2} \times \log 2 (\frac{s21}{n2}) - (\frac{s22}{n2}) \times \log 2 (\frac{s22}{n2})$$

- The Results of C4.5 Rules:

The knowledge extracted from decision tree is classified as IF-THEN rules and when these rules are used on diabetic database, we obtained more of 100 rules but choose only 100 rules to use it for find the probability of infection as follows:
Test data person (which is a possibility that become infected) with resulting from the decision tree rules,
Where

$$\text{infection rate} = (\ number\ of\ \ rules\ achieved \Big/ 100\ ) \times 100$$

The results of some the rules that obtained from decision tree as follows Figure 4:

Fig. 4. A Sample of the Expert System's Rules

Rule condition_ 1

IF Pregnant <=1 and Plasma- Glucose <95 and Diastolic BP <=80 and BMI <24.9 and DPF<0.42 and Age <=40 THEN probability of infection =33

IF Pregnant>=2 and Plasma- Glucose <95 and Diastolic BP <=80 and BMI <24.9 and DPF<0.42 and Age <=40 THEN probability of infection =41

Rule condition_2

IF Pregnant <=1 and Plasma- Glucose >= 95 and Plasma- Glucse <=140 and Diastolic BP <=80 and BMI <24.9 and DPF<0.42 and Age <=40 THEN probability of infection =42

IF Pregnant>=2 and Plasma- Glucose >= 95 and Plasma- Glucse <=140 and Diastolic BP <=80 and BMI <24.9 and DPF<0.42 and Age <=40 THEN probability of infection = 50

Rule condition_ 3

IF Pregnant <=1 and Plasma- Glucse >140 and Diastolic BP <=80 and BMI <24.9 and DPF<0.42 and Age <=40 THEN probability of infection =44

IF PREGNANT>=2 AND PLASMA- GLUCSE >140 AND DIASTOLIC BP <=80 AND BMI <24.9 AND DPF<0.42 AND AGE <=40 THEN PROBABILITY OF INFECTION =52

Figure (5) shows how the result appears:

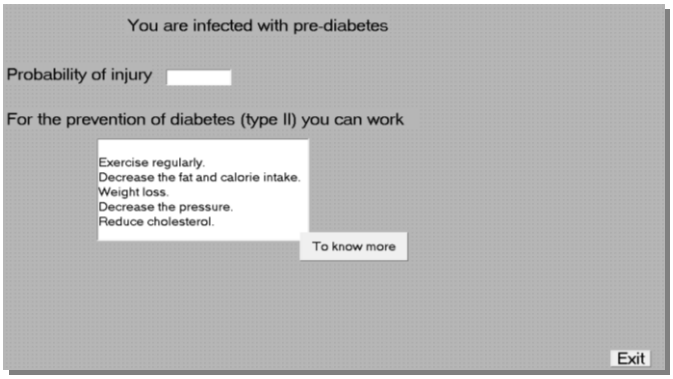


Fig. 5. Infected with Pre_diabetes

When you click on an icon To Know more, Figure (6) shows more information

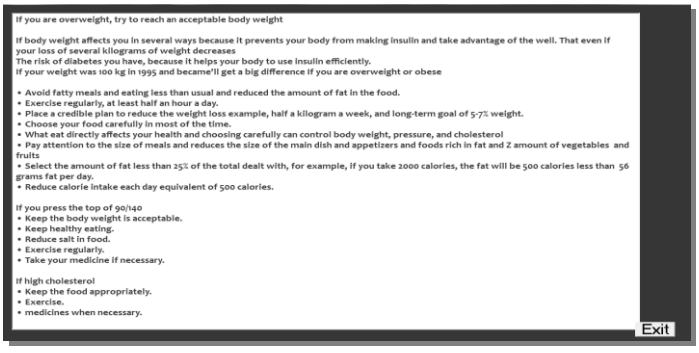


Fig. 6. More Information about Diabetes

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Configurations

The framework of project work is designed by (Delphi Ver.8) on a PC with the following configurations Core i7 laptop with 8GB of RAM for implementing the algorithms, running under Microsoft Windows 8_64 bits.

B. Testing Results

It was ascertained the validity of the results by comparing rules Decision Tree (C4.5) with the Indian diabetes database and as if the injury rate of less than 50 is not infected and most or equal to 50 is infected. The results which showed us are very good. It was divided into 70% of training and 30% of testing data. The result as the following:

TABLE 4

Performance	**Decision Tree**
Precision	0.8783
Recall(Sensitivity)	0.8783
Specificity	0.70886
Accuracy %	82.83582%
F_measure	0.8783
AUC	0.75996

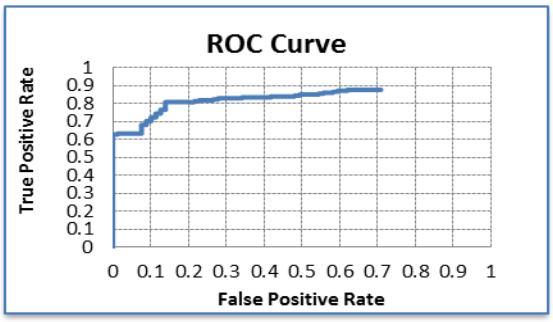


FIG. 7. the results

VI. CONCLUSION AND FUTURE WORK

The research in diabetic systems is important for both medical industry and diabetes patients. Reasoning techniques for diagnosing diabetes is urgently needed for helping both specialist doctors and patients. The abilities of inference, reasoning, and learning are the main features of any expert system. C4.5 algorithms in the medical field extracts distinctive concealed patterns from the medical data. They can be utilized for the examination of vital clinical parameters, expectation of different diseases, The proposed approach has shown that mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class we are trying to predict. Now we are developing expert system for detection and find probability of infection diabetes. In future we need develop a new approach and then compare it with the current methods in this paper and moreover these data analysis results can be used for further research in enhancing the accuracy of the prediction system in future.

REFERENCES

[1] Ibrahim M.Ahmeda, Marco Alfonseb, Mostafa Arefc ,Abdel-Badeeh M.Salemd, "Reasoning Techniques for Diabetics Expert Systems", Procedia Computer Science (813 – 820),2015.

[2] Ibrahim M. Ahmed, Abeer M. Mahmoud, "Development of an Expert System for Diabetic Type-2 Diet", International Journal of Computer Applications (0975 – 8887) Volume 107 – No.1, December 2014.

[3] Dr. Abdullah Al-Malaise Al-Ghamdi, Majda A.Wazzan, Fatimah M. Mujallid, Najwa K.Bakhsh, "An Expert System of Determining Diabetes Treatment Based on Cloud Computing Platforms", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2011.

[4] Ibrahim M.Ahmed, Abeer M.Mahmoud, Mostafa Aref, Abdel-Badeeh M.Salem, "A study on Expert Systems for Diabetic Diagnosis and Treatment", Recent Advances in Information Science, ISBN: 978-960-474-304-9,2012.

[5] P. M. Beulah Devamalar, V. Thulasi Bai, andSrivatsa S. K., "An Architecture for a FullyAutomated Real-Time Web-Centric Expert System", World Academy of Science, Engineering and Technology, 2007.

[6] Cindy.M, Jay. Sand Frank. S,"Toward Case Based Reasoning For Diabetes Management", Computational Intelligence Journal, Volume 25, Number 3,9p 165-179, 2009.

[7] Matthew Wiley and Razvan Bunescu. "Emerging Applications for Intelligent Diabetes Management Cindy Marling", Association for the Advancement of Artificial Intelligence, 2011.

[8] Wioletta SZAJNAR and Galina SETLAK., "A concept of building an intelligence system to support diabetes diagnostics", Studia Informatica, 2011.

[9] M.Kalpana and A.V Senthil Kumar,"Fuzzy Expert System for Diabetes using Fuzzy Verdict Mechanism", Int. J. Advanced Networking and Applications Volume: 03, Issue: 02, Pages: 1128-1134 , 2011.

[10] Sanjeev Kumar and Babasaheb Bhimrao, "Development of knowledge Base Expert System for Natural treatment of Diabetes disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 3, 2012.

[11] N.Nnamoko, F.Arshad, D.England and J. Vora, "Fuzzy Expert System for Type 2 Diabetes Mellitus (T2DM) Management Using Dual Inference Mechanism", AAAI Spring Symposium, p 67-70, 2013.

[12] Margret Anouncia S., Clara Madonna L., Jeevitha P.and Nandhini R.,"Design of a Diabetic Diagnosis System Using Rough Sets, Cybernetics And Information Technologies" ,Volume 13, No 3, pp 124-139,Sofia ,2013.

[13] Abdelhak .M, Baghdad .A, and Sofia. B, "A Hybrid Decision Support System :Application On Healthcare ", Corr, 2013.

[14] C. Marling, S.Montani , I. Bichindaritz and Peter Funk, "Synergistic case-based reasoning in medical domains, Expert Systems with Applications", Volume 41, Issue 2, 1, Pages 249–259, February 2014.

[15] VelidePhani Kumar 1 and Lakshmi Velide2, "A Data Mining Approach For Prediction And Treatment Ofdiabetes Disease" , VelidePhani Kumar-et al., IJSIT, 3(1),073-0792014.

[16] M. Berry and M. Browne, "Lecture Notes in Data Mining", published by World Scientific, United States of America, 2006.

[17] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Second Edition, published by Elsevier, United States of America, 2006.

[18] M. L. Wong and K. S. Leung , "Data Mining Using Grammar Based Genetic Programming and Applications", published by Kluwer Academic, United States of America, 2002.

[19] T. S. Korting , "C4.5 algorithm and Multivariate Decision Trees", In proceeding on National Institute for Space Research, PP.(1-5), Brazil, tkorting@dpi.inpe.br, 2006.

[20] D. T. Larose, "Discovering Knowledge in Data", published by JohnWiley & Sons, United States of America, 2005.

[21] D. Benhaddouche and A. Benyettou, "Data mining by the decision tree and support vector machine (SVM)", In proceeding of 5th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications(SETIT), TUNISIA, PP. (1-5), (22-26) March 2009.

[22] S. Theodoridis and K. Koutroumbas, "Pattern Recognition", Second Edition, published by Elsevier, United States of America, 2003.

[23] Z. Markov and D. T. Larose, "Data Mining The Web", published by John Wiley & Sons, United States of America, 2007.

[24] R. Devi and V. Khemchandani, "Application of Data Mining Techniques For Diabetic Dataset", Proceedings of the 4th National Conference, Computing For Nation Development, Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi, February 25 – 26, 2010.

[25] J.R. Quinlan , "Induction of Decision Trees", In proceeding of Centre for Advanced Computing Sciences, New South Wales Institute of Technology, Australia, PP.(81-106), 2007.

[26] Sumathy, Mythili, P. Kumar, T. M. Jishnujit, and K.R. Kumar, "Diagnosis of Diabetes Mellitus based on Risk Factors", In proceeding of International Journal of Computer Applications, Vol. (10), No. (4), PP. (1-4), 2010.

[27] A. R. Webb, "Statistical Pattern Recognition", Second Edition, published by John Wiley & Sons, England, 2002.

[28] W. Szajnar and G. Setlak, "A Concept of Design Process of Intelligent System Supporting Diabetes Diagnostics", In proceeding of Methods and Instruments of Artificial Intelligence, PP. (168-178), 2010.