ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(14), DOI: 10.17485/ijst/2015/v8i14/72688, July 2015* ISSN (Online) : 0974-5645

Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques

C. Kalaiselvi[1,2*] and G. M. Nasira[3]

[1]Karpagam University, Coimbatore-641021, Tamilnadu, India; kalaic29@gmail.com
[2]Department of Computer Application, Tiruppur Kumaran College for Women, Tiruppur-641687, Tamilnadu, India
[3]Department of Computer Science, Chikkanna Govt Arts College, Tiruppur-641602, Tamilnadu, India; nasiragm99@yahoo.com

Abstract

Background: The heterogeneous, chronic diseases like heart diseases and cancer are commonly occur and increased nowadays in diabetic patients. Most of the people do not know the symptoms of these diseases and its chronic complications. **Objective:** The aim of this paper is to predict the diseases such as heart diseases and cancer in diabetic patients. The association between these diseases can be analyzed based on the factors that cause these diseases which include obesity, age, associated diabetic duration, and some other life style factors. **Methods:** This work consists of two stages. In the first stage, the attributes are identified and extracted using Particle Swarm Optimization (PSO) algorithm. In the second stage, ANFIS (Adaptive Neuro Fuzzy Inference System) with Adaptive Group based K-Nearest Neighbor (AGKNN) algorithm has been used to classify the data. **Findings:** The experimental results show a very good accuracy and signify the ANFIS with AGKNN along with feature subset selection using PSO. The performance is evaluated using performance metrics and proved this classifiers efficiency for the prediction of heart disease and cancer in diabetic patients. **Application/ Improvement:** This work demonstrates the diagnosis of diseases and its importance to predict it earlier. In future it can be implemented for other related diseases in medical data mining and healthcare.

Keywords: Classification, Data Mining, Disease Prediction, Feature Selection, Normalization

1. Introduction

Medical data mining has the ability to explore hidden patterns from the datasets obtained from the medical domain. The raw medical data are heterogeneous and voluminous and it needs to extracted and integrated in an organized manner to explore hidden patterns in the data. Diabetes mellitus is a metabolic defect in the body's ability to convert glucose to energy. When food is digested, it is converted into fats, protein, carbohydrates. Carbohydrates are converted to glucose in blood and are the main source of energy to the body. To transfer glucose into blood cells, the hormone called insulin is required and is produced by the organ pancreas fails to produce enough insulin causes diabetes. There are three main types of diabetes. They are Type1 diabetes, Type2 diabetes and Gestational diabetes. Type1 diabetes is mostly occurring in children. Type2 diabetes is called adult-onset diabetes. It is common in adults. Gestational diabetes is only in women during pregnancy. Diabetes mellitus cannot be healed fully but it can be controlled by using medicines such as insulin, food items. The diabetes mellitus causes nerve damages, kidney disease, and also cancer[1,2]. Many researches are going on this area and some of the statistics report shows that most of the people will die because of the high glucose level lead to either cancer or heart diseases. The effects are uncontrolled and over time lead to serious damage to many human body systems, especially for the nerves and blood vessels[3,4].

It is very common knowledge that patients with type2 diabetes have high risk of developing heart disease. But recent research reveals that there exist interrelationships

* *Author for correspondence*

among diabetes, heart disease and cancer. These interrelationships may seem coincidental and based only on the fact that these conditions share common risk factors of these diseases in particular[5,6]. The current research and continuing education activities will evaluate how type2 diabetes and cancer is associated and suggest recommendations to reduce cancer risk intersect and to care for patients who are at high risk of diabetes and heart disease[7,8]. The risk factors of heart diabetes among diabetes patients includes elevated serum triglycerides and decreases HDL cholesterol higher fasting insulin and etc. Research has found that diabetes, cancer and heart disease have common physiological associations. These associations overlap with one another and all remain active areas of research[9].

1.1 Diabetes and Cancer

A number of epidemiological studies investigated the association between diabetes and cancer and also risks have been found for certain types of cancers such as endometrial cancer, pancreatic cancer, liver cancer and etc. Various epidemiological studies have found that the association of diabetes and cancer in various population society. The major strength of association may depend on the diabetic duration and specific cancer types. The reason by which diabetes can cause cancer is uncertain. It is based on some important risk factors such as age, obesity and life style factors that may increase the chance of diabetes. A research found that there is considerable evidence linking type2 diabetes and cancer incidence based on some direct and indirect risk factors like glucose lowering therapies modify the risk of cancer[10,11]. There can be an association found with pancreatic and liver cancer and may reflect in reverse casualty and may lead to onset diabetes. The risk of endometrial cancer doubled in women with diabetes in every year. The various risks of breast cancer, colorectal cancer, bladder cancer and kidney cancer are about 20- 40% in people with diabetes. But there exists no consistent association with lung and ovarian cancers. A research indicates that there exists an association between type2 diabetes and cancer compared to type1 diabetes[12].

The production of insulin and insulin factor (IGF)–I may cause tumor development through cell proliferation the Glycosylated hemoglobin (HbA1c) level can be found through blood tests. The abnormality in glucose metabolism may result into an increased risk of cancers and people with HbA1c level>7% or above may have increased risk of cancer incidence with age and tobacco smoking habits. A recent meta-analysis reveals that elevated serum insulin or c-peptide levels are associated with certain cancer types. An elevated blood glucose levels called as hyperglycemia would give cancer cells a relative growth is one of the factor that cause an increased risk of cancer in diabetes. The elevated risk of cancer persists during longer period of follow-up up to 10 years after diabetes onset in cancer types such as pancreatic cancer[13]. A study indicates that diabetes can be induced by cancers such as liver and pancreatic cancer. Among diabetic patients an increased risk of death due to cardiovascular disease and cancer may have a longer latency period. Bias in cancer screening procedure among diabetic patients may lead to decreased or elevated risk of cancer incidence[14,15] .

1.2 Diabetes and Heart Disease

Various researches have been conducted to find the relationship between diabetes mellitus and heart diseases recently in diabetic patients. The diabetes is a group of metabolic disorders that affects life of human. A reinterpretation is required, so, the possibilities of heart diseases in diabetic patients are high[16,17]. Most of the diabetic patients are affected by metabolic and hormonal disorders. Some of the risk factors of heart diseases and cancer among diabetic patients are lack of physical in activity, alcohol, tobacco smoking and obesity[18,19]. The biological link mechanisms behind heart diseases and diabetes are hyperinsulinemia, hyperglycemia and inflammation in the blood[20–22]. Very high blood sugar damages blood vessels can lead to blockage. People with diabetes have two to four times risk of developing heart disease. High blood sugar causes blocks in leg vessels can cause pain and also impair circulation. Some of the symptoms of heart failure are problems inbreathing, swelling in the ankles, feet, legs, abdomen, and veins in neck[23]. People with heart failure can live longer and more active lives if the condition is diagnosed earlier and if they follow their treatment plans regularly. People having diabetes can have a chance of developing heart disease and stroke[24]. Several new techniques are used for diagnosing both heart diseases and diabetes. The machine learning techniques are one of the existing techniques which have a transparent diagnostic knowledge to diagnose diseases[25].

1.3 Organization of the Study

In this paper, the data mining techniques are used to diagnose the diseases like heart disease and cancer in diabetic patients. The relationship between diabetes, heart disease and cancer are examined by taking into account of age, associated diabetic duration, and other life style factors. Section 1 deals with Introductory part of the paper. In Section 2, the methods and materials used, data set description is discussed. The data collected from diabetic patients are stored in a database. From these database, features are selected, normalized and those selected features are given to ANFIS (Adaptive Neuro Fuzzy Inference System) with Adaptive Group based K-Nearest Neighbor (AGKNN) algorithms to classify normal and abnormal data. The performance is evaluated using performance metrics and proved this classifiers efficiency for the prediction of heart disease and cancer. The section 4 describes the sampling results and finally section 5 gives the conclusion of the paper.

2. Proposed Methodology

The entire process of classification and prediction of diseases is diagrammatically represented in the following Figure 1. The data collected from diabetic patients are stored in a database. From these databases, features are selected, normalized based on particle swarm optimization algorithm.

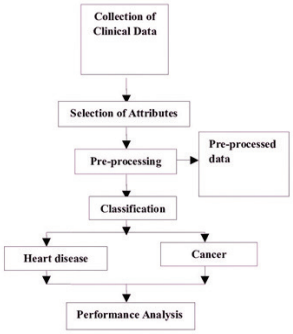

Figure 1. Framework of the proposed system.

2.1 Feature selection

The feature selection is done from the large number of attributes to select best attributes to achieve high performance. The attributes taken to predict heart disease and cancer in diabetic patients are depicted in the Table 1 below.

Table 1. Attributes used for classification

Attribute	Description
Age	Age of the patient
Family Heredity	Previous history
BP	Blood pressure
HBA1C	HbA1c level
Tot Cho	Total cholesterol level
Trigly	Triglycerides
LDL	LDL cholesterol
HDL	HDL cholesterol
Pul	Pulse
Obese	Obesity
ST	Smoking/ Tobacco
Herid	Heridity
Dur	Diabetic duration

Feature selection is used to select a subset of relevant features for building robust and best learning models from the data available. These features are used in machine learning processes and provide better understanding of the data by selecting important features within the data.

2.2 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a kind of population-based optimization algorithms. It is modeled using the simulation of social behavior of birds in a flock. This algorithm is initialized with particles and it searches optimal value through updating process in generations. Each of the particles is flown a group of random search space where the position is calculated by adjusting and is based on its distance from its own personal best position and the distance from the best particle of the swarm. The performance of each particle can be measured using a fitness function that depends on the optimization problem. The data normalization can also be done for the selected attributes. Each particle 'i' is an n-dimensional search space, R^n and maintain the following function: x_i the current position of ith particle (x-vector), p_i the personal best position of ith particle (p-vector), and v_i, the current velocity of ith particle (v-vector). The best position of the particle *m* is the best position that the particle has visited so far. If g denotes the fitness function, then the personal best of *m* at a time step t is updated as:

$$p_m(t+1)=\{(p_m(t) \text{ if } g[(x]_m(t+1)) \geq g(t)))x_m(t+1) \text{ if } g(x_m(t+1)) < g\,(p_m(t)) \quad (1)$$

If the position of the global best particle is denoted by gbest, then

$$gbest \in \{p_1(t),p_1(t),....p_m(t)\}=\min\{f[(p]_1(t)),f[(p]_2(t)),.... f[(p]_m(t))\} \quad (2)$$

The updates of velocity are calculated as a linear combination of position vectors and velocity vectors. Thus, the velocity of particle *m* is updated and the position of particle i is updated by the following equations.

$$v_m(t+1) = w.v_m(t)+c_1r_1(p_m(t)-x_m(t))+ c_1r_1(gbest-x_m(t)) \quad (3)$$

$$x_m(t+1) = x_m(t) + v_m(t+1) \quad (4)$$

In the formula, w is the inertia weight, c_1 and c_2 are the acceleration constants, r_1 and r_2 are random numbers in the range [0,1] and V_{i1} must be in the range$[-V_{max}, V_{max}]$, where V_{max} is the maximum velocity. Kernel methods are used in general due to the grown of popularity of the Support Vector Machines. These are linear classifiers and regressors and are able to perform non-linear classification and regression in their input space. Here Radial Bias Function Kernel is used and it is expressed as:

$$\mathbf{RBF} = \exp\left(\frac{1}{2\sigma^2\|\mathbf{x} - \mathbf{x}_i\|^2}\right) \quad (5)$$

2.3 Adaptive Neuro Fuzzy Inference System (ANFIS)

Adaptive Neuro Fuzzy Inference System is the combination of fuzzy inference system and learning power of artificial neural network. It incorporates the best features of the fuzzy systems and the neural network. The algorithms such as gradient descent and back propagation are used to train the artificial neural network systems by regulating the membership functions and weights. The first order fuzzy inference system based on if then rules is used in ANFIS architecture.

Rule 1: If $(x \text{ is } A_1)$ and $(y \text{ is } B_1)$
Then $(f_1 = p1x + q1y + r1)$ (6)

Rule 2: If $(x \text{ is } A_2)$ and $(y \text{ is } B_2)$
Then$(f_2 = p2x + q2y + r2)$ (7)

Where x and y are the inputs.A_i and B_i are the fuzzy sets f_i are the outputs within the fuzzy region specified by the fuzzy rules.p_i, q_i and r_i are the design parameters that are determined in the training process. The architecture of ANFIS to implement these two rules is shown in Figure 2, in which a circle indicates a fixed node,

In the first layer, all the nodes are adaptive nodes. The outputs from the layer 1 are the fuzzy membership grade of the inputs, which are given by:

$$O^1_i = \mu A_i(x),\ i = 1,2 \quad (8)$$

$$O^1_i = \mu B_{i-2}(y),\ i = 3,4 \quad (9)$$

Where $\mu A_i(x)$, $\mu B_{i-2}(y)$ adopt any fuzzy membership function. For example, if the bell shaped membership function is employed, $\mu A_i(x)$ is given by:

$$\mu A_i(x) = \frac{1}{1 + \left\{\left(\frac{x - c_i}{a_i}\right)^2\right\}^{b_i}} \quad (10)$$

Where A_i, B_i and C_i are the parameters of the membership functions. In the second layer, the nodes are fixed nodes. They are labeled with m and it indicates a simple multiplier. The output of this layer can be represented as:

$$O^2_i = w_i = \mu A_i(x)\mu B_i(y)\ i = 1,2 \quad (11)$$

Which are so-called firing strengths of the rules.

In the third layer, the nodes are fixed nodes and are labeled with N. They play a normalization role to the firing strengths from the previous layer. The output can be represented as:

$$O^3_i = \bar{w}_i = \frac{w_i}{w_1 + w_2}\ i = 1,2 \quad (12)$$

And which are the called as normalized firing strengths. In the fourth layer, the nodes are adaptive nodes. The output of each node in this layer is simply the product of the normalized firing strength and a first order polynomial. The outputs of this layer are given by:

$$O^4_i = \bar{w}_i f_i = \bar{w}_i(p_i x + q_i y + r_i)\ i = 1,2 \quad (13)$$

In the fifth layer, there is only one single fixed node labeled with *S* and this performs the summation of all incoming signals. Thus, the overall output of the model is given by:

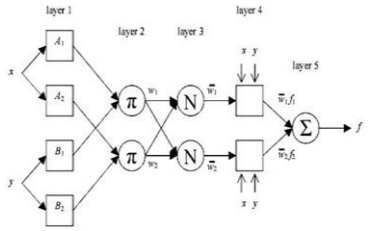

Figure 2. ANFIS architecture.

The above diagram (Figure 2) depicts the typical model of Adaptive Neuro Fuzzy Inference System (ANFIS) with 5 layers. In each layer circle depicts fixed node, square shows that adaptive node and each layer is used for different purpose. To obtain the preferred performance number, type, parameter and rules of fuzzy membership functions are used and it is selected based on the trial and error method. Although the better performances are achieved, these parameters are hard to use in some situations. To overcome this problem ANFIS is trained to get optimal premise and consequent parameters.

It can be observed that there are two adaptive layers in this ANFIS architecture. They are the first layer and the fourth layer. The first layer contains three modifiable parameters $\{a_i, b_i, c_i\}$ and they are related to the input membership functions and are called as premise parameters. The fourth layer contains three modifiable parameters $\{p_i, q_i, r_i\}$, and pertaining to the first order polynomial called as consequent parameters. In ANFIS assume the model structure is used to relate the membership function, rules input and output in cyclic process. The collection of input and output data which is used to train the fuzzy inference model by regulating the membership function based on the trial and error method.

2.4 Adaptive Group-based K-Nearest Neighbor Algorithm(AGKNN)

The AGKNN algorithm is used for recording the set of observations or training set. The k number of inputs and the desired outputs that should be computed in network, and then the error (difference between actual and expected results) is calculated. The data are classified using ANFIS simultaneously in each group with random value of k and compare the results. The k value is changed according to compared results. If the result is reliable then group size and k value are unchanged. If the result is similar emerge the group and reduce the k value. If the results are different then increase the k value. The groups should increase when the two opposing are offset.

$N_g(i)$ - Number of training group during i[th] data processed. $C_j(i)$ - Categorization result of i[th] document by j[th] group.

The Adaptive training group is determined as

$$N_g(i+1)=\begin{cases} N_g + int\left(\frac{1}{N_g}\sum_{j=1}^{N_g}\left(C_j(i)-\bar{C}\right)^2\right), \sum_{j=1}^{N_g}\left(C_j(i)-\bar{C}\right)^2 > \bar{C} \\ N_g \frac{1}{C} < \sum_{j=1}^{N_g}\left(C_j(i)-\bar{C}\right)^2 \le \bar{C} \\ N_g - int\left(\frac{1}{\bar{C}}\right)\sum_{j=1}^{N_g}\left(C_j(i)-\bar{C}\right)^2 \le \frac{1}{\bar{C}} \end{cases} \quad (14)$$

The samples data's are suggested to the variance of different groups by adjusting the grouping situation. If the variance of the grouping data is higher than the threshold then the categorization results are inaccurate. The reason for inaccurate results is more groups are required for final decision. If the variance is low means then the sample groups are merged without any disputes in classification results. Threshold value can be calculated as a by using lower and higher bound ($\frac{1}{C}$ and C).

The k value can be calculated adaptively as:

$$k=\begin{cases} \gamma_i \frac{1}{N_g}\sum_{j=1}^{N_g}(C_j(i)-C)^2 < 1 \\ \gamma_i + int\left(\frac{1}{n}\sum_{j=1}^{N_g}(C_j(i)-C)^2\right) \\ \quad \cdot \frac{1}{N_g}\sum_{j=1}^{N_g}(C_j(i)-C)^2 \ge 1 \end{cases} \quad (15)$$

γ_i- Random initial value of k. The random value can be tested by the system to check whether it is suitable or not. To obtain the exact categorization results the value of k should be adjusted.

The k value can be tested by the system to check whether it is suitable for group or not and can be set by algorithm with the help of training set. Based on the variance of categorization results the groups are adjusted by different groups in real time. Runtime complexity for n elements are $3n$-1 and computational complexity is calculated as a

$$T = N_g \,.\, k \,.(3N_g - 1) \quad (16)$$

Training of network i.e. error correction is stopped when the value of the k has become sufficiently small and as desired in the required limits[4]. Total error for p^{th} observation of data set and j^{th} neuron in the output layer can be computed as:

$$E_k = t_k - y_k \quad (17)$$

Where, t_k represents the desired target output, y_k represents the predicted from the system and E_k error correction. The problems of k nearest neighbor is reduced in adaptive group based KNN. When the AGKNN is compared with the traditional KNN the proposed algorithm shows the higher efficiency.

3. Simulation Results

To predict heart disease and cancer among diabetic patients the diabetic dataset is collected from one of the leading diagnostic centre as the dataset with important attributes of diabetes are not publicly available. This work was designed as a prospective, single centre study. Totally the around 500 patient's data are collected. The data is initially normalized to get the values in a range.

The formula for normalization is:

$$X = X - X_{min} / X_{max} - X_{min} \quad (18)$$

After normalization the data set is divided into two. 30% of the data is used for testing and 70% of data is used for training. The data is classified using ANFIS with AGKNN. It uses member functions for each of the input and predicts the results. The experiment is done using MATLAB 7.14 and the simulation results are obtained and grouped into four categories. Category1 is for diabetes and heart disease. Category2 is for diabetes and cancer disease. Category3 is for heart disease and cancer in diabetic patients. The remaining others placed in the Category0 with only diabetes. The effectiveness of models was tested using different methods. Confusion matrix is one of this. The purpose of this is to determine which model gives the highest percentage of correct predictions for diagnosing diabetic patients with heart disease. According to our experiment 70% of people are under the risk of heart disease and 10% of people are under the risk of cancer. Only 0.4% of people are under the risk of both heart disease and cancer when they are having diabetes for long time duration based on lifestyle modifications. The sampling results of the proposed methodology are shown in Table 2.

Table 2. Sampling results of proposed methodology

Proposed ANFIS with AGKNN	Accuracy	Sensitivity	Specificity
Diabetic / Heart disease	98	98.99	98.98
Diabetic / Cancer disorder	96.3	96.6	96.7
Diabetes	99.8	99.8	99.7

4. Conclusion

In this research work, to improve the classification accuracy and to achieve better efficiency a new approach like Adaptive Neuro Fuzzy Inference System (ANFIS) is proposed. ANFIS is used to train the neural network. The input nodes in neural network are constructed based on the input attributes. The hidden nodes are used to classify given input based on the training dataset with the help of AGKNN. The experimental results show that the classification accuracy is better than existing approaches. The proposed approach gives higher efficiency and reduces complexity. The algorithm performs well and classifies the dataset well compared to traditional methods. The proposed work reduces the cost for different medical tests and helps the patients to take precautionary measures well in advance.

5. References

1. From AM, Leibson CL, Bursi F et al. Diabetes in heart failure: prevalence and impact on outcome in the population. Am J Med. 2006; 119:591–9.
2. Kalaiselvi C, Nasira GM. A new approach for the diagnosis of diabetes and cancer using ANFIS. World Congress on computing and communication technologies, WCCCT-IEEE conference; 2014 Feb. p. 188–90.
3. Weiderpass E, Gridley G, Persson. Risk of endometrial and breast cancer in patients with diabetes mellitus. Int J Cancer. 1997 May 2; 71(3):360–3.
4. Kalaiselvi C Nasira GM. Classification and Prediction of heart disease from diabetes patients using hybrid particle swarm optimization and library support vector machine algorithm. International Journal of Computing Algorithm (IJCOA). 2015 Mar; 4:1403–7.
5. Bakris GL. Preclinical diabetic cardiomyopathy: prevalence, screening and outcome. Internal Medicine University Department; 2009. p. 1–27.
6. Ali MK. Diabetes and coronary heart disease: Current perspectives. Indian journal of medical research. 2010; 132(5):584–97.

7. Collins K, MS, RD. The cancer, diabetes and heart disease Link. Today's Dietitian. 2013 Mar; 15(3):46.
8. Kalaiselvi C, Nasira GM. A Novel Approach for the Diagnosis of Diabetes and Liver Cancer using ANFIS and Improved KNN. Research Journal of Applied Sciences, Engineering and Technology. 2014; 8(2):243–50.
9. Johnson JA, Cartensen B. Diabetes and Cancer (1): evaluating the temporal relationship between type2 diabetes and cancer incidence. Diabetologia. 2012; 55(6):1607–18..
10. Travier N, Jeffreys M, Brewer N, Wright CS. Association between Glycosylated Hemoglobin and Cancer risk. Ann Oncol. 2007 Aug; 18(8):1414–9.
11. Giovannucci E, MD, SCD, Harlan DM. Diabetes and cancer. A consensus report. 2010 Jul.
12. La Vecchia C, Nagri E, Decarli A, Franceschi S. Diabetes mellitus and the risk of primary liver cancer. Int J Cancer. 1997 Oct 9; 73(2):204–7.
13. Everhart J, wright D. Diabetes mellitus as a risk factor for pancreatic cancer- A meta analysis. Jama. 1995 May 24-31; 273(20):1605–9.
14. Vigneri1 P, Frasca F, Sciacca L, Pandini G. Diabetes and cancer. Endocrine Related cancer. 2009; 16:1103–23.
15. Kalaiselvi C, Nasira. Fuzzy and Rough Set Theory Based Gene Selection Method. IJSR. 2014 Oct; 3(10):764–7.
16. Shi Y, Eberhart R. A modified particle swarm optimizer. The Proceedings of the IEEE International Conference on Evolutionary Computation; 1998; Piscataway, NJ. p. 69–73.
17. Bianchini F, Kaaks R, Vainio H. Overweight, Obesity and Cancer risk. Lancet Oncol. 2002 Sep; 3(9):565–74.
18. M.Harris M. The role of primary health care in preventing the onset of chronic disease, with a particular focus on the lifestyle risk factors of obesity, tobacco and alcohol. Centre for Primary Health Care and Equity, UNSW. 2008; 1–21.
19. Anderson KM. Correlation of regional cardiovascular disease mortality in India with lifestyle and nutritional factors. International J Cardiol. 2006; 108: 291–300.
20. Alessandra Saldanha de. Impact of Diabetes on Cardiovascular Disease: An Update. International Journal of Hypertension. 2013; 653789.
21. Bertoluci MC, Pimazoni A. Diabetes and cardiovascular disease: from evidence to clinical practice. Diabetology and Metabolic syndrome. 2014; 6:58.
22. Anandhapadmanabhan KR, Parthiban G. Prediction of chances Diabetic retinopathy using data mining classification techniques. IJST. 2014 Oct; 7(10).
23. Ronald M. Witteles. Insulin resistant cardiomyopathy. Journal of American College of Cardiology. 2008 Jan 15; 51(2):93–102.
24. Santhanam T, Ephzibah. Heart disease prediction using hybrid genetic fuzzy model. IJST. 2015 May; 8(9).
25. Nagarajan S, Chandrasekaran RM. Design and implementation of expert clinical system for diagnosing diabetes using datamining techniques. IJST. 2015 Apr; 8:771–6.