International Journal of Bio-Science and Bio-Technology
Vol.7, No.5 (2015), pp.135-150
http://dx.doi.org/10.14257/ijbsbt.2015.7.5.13

GA_J48graft DT: A Hybrid Intelligent System for Diabetes Disease Diagnosis

Dilip Kumar Choubey[1] and Sanchita Paul[2]

[1,2]CSE, BIT, Mesra, Ranchi, India
dilipchoubey_1988@yahoo.in
sanchita07@gmail.com

Abstract

Diabetes is a condition in which the amount of sugar in the blood is higher than normal. Classification systems have been widely used in the medical domain to explore patient's data and extract a predictive model or set of rules. The prime objective of this research work is to facilitate a better diagnosis (classification) of diabetes disease. There are already several methodologies which have been implemented on classification for the diabetes disease. The proposed methodology implemented work in 2 stages: (a) In the first stage Genetic Algorithm (GA) has been used as a feature selection method on Pima Indian Diabetes Dataset (PIDD) (b) In the second stage, J48graft Decision Tree (J48graft DT) has been used for the classification and prediction on the selected feature. Early diagnosis of any disease with less cost is preferable. Diabetes is also one of such diseases. GA is noted to reduce not only the storage capacity, cost and computation time of the diagnostic process, but the proposed approach also improved the ROC of classification. The experimental results obtained classification accuracy (74.7826%) and ROC (0.786) show that GA and J48graft DT can be successfully used for the diagnosing of diabetes disease.

Keyword: *PIDD, GA, J48graft DT, Diabetes Disease Diagnosis, Feature Selection, Classification*

1. Introduction

Diabetes is a chronic disease and a major public health challenge worldwide. Diabetes happens when a body is not able to produce or respond properly to insulin which is needed to maintain the rate of glucose. Diabetes can be controlled with the help of insulin injections, a controlled diet (changing eating habits) and exercise programs, but no whole cure is available.

GA were introduced by John Holland in the 1970 at University of Michigan (US). GA is an adaptive population based optimization method which is inspired by Darwin's theory about survival of the fittest [10]. GA mimics the natural evolution process gave the Darwin *i.e.*, in GA the next population is evolved through simulating operators of selection, crossover and mutation. John Holland is known as the father of the original genetic algorithm who first introduced these operators in [16]. These operators were later improved by Goldberg [13] and Michalewicz [18]. In this paper, GA has been used as a feature selection method in which among 8 attributes, 4 attributes have been selected.

The main purpose of feature selection method is to reduce the number of features used in classification while maintaining acceptable ROC, and classification accuracy. Reducing the number of features (dimensionality) is important in statistical learning. With the help of the feature selection process we can save storage capacity, computation time (shorter training time and test time), computation cost and increases Classification rate, comprehensibility.

ISSN: 2233-7849 IJBSBT

A decision tree is graph or tree based classifier used for the classification of uncertain data or huge data. The methodology of decision tree is made up of a unique root node (one root), a set of terminal nodes called leaves (number of leaves), a set of decision nodes (internal nodes), and branches. A node consists of one attribute, one branch refers to a chain of nodes, Its path is from root to leaf. In decision tree oval shaped represents internal nodes, and rectangular represents leave nodes. The procedures of creating a decision tree algorithm are summarized: (a) The available feature set provides input to the algorithm and decision tree is output. (b) The decision tree possesses leaf node, which expresses class label related with classes being classified. (c) Branches of tree shows every single possible value of a node from where they are originated. In Data mining there are various classifiers used for the generation of classification tree such as using ID3, CART, C4.5 and J48graft etc. J48graft algorithm is one of the preparing decision trees. J48graft algorithm is used for generating grafted C4.5 algorithm [8]. J48graft is an algorithm having aim to increase the probability of classifying rightly the instances. This algorithm creates only single tree and it reduces prediction error [17] or to frequently improve predictive accuracy. J48graft DT has been used for the classification of the diabetes disease diagnosis.

The rest of the paper is organized as follows: Related work is presented in Section 2, Proposed methodology is discussed in Section 3, Results and discussion are devoted to Section 4, Conclusion and future work are discussed in Section 5.

2. Related Work

Kemal Polat, Salih Gunes [2] stated Principal component analysis (PCA) and Adaptive neuro – fuzzy inference system (ANFIS) to improve the diagnostic accuracy of diabetes disease in which PCA is used to reduce the dimensions of diabetes disease datasets features and ANFIS diagnosis of diabetes disease mean apply classification on that reduced features of diabetes disease datasets. Manjeevan Seera, Chee Peng Lim [3] introduced a new way of classification of medical data using hybrid intelligent system. The methodology implemented here is based on the hybrid combinatorial method of Fuzzy max-min based neural network and classification of data using Random forest regression tree. The methodology is implemented on various datasets including Breast Cancer and PIDD and performs better as compared to other existing techniques. Esin Dogantekin, Akif Dogantekin, Derya Avci, Levent Avci [1] used Linear discriminant analysis (LDA) and ANFIS for diagnosis of diabetes. LDA is used to separate feature variables between healthy and patient (diabetes) data, and ANFIS is used for classification on the result produced by LDA. The techniques used provide good accuracy then the previous existing results. So, the physicians can perform very accurate decisions by using such an efficient tool.

H. Hasan Orkcu, Hasan Bal [4] compares the performance of back propagation and GA for the classification of data. Since Back propagation is used for the efficient training of data in Artificial neural network (ANN) but contains some error rate, hence GA is implemented for the binary and real-coded so that the training is efficient and more number of features can be classified. Muhammad Waqar Aslam, Zhechen Zhu, Asoke Kumar Nandi [7] implemented an expert system for the classification of diabetes data using Genetic programming (GP). The technique implemented here consists of three stages: the first stage includes feature selection using t-test and kolmogorov-smirnov test and kulback-Leibler divergence test, the next stage uses GP which is used for the non-linear combination of selected attributes from the first stage. At the final stage the generated features using GP is compared with K-nearest neighbor (KNN) and SVM. The classification is done on PIDD consists of 768 instance values in the dataset and 8 attributes and one output variable (class variable) which have either a value ‘1’ or ‘0’ available in the dataset. The selected features are then used for the classification of diabetes patients with high accuracy of classification.

Pasi Luukka [5] used Fuzzy entropy measure, similarity classifier for the better classification of diabetic disease. Fuzzy entropy used as a feature selection and similarity classifier used for the classification on that selected features. The techniques used provide computation time much lower, enhance classification accuracy by the process to reduce noise, reduced computational cost, more transparent and comprehensible by removing insignificant features from the dataset. Kemal Polat, Salih Guneh, Ahmet Arslan [12] proposed uses a new approach of a hybrid combination of Generalized discriminant analysis (GDA) and Least square support vector machine (LS – SVM) for the classification of diabetes disease. Here the methodology is implemented in two stages: in the first stage pre-processing of the data is done using the GDA such that the discrimination between healthy and patient disease can be done. In the second stage LS-SVM technique is applied for the classification of Diabetes disease patient’s. The methodology implemented here provides accuracy about 78.21% on the basis of 10 fold-cross validation from LS-SVM and the obtained accuracy for classification is about 82.05%.

E.P. Ephzibah [27] used GA and Fuzzy Logic (FL) for diabetes diagnosis in which GA has been used as a feature selection method and FL is used for classification. The used techniques improve the accuracy and reduced the cost.

3. Proposed Methodology

Here, the proposed approach is implemented and evaluated by GA as a feature selection and J48graft DT for classification on PIDD from UCI repository of machine learning databases.

The proposed system of block diagram and the next proposed algorithm is shown below:

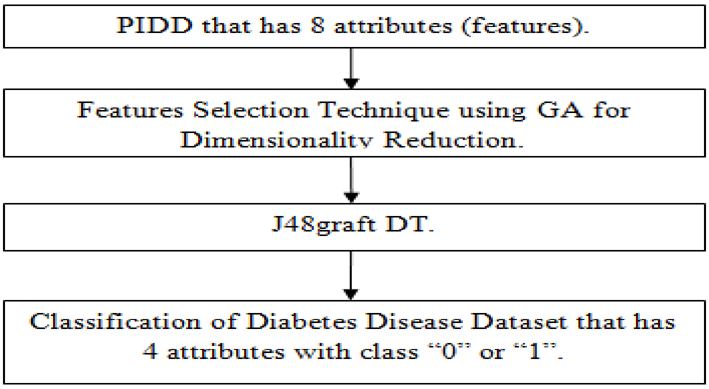

PIDD that has 8 attributes (features).
Features Selection Technique using GA for Dimensionality Reduction.
J48graft DT.
Classification of Diabetes Disease Dataset that has 4 attributes with class “0” or “1”.

Figure 1. Block diagram of proposed system

3.1. Proposed Algorithm

Step1: Start

Step2: Load Pima Indian Diabetes Dataset

Step3: Initialize the parameters for the GA

Step4: Call the GA

Step5.1: Construction of the first generation

Step5.2: Selection

```
        While stopping criteria not met do
  Step5.3: Crossover
  Step5.4: Mutation
  Step5.5: Selection
          End
  Step6: Apply J48graft DT Classification
  Step7: Training Dataset
  Step8: Calculation of error and accuracy
  Step9: Testing Dataset
  Step10: Calculation of error and accuracy
  Step11: Stop
```

The proposed approach works in the following phases:

1. Take PIDD from UCI repository of machine learning databases.
2. Apply GA as a Feature selection on PIDD.
3. Do the Classification by using J48graft DT on selected features in PIDD.

3.1.1. Used Diabetes Disease Dataset

The PIDD was obtained from the UCI Repository of machine learning databases [14]. The same dataset used in the reference [1-7] [9] [11] [15] [19] [21-27]. The National Institute of Diabetes and Digestive and Kidney Diseases originally owned this data and received in 9 May 1990. All patients in this database are Pima Indian Woman at least 21 years old and living near Phoenix, Arizona, USA. The features of this database are given in below:

- Number of instances: 768
- Number of attributes: 8
- Attributes:
  1. Number of times pregnant
  2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  3. Diastolic blood pressure (mm Hg)
  4. Triceps skin fold thickness (mm)
  5. 2 – hour serum insulin (mu U/ml)
  6. Body mass index (weight in kg/ (height in m)2)
  7. Diabetes pedigree function
  8. Age (years)
  9. Class variable (0 or 1)

There are eight all numeric - valued attributes and one output variable (class variable) which has either a value ‘1’ or ‘0’.

Class distribution (class value 1 is interpreted as “tested positive for diabetes”):

Class value	Number of instances
0	500 (65.1%)
1	268 (34.9%)

3.1.2. GA

The GA is a repetitive process of selection, crossover and mutation with the population of individuals in each iteration called a generation. Each chromosome or individual is encoded in a linear string (generally of 0s and 1s) of fix length in genetic analogy. In search space, first of all, the individual members of the population are randomly initialized. After initialization each population member is evaluated with respect to objective function being solved and is assigned a number (value of the objective function) which represents the fitness for survival for the corresponding individual. The GA maintains a population of a fix number of individuals with the corresponding fitness value. In each generation, the more fit individuals (selected from the current population) go in the mating pool for crossover to generate new off-springs, and consequently individuals with high fitness are provided more chance to generate off-springs. Now, each new offspring is modified with a very low mutation probability to maintain the diversity in the population. Now, the parents and off-springs together forms the new generation based on the fitness which will treated as parents for next generation. In this way, the new generation and hence successive generations of individual solutions is expected to be better in terms of average fitness. The algorithms stops forming new generations when either a maximum number of generations has been formed or a satisfactory fitness value is achieved for the problem.

The standard pseudo code of GA is given in Algorithm 1

3.1.2.1. Algorithm 1 GA

```
Begin
r = 0
Randomly initialize individual members of population P(r)
Evaluate fitness of each individual of population P(r)
while termination condition is not satisfied do
r = r+1
            selection (of better fit solutions)
       crossover (mating between parents to generate off-springs)
       mutation (random change in off-springs)
end while
Return best individual in population;
```

In Algorithm 1, r represents the generation counter. In Algorithm 1, initialization is done randomly in search space and corresponding fitness is evaluated based on objective function. After that GA algorithm requires a cycle of three phases: selection, crossover and mutation.

In medical world, If we have to be diagnosed any disease then there are some tests to be performed. After getting the results of the tests performed, the diagnosed could be done better. We can consider each and every test as a feature. If we have to do a particular test then there are certain set of chemicals, equipment, may be people, more time required which can be more expensive. Basically, feature selection informs whether a particular test is necessary for the diagnosis or not. So, if a particular test is not required that can be avoided. When the number of tests gets reduced the cost that is required also gets reduced which helps the common people. So, here that is why we have applied GA as a feature selection by which we reduced 4 features among 8 features. So from the above it is clear that GA is reducing the cost, storage capacity, and computation time by selected some of the feature.

3.1.3. J48graft DT

A Decision tree is a recursive form of the tree consisting of nodes and leaves on the basis of which a decision is taken from the dataset. It is constructed on the basis of the attribute with the highest normalized information gain ratio which will be taken as the root node and the dataset is split based on the root element values. Again the information gain ratio is calculated for all the sub nodes individually and the process is repeated until the prediction is completed. V. K. Pachghare [20] have already used the same procedure for the calculation of information gain ratio.

Algorithms

The algorithm for the J48graft DT are as follows:

J48graft Decision Tree (T, F, O)
T- Training Dataset Features
F- Pima Indian Input Dataset features selected from GA
O- Output classified features

1. Initially an empty Tree ‘t’ $t \rightarrow \varphi$

2. For each of the selected training features present in the dataset

3. Compute the calculation of information gain ratio:

   (a) On the assumption that S is the collection of data samples, its size is s. They respectively belong to category C_i (i=1,2…….,m). Supposing s_i is sample collection of category $C_{i,}$ then the information entropy of S is defined as follows:

   $$H(s) = H(£_1, \qquad £_2, £_3, \dots \dots \dots . £_n) = -\sum_{i=1}^{m} p_i \log_2 p_i \qquad (1)$$

   p_i is the probability of sample belonging to category C_i. we use $|s_i|/|s|$ to estimate it.

   (b) On the assumption that attribute A has v different values $\{a_{1,} a_2 \dots\dots\dots\dots\dots a_v\}$. We use attribute A to divide S into v subsets $\{S_1, S_{2,} \dots\dots\dots S_v\}$. $S_{j\,=}$ {x |x €S & x.A = a_j}. If we select attribute A as testing attribute (best splitting attribute), then these subsets correspond to branches are generated by collection S. Supposing s_{ij} is the sample collection which belongs to category C_i in subset S_j, then the entropy of S divided by attribute A can be given by the following formula:

   $$H(A, S) = \sum_{j=1}^{v} \frac{(s_{1j} + s_{2j} + s_{3j} + \ldots + s_{mj})}{s} * H(s_{1j}, s_{2j}, s_{3j}, \ldots, s_{mj}) \qquad (2)$$

   $(s_{1j\,+}\, s_{2j,\,+}\, s_{3j,}\, \dots\dots\dots\, {}_{+}s_{mj})$s express the weight of j^{th} subset. It equals the number of subsets (the value of attribute A is $a_{j)}$ divided by the total number (s) of S. The smaller the entropy is, the higher the division level of subset is. The expected information of subset Qj is given by the following formula:

$$H(£_{1j}, £_{2j}, £_{3j,..............}£_{nj}) = -\sum_{i=1}^{m} p_{ij} \log_2 p_{ij} \quad (3)$$

P_{ij} is the probability os samples S_j belonging to category $C_{i,}$ $p_{ij} = |s_{ij}| / |s_j|$.

(c) On the branch of attribute A, the information gain can be obtained from the following formula:

$$\mathrm{Gain}\ (A, S) = H(S) - H\ (A, S) \quad (4)$$

(d) For sample set S, on the assumption that attribute A has v different discrete values. Then the partition information divided by attribute A can be given by the following formula:

$$\mathrm{SplitInfo}\ (A, S) = -\sum_{j=1}^{v} \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (5)$$

(e) The following gain ratio of attribute A is as follows:

$$\mathrm{GainRatio}(A, S) = \frac{\mathrm{Gain}(A, S)}{\mathrm{SplitInfo}\ (A, S)} \quad (6)$$

4. Select the attribute with highest information gain
5. Update the tree ‘t’ the attribute as the root node to ‘t’.
6. Remove the attribute from the relation set
7. End

4. Results and Discussion

The work was implemented on i3 processor with 2.30GHz speed, 2 GB RAM, 320 GB external storage and software used JDK 1.6 (Java Development Kit), NetBeans 8.0 IDE and have done the coding in Java. For the computation of J48graft DT, various parameters of weka library are used.

In Experimental studies we have partition 70-30% for training & test of GA _ J48graft DT system for diabetes disease diagnosis. We have performed the experimental studies on PIDD mentioned in Section 3.1.1. We have compared the results of our proposed system *i.e.*, GA _ J48graft DT with the previous results reported by earlier methods [3].

The parameters for the Genetic Algorithm for our task are:

Population size	20
Number of generations	20
Probability of crossover	0.6
Probability of mutation	0.033
Report frequency	20

Random number seed 1

As per the Table no.3 we may see that by applying the GA approach, we have obtained 4 features among 8 features. This means we have reduced the cost to s(x) = 4/8 = 0.5 from 1. This means that we have obtained an improvement on the training and classification by a factor of 2.

As we know that the Diagnostic performance is usually evaluated in terms of Classification accuracy, Precision, Recall, Fallout and F – Measure, ROC, Confusion matrix. These terms are briefly explained below:

Classification Accuracy: Classification accuracy may be defined as the probability of it in correctly classifying records in the test datasets or Classification accuracy is the ratio of total number of correctly diagnosed cases to the total number of cases.

$$\text{Classification accuracy (\%)} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \qquad (7)$$

TP (True Positive): Sick people correctly detected as sick.

FP (False Positive): Healthy people incorrectly detected as diabetic people.

TN (True Negative): Healthy people correctly detected as healthy.
FN (False Negative): Sick people incorrectly detected as healthy.

Precision: Precision may define to measures the rate of correctly classified samples that are predicted as diabetic samples or precision is the ratio of number of correctly classified instances to the total number of instances fetched.

$$\text{Precision} = \frac{\text{No. of correctly classified instances}}{\text{Total No. of instances fetched}} \qquad (8)$$

$$\text{or Precision} = \text{TP}/\ \text{TP} + \text{FP} \qquad (9)$$

Recall: Recall may define to measures the rate of correctly classified samples that are actually diabetic samples or recall is the ratio of Number of Correctly Classified Instances to the Total Number of Instances in the Dataset.

$$\text{Recall} = \frac{\text{No. of correctly classified instances}}{\text{Total No. of instances in the Dataset}} \qquad (10)$$

$$\text{or Recall} = \text{TP}/\ \text{TP} + \text{FN} \qquad (11)$$

Fallout: The term fallout is used to check true negative of the dataset during classification.

F - Measure: The F – Measure computes some average of the information retrieval precision and recall metrics. The F – Measure (F – Score) is calculated based on the precision and recall.

The calculation is as follow:

$$\text{F} - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (12)$$

Area under curve (AUC): It is defined as the metric used to measure the performance of classifier with relevant acceptance. It is calculated from area under curve (ROC) on the basis of true positives and false positives.

$$\mathrm{AUC} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right) \quad (13)$$

ROC (Receiver operating curve graph) is an effective method of evaluating the performance of diagnostic tests.

Confusion matrix: A confusion matrix [12][2] contains information regarding actual and predicted classifications done by a classification system.

The following terms is briefly explained which is used in implementation result part.

Kappa Statistics: It is defined as performance to measure the true classification or accuracy of the algorithm.

$$\mathrm{K} = \frac{\mathrm{P0}-\mathrm{Pc}}{1-\mathrm{Pc}} \quad (14)$$

Where, P0 is the total agreement probability and Pc is the agreement probability due to change.

Root mean square error (RMSE): It is defined as the different between actual predicted value and the actual predicted value in the learning.

$$\mathrm{RMSE} = \sqrt[2]{\frac{1}{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}}(\mathrm{Ere}-\mathrm{Eacc})^2} \quad (15)$$

Where, Ere is the resultant error rate and Eacc is the actual error rate

Mean absolute error: It is defined as:

$$\mathrm{MAE} = \frac{|\mathrm{p}_1-\mathrm{a}_1|+\cdots+|\mathrm{p}_\mathrm{n}-\mathrm{a}_\mathrm{n}|}{\mathrm{n}} \quad (16)$$

Root mean-squared error: it is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{(\mathrm{p}_1-\mathrm{a}_1)^2+\cdots+(\mathrm{p}_\mathrm{n}-\mathrm{a}_\mathrm{n})^2}{\mathrm{n}}} \quad (17)$$

Relative squared error: It is defined as:

$$\mathrm{RSE} = \frac{(\mathrm{p}_1-\mathrm{a}_1)^2+\cdots+(\mathrm{p}_\mathrm{n}-\mathrm{a}_\mathrm{n})^2}{(\bar{\mathrm{a}}-\mathrm{a}_1)^2+\cdots+(\bar{\mathrm{a}}-\mathrm{a}_\mathrm{n})^2} \quad (18)$$

Relative absolute error: It is defined as:

$$\mathrm{RAE} = \frac{|\mathrm{p}_1-\mathrm{a}_1|+\cdots+|\mathrm{p}_\mathrm{n}-\mathrm{a}_\mathrm{n}|}{|\bar{\mathrm{a}}-\mathrm{a}_1|+\cdots+|\bar{\mathrm{a}}-\mathrm{a}_\mathrm{n}|} \quad (19)$$

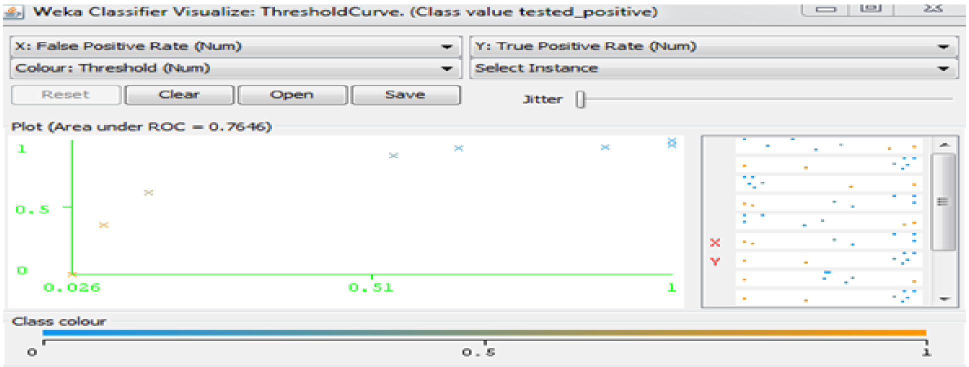
Where, 'a1,a2….an' are the actual target values and 'p1,p2….pn' are the predicted target values.

The Table 1 shows the analysis of evaluation of result on J48graft DT for PIDD.

Table 1. Evaluation of J48graft DT Performance for PIDD

Measure	Training set evaluation	Testing set evaluation
Precision	0.797	0.761
Recall	0.799	0.765
F - Measure	0.792	0.762
Accuracy	79.9257% (0.799257)	76.5217% (0.765217)
ROC	**0.859**	**0.765**

The figure shown below is the analysis of False positive rate Vs True positive rate. The PIDD classified using J48graft DT generates less error rate as shown in Figure 2.



Weka Classifier Visualize: ThresholdCurve. (Class value tested_positive)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.7646)

Figure 2. Analysis of positive rate for PIDD without GA

The Table 2 shows the result of training and testing accuracy, ROC by applying J48graft DT methodology.

Table 2. Classification with J48graft DT

Dataset	**Instances**	**Training data**	**Testing data**	**Attributes**	**Training accuracy**	**Testing accuracy**	**Training ROC**	**Testing ROC**
PIDD	768	538	230	8	79.9257 %	76.5217%	0.859	0.765

The Table 3 shows the feature reduction by using GA on PIDD which is just below.

Table 3. GA Feature Reduction

Data set	**Number of attributes**	**Feature set (Name of attributes)**	**No. of instances**	**No. of classes**
PIDD (Without GA)	8	1. Number of times pregnant 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test 3. Diastolic blood pressure 4. Triceps skin fold thickness 5. 2 – hour serum insulin 6. Body mass index 7. Diabetes pedigree function 8. Age (years)	768	2
PIDD (With GA)	4	2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test 5. 2 – hour serum insulin 6. Body mass index 8. Age (years)	768	2

The Table 4 shows the analysis of evaluation of result on J48graft DT for the selected features by using GA in the PIDD.

Table 4. Evaluation of GA_J48graft DT Performance for PIDD

Measure	**Training set evaluation**	**Testing set evaluation**
Precision	0.819	0.789
Recall	0.783	0.748
F - Measure	0.787	0.754
Accuracy	78.2528%	74.7826%
ROC	**0.862**	**0.786**

The Table 5 shows the result of training and testing accuracy, ROC by applying GA_J48graft DT methodology.

Table 5. Classification with GA_J48graft DT

Dataset	**Instances**	**Training data**	**Testing data**	**Attributes**	**Reduced attributes**	**Training accuracy**	**Testing accuracy**	**Train ROC**	**Test ROC**
PIDD	768	538	230	8	4	78.2528 %	74.7826 %	0.862	0.786

The Table 6 shows the analysis of comparison result with and without GA on J48graft DT for PIDD by several measure *i.e.*, noted in table.

Table 6. Evaluation of J48graft DT & GA _ J48graftDT Performance for PIDD

Measure	**J48graft DT**	**GA _ J48graft DT**
Precision	0.761	0.789
Recall	0.765	0.748
F – Measure	0.762	0.754
Accuracy	76.5217%	74.7826%
ROC	**0.765**	**0.786**

As we may see in Table 6, with GA the improvement has occurred in ROC, precision. We achieved slightly less accuracy, recall, and f-measure, may be by applying this approach only on this dataset but mostly in any cases by applying feature selection approach improvement occur in every parameter.

The figure shown below is the analysis of comparison result with and without GA on J48graft DT for PIDD.

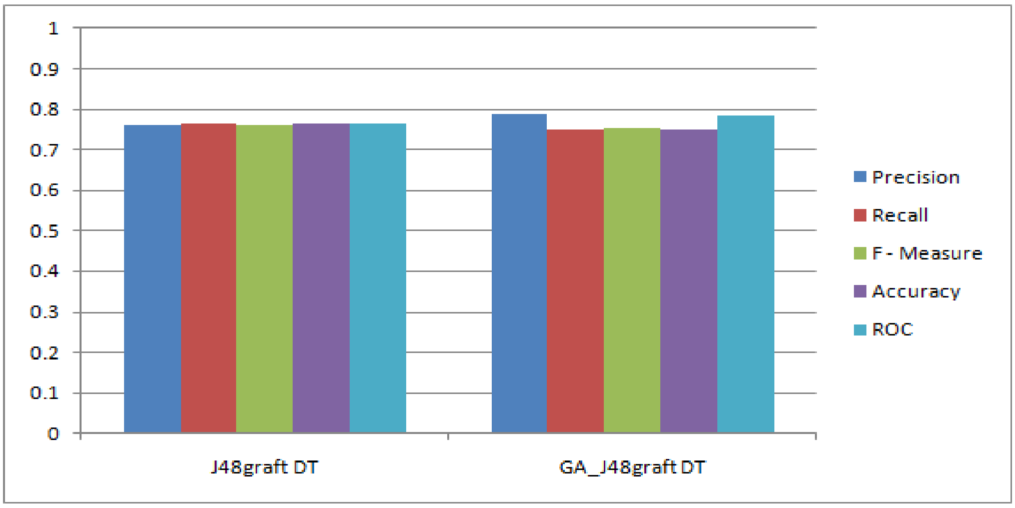
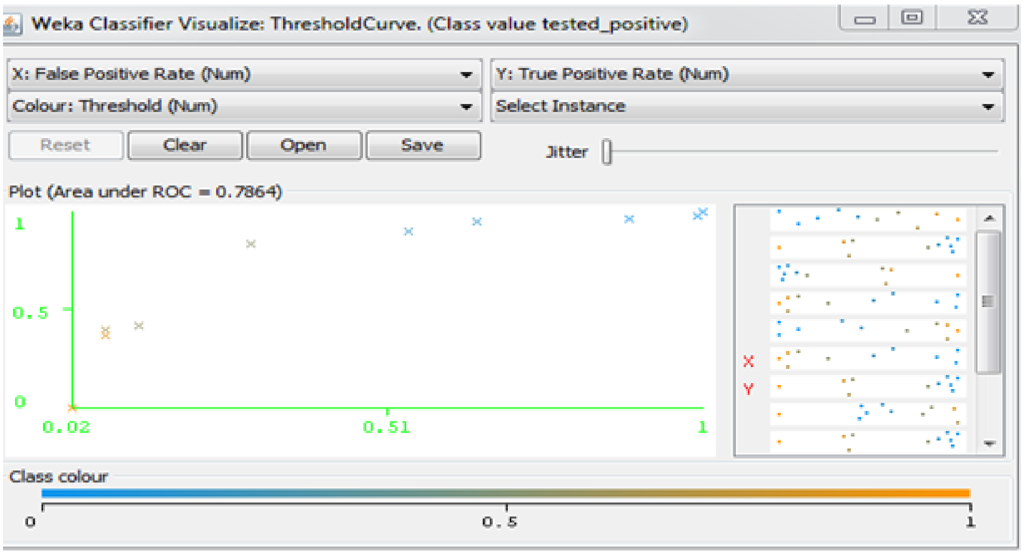


Figure 3. Evaluation of J48graft DT & GA _J48graft DT Performance for PIDD

The above Figure 3 is representing the above Table 6 parameters (measure) in chart graphical form and this is indicating the difference in more precise form between J48graft DT and GA _ J48graft DT.

The figure shown below is the analysis of False positive rate Vs True positive rate. The PIDD classified using J48graft Decision tree on selected feature by GA and generates less error rate as shown in Figure 4.



Weka Classifier Visualize: ThresholdCurve. (Class value tested_positive)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.7864)

Figure 4. Analysis of positive rate for PIDD with GA

The Figure 5 Shown below is the Decision tree generated when classification is done using J48graft DT on PIDD. The figure shows the dependent attributes of the dataset so that the classification is done easily and quickly.

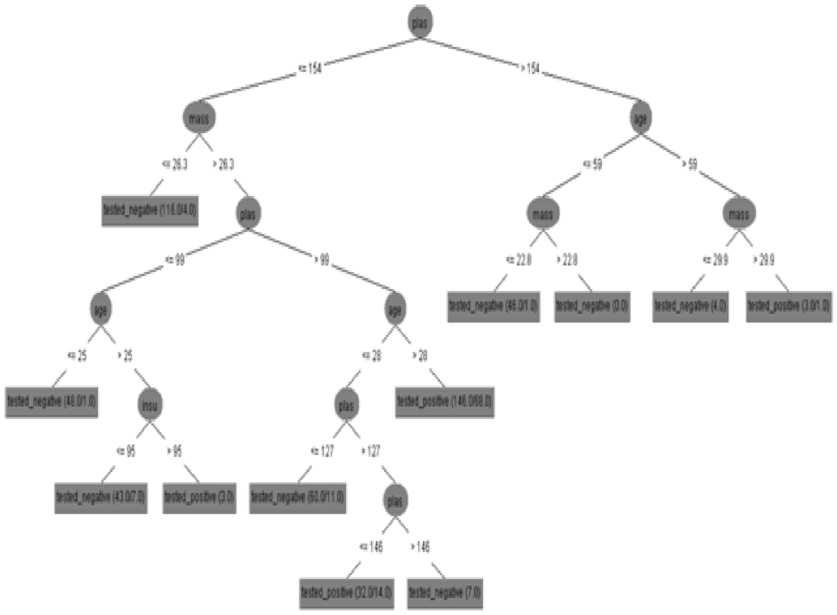


Figure 5. Generation of Decision Tree using J48graft

The rules generated using J48graft DT on selected features by using GA which is shown in Figure 5 for the classification of PIDD are as follows:

Rule1: If (plas >154) If (age >59) If (mass >29.9) “Tested Positive”	**Rule2:** If (plas >154) If (age >59) If (mass <=29.9) “Tested Negative”	**Rule3:** If (plas >154) If (age <=59) If (mass >22.8) “Tested Negative”	**Rule4:** If (plas >154) If (age <=59) If (mass <=22.8) “Tested Negative”
Rule5: If (plas <=154) If (mass >26.3) If (plas >99) If (age >28) “Tested Positive”	**Rule6:** If (plas <=154) If (mass >26.3) If (plas >99) If (age <=28) If (plas >127) If (plas >146) “Tested Negative”	**Rule7:** If (plas <=154) If (mass >26.3) If (plas >99) If (age <=28) If (plas >127) If (plas <=146) “Tested Positive”	**Rule8:** If (plas <=154) If (mass >26.3) If (plas >99) If (age <=28) If (plas <=127) “Tested Negative”
Rule9: If (plas <=154) If (mass >26.3) If (plas <=99) If (age >25) If (insu >95) “Tested Positive”	**Rule10:** If (plas <=154) If (mass >26.3) If (plas <=99) If (age >25) If (insu <=95) “Tested Negative”	**Rule11:** I f(plas <=154) If (mass >26.3) If (plas <=99) If (age <=25) “Tested Negative”	**Rule12:** If (plass <=154) If (mass <=26.3) “Tested Negative”

The Table 7 shows the result comparison in terms of accuracy on PIDD for the diagnosis of diabetes disease. As we may see in Table 7 that several methods have been employed.

Table 7. Results and Comparison with other Methods for the PIDD

Source	Method	Accuracy (%)
Pasi Luukka (2011)	Sim	75.29%
	Sim + F1	75.84%
	Sim + F2	75.97%
H. Hasan Orkcu et al. (2011)	Binary – coded GA	74.80%
	BP	73.80%
	Real – coded GA	77.60%
Manjeevan Seera et al. (2014)	FMM	69.28%
	FMM – CART	71.35%
	FMM-CART – RF	78.39%
Our Study	**GA _J48graft DT**	**74.7826%**

The Table 8 shows the result comparison in terms of ROC on PIDD for the diagnosis of diabetes disease. As we may see in Table 8 that the proposed method provide better ROC than other existing technique.

Table 8. Results and Comparison of ROC with other Methods for PIDD

Source	Method	ROC
Pasi Luukka (2011)	Sim	0.762
	Sim + F1	0.703
	Sim + F2	0.667
Manjeevan Seera et al. (2014)	FMM	0.661
	FMM - CART	0.683
	FMM-CART - RF	0.732
Our Study	**GA _J48graft DT**	**0.786**

5. Conclusion and Future Work

Diabetes is a chronic disease that occurs due to high blood glucose level in the body. Diabetes also contributed to so many diseases i.e. blindness, blood pressure, heart disease, kidney disease and nerve damage, etc. which is hazardous to health. The proposed approach implemented here for the feature selection, classification and prediction of Diabetes Patient's using GA_J48graft DT on PIDD. The proposed work minimizes the computation cost, computation time and maximizes the ROC and achieved slightly less accuracy compare than other existing methods. With features selection method (GA), we improve the ROC but achieved slightly less accuracy may by applying this method only on this dataset but mostly in any cases by applying feature selection method the accuracy also improved however the ROC has been improved.

For the future research work, we suggest to develop such a classification system of diabetes disease which provides good ROC, classification accuracy, precision, recall, f -measure which could significantly decrease healthcare costs via early prediction and diagnosis of diabetes disease, This approach of a classification system can also be used for other kinds of medical disease diagnosis.

References

[1] E. Dogantekin, A. Dogantekin, D. Avci and L. Avci, "An Intelligent Diagnosis System For Diabetes On Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA – ANFIS", Elsevier: Digital Signal Processing, vol. 20, **(2010)**, pp. 1248-1255.

[2] K. Polat and S. Gunes, "An Expert System Approach Based On Principal Component Analysis and Adaptive Neuro – Fuzzy Inference System to Diagnosis Of Diabetes Disease", Elsevier: Digital Signal Processing, vol. 17, **(2007)**, pp. 702-710.

[3] M. Seera and C. Peng Lim, “A Hybrid Intelligent System for Medical Data Classification”, Elsevier: Expert Systems with Applications, vol. 41, **(2014)**, pp. 2239-2249.

[4] H. Hasan Orkcu and H. Bal, “Comparing Performances of Backpropagation and Genetic Algorithms in the Data Classification”, Elsevier: Expert Systems with Applications, vol. 38, **(2011)**, pp. 3703-3709.

[5] P. Lukka, “Feature Selection using fuzzy entropy measures with similarity classifier”, Elsevier: Expert Systems with Applications, vol. 38, **(2011)**, pp. 4600-4607.

[6] H. Temurtas, N. Yumusak and F. Temurtas, “A Comparative Study On Diabetes Disease Diagnosis Using Neural Networks”, Elsevier: Expert Systems With Applications, vol. 36, **(2009)**, pp. 8610-8615.

[7] M. Waqar Aslam, Z. Zhu and A. Kumar Nandi, “Feature Generation Using Genetic Programming With Comparative Partner Selection For Diabetes Classification”, Elsevier: Expert Systems With Applications, vol. 40, **(2013)**, pp. 5402-5412.

[8] L. Rokach and O. Maimon, ‘Data mining and knowledge discovery handbook”, Chapter 9-decision trees.

[9] K. Selvakuberan, D. Kayathiri, B. Harini and Dr M. Indra Devi, “An efficient feature selection method for classification in Health care Systems using Machine Learning Techniques”, **(2011)**, IEEE.

[10] C. Darwin, “On the origins of species by means of natural selection”, **(1859)**, London : Murray.

[11] K. Kayaer and T. Yildirim, “Medical Diagnosis On Pima Indian Diabetes Using General Regression Neural Networks”, Yildiz Technical University, Department Of Electronics and Comm. Eng. Besiktas, Istanbul 34349 Turkey IEEE, **(2003)**.

[12] K. Polat, S. Guneh and A. Arslan, “A Cascade Learning System for Classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine”, Elsevier: Expert Systems with Applications, vol. 34, **(2008)**, pp. 482-487.

[13] D. E. Goldberg, “Genetic algorithms in search, optimization, and machine learning”, Addison - wesley, reading, ma, NN Schraudolph and J., vol. 3, no. 1, **(1989)**.

[14] UCI Repository of Bioinformatics Databases [online] Available: http://www.ics.uci.edu./~mlearn/ML Repository.html.

[15] M. Fathi Ganji and M. Saniee Abadeh, “Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease”, Proceedings of ICEE, IEEE, **(2010)**, May 11-13.

[16] J. H. Holland, “Adaptation in natural and artificial systems”, Ann Arbor MI: The University of Michigan Press, **(1975)**.

[17] E. Brissman and K. Eriksson, “Classification: Grafted decision trees”.

[18] Z. Michalewicz, “Genetic algorithms + data structures = evolution programs”, Springer, **(1996)**.

[19] H. Kahramanli and N. Allahverdi, “Design of a Hybrid System for the Diabetes and Heart Diseases”, Elsevier: Expert Systems with Applications, vol. 35, **(2008)**, pp. 82-89.

[20] V. K. Pachghare and P. Kulkarni, “Pattern Based Network Security using Decision Trees and Support Vector Machine”, **(2011)**, IEEE.

[21] K. Selvakuberan, D. Kayathiri, B. Harini and Dr M. Indra Devi, “An efficient feature selection method for classification in Health care Systems using Machine Learning Techniques”, **(2011)**, IEEE.

[22] C.-S. Lee, “A Fuzzy Expert System for Diabetes Decision Support Application”, IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, vol. 41, no. 1, **(2011)**.

[23] S. Karatsiolis and C. N. Schizas, “Region based Support vector machine algorithm for Medical Diagnosis on Pima Indian Diabetes Dataset”, Proceedings of the IEEE 12th International Conference on Bioinformatics& Bioengineering (BIBE), Larnaca, Cyprus, **(2012)** November 11-13.

[24] T. Jayalakshmi and Dr. A. Santhakumaran, “A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks”, International Conference on Data Storage and Data Engineering, IEEE, **(2010)**.

[25] C. Kalaiselvi and Dr. G. M. Nasira, Ph.D., “A New Approach for Diagnosis of Diabetes and Prediction of Cancer using ANFIS”, World Congress on Computing and Communication Technologies, IEEE, **(2014)**.

[26] S. Noman Qasem and S. Mariyam Shamsuddin, “Radial basis function network based on time variant multi-objective particle swarm optimization for medical diseases diagnosis”, Elsevier: Applied Soft Computing, vol. 11, **(2007)**, pp. 1427-1438.

[27] E. P. Ephzibah, “Cost Effective Approach on Feature Selection using Genetic Algorithms and Fuzzy Logic for Diabetes Diagnosis”, International Journal on Soft Computing (IJSC), vol. 2, no. 1, **(2011)**.

Authors

Dilip Kumar Choubey, received his M.Tech in Computer Science and Engineering from Oriental College of Technology (O.C.T), Bhopal, India and B.E. in Information Technology from Bansal Institute of Science and Technology (B.I.S.T), Bhopal India. Currently, He is Pursuing Ph.D from

Birla Institue of Technology (B.I.T), Mesra, Ranchi India. He worked as an Asst. Prof. in Lakshmi Narain College of Technology (L.N.C.T), Bhopal India and Oriental College of Technology (O.C.T), Bhopal India. He has 4 years of teaching and research experience. His research interests include soft computing, Bioinformatics, Data Mining and warehousing and Database Management System, etc. He has 5 International and 1 national publications. He is also the student member of Soft Computing Research Society. Ph. No. 7033789676, Email Id: dilipchoubey_1988@yahoo.in

Dr. Sanchita Paul, received her Ph.D degree and M.E. degree in Computer Science & Engineering from Birla Institute of Technology, Mesra, Ranchi, India and she has received B.E. degree in Computer Science & Engineering from Burdwan university, West Bengal, India. She has approximately 9 years of teaching and research experiences. She has 30 international publications. Her research areas include Machine learning, NLP, cloud computing, Bioinformatics, etc. Email: sanchita07@gmail.com