International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 8621-8633


# Gene Optimized Association Rule Generation Based Integral Derivative Gradient Boost Classification for Disease Diagnosis

**Sasirekha D[a*], Dr.Punitha A[b]**

[a]*Department of Computer Science, Bharathiyar University, Coimbatore - 641046, Tamil Nadu, India.*
[b]*Department of Computer Applications, Queen Mary's College, Chennai - 600004, Tamil Nadu, India.*

**Abstract**

Associative classification is a significant technique used for disease diagnosis. Few research works has been developed for associative classification to predict the disease patients. However, the performance of conventional associative classification technique was not efficient. In order solve this limitation, A Gene Optimized Association Rule Generation based Integral Derivative Gradient Boost Classification (GOARG-IDGBC) technique is proposed. The GOARG-IDGBC technique is designed for diagnosing the disease with minimal time consumption and higher classification accuracy. Initially GOARG-IDGBC technique used Optimized Genetic Algorithm (OGA) to generate the association rules from attributes in a medical dataset using support and confidence value. By using generated association rules, classification is then carried out using Integral Derivative Gradient Boost Classifier (IDGBC) in GOARG-IDGBC technique. Integral Derivative Gradient Boost Classifier classifies the patients in a medical dataset as normal or abnormal with higher classification accuracy through constructing strong classifier. Experimental evaluation of GOARG-IDGBC technique is carried out on factors such as classification accuracy, disease diagnosing time and false positive rate with respect to different number of patients. The experimental results show that the GOARG-IDGBC technique is able to improve the classification accuracy and also minimizes the time of disease diagnosing when compared to state-of-the-art works.

**Keywords:** Association rules, decision tree, Gradient Boost Classifier, Medical dataset, Optimized Genetic Algorithm, Strong classifier

## INTRODUCTION

Medical datasets contains huge amount of information regarding the patients, diseases and the physicians. Therefore, diseases diagnosis is required to predict the diseases. Disease prediction and decision making plays considerable role in medical diagnosis. Associative classification is one of the data mining techniques employed for disease diagnosis. Associative classification unites the concepts of association and classification for efficient diagnosis of disease. A lot of research works have been designed for associative classification. But, the performance of existing associative classification was not sufficient for achieving very higher classification accuracy for disease diagnosis.

A Neighborhood Rough Set based Classification (NRSC) algorithm was designed in [1] for disease prediction and to solve the medical diagnosis problems. However, classification performance was poor for disease prediction. A hybrid decision support system based on Rough Set Theory (RST) and Bat optimization Algorithm (BA) named RST-BatMiner [2] in order to improve the classification performance for detection of diabetes disease with higher classification accuracy. But, disease diagnosis time was more.

Multiple supports Classification Based on Associations (MSCBA) algorithm was designed in [3] for improving classification accuracy. However, time complexity of classification was more. A new algorithm was developed in [4] to generate hidden rules of gastric cancer data depends on ontology. But, disease prediction using generated association rules was remained unsolved.

An improved predictive association rule based classifier was intended in [5] for health care disease diagnosis and prognosis with higher classification accuracy. However, classification performance was not sufficient for accurate disease diagnosis. A combination of discretization and association rule-based classification was presented in [6] for Alzheimer's disease diagnosis which resulted in increased classification accuracy. But, sensitivity of disease prediction was poor.

An interval-valued fuzzy rule-based classification system was designed in [7] to correctly distinguish patients of the diverse risk categories. But, classification accuracy was poor. A hybrid Association Rule Mining and Artificial Neural Network was used in [8] in order to improve the performance of classification for Complex diseases prediction. However, time taken for diagnosing disease was very high.

A fuzzy classification system was presented in [9] for diagnosis of diabetes disease. But, false positive rate of classification was higher. An enhanced rule-based classification was developed in [10] to predict malaria disease with higher sensitivity and accuracy levels. However, diagnosis time was more.

In order to overcome the above mentioned existing issues, A Gene Optimized Association Rule Generation based Integral Derivative Gradient Boost Classification (GOARG-IDGBC) technique is designed. The major contribution of GOARG-IDGBC technique is formulated as,

- To improve the performance of disease diagnosis at an earlier stage with minimum time, GOARG-IDGBC technique is developed with application of Optimized Genetic Algorithm (OGA) and Integral Derivative Gradient Boost Classifier (IDGBC).
- To generate the optimal association rules for disease diagnosis, Optimized Genetic Algorithm (OGA) is used in GOARG-IDGBC technique. OGA selects the

attributes with minimum support and confidence value in a medical dataset for generating association rules. The generated association rules are employed to make the accurate classifier for diagnosing disease.

- To achieve very high classification accuracy for disease diagnosis, Integral Derivative Gradient Boost Classifier (IDGBC) is employed in GOARG-IDGBC technique. The GBC unites results of all base decision tree classifier into a strong classifier for accurately predicting the diseases patients with minimum time.

The rest of paper is arranged as follows. Section 2 explains a Gene Optimized Association Rule Generation based Integral Derivative Gradient Boost Classification (GOARG-IDGBC) technique with the aid of architecture diagram. Section 3 and Section 4 explains the experimental settings and details performance analysis with the aid of parameters. Section 5 reviews the different related works. Finally, a Section 6 concludes the paper.

## GENE OPTIMIZED ASSOCIATION RULE GENERATION BASED GRADIENT BOOST CLASSIFICATION TECHNIQUE

A Gene Optimized Association Rule Generation based Integral Derivative Gradient Boost Classification (GOARG-IDGBC) technique is designed with objective of improving the classification performance for disease diagnosis. The GOARG-IDGBC technique includes the two processes such as association rule generation and classification for efficiently diagnosing the disease of patients in a medical dataset with higher classification accuracy and minimum false positive rate. At first, the Optimized Genetic Algorithm (OGA) is applied in GOARG-IDGBC technique in order to generate the optimal number of association rules from the attributes in a medical data for efficient disease diagnosis. With the help of the generated association rules, then Integral Derivative Gradient Boost Classifier (IDGBC) classifies the patients in a medical dataset as normal or abnormal condition. The IDGBC is an ensemble of base classifier (i.e. decision tree).

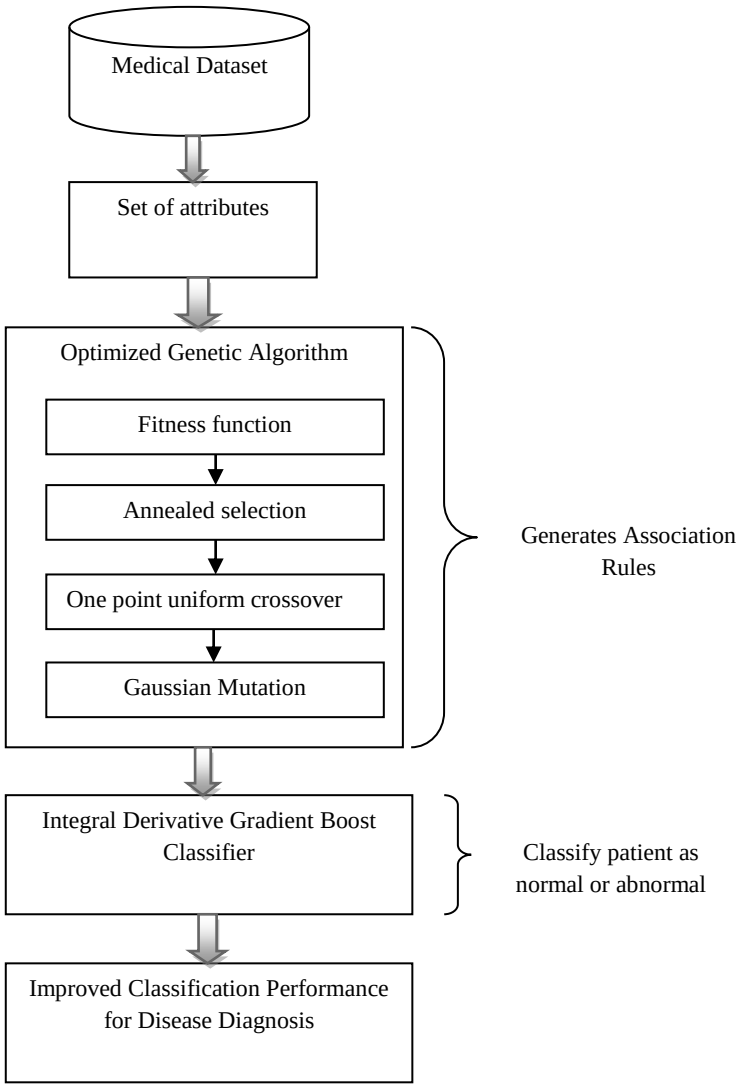


**Figure 1.** Flow Diagram of the Proposed Technique

The IDGBC combines results of all base classifier into a strong classifier for predicting occurrence of disease in a patient with minimum false positive rate. This helps for GOARG-IDGBC technique to predict the disease at an earlier stage with minimum time. The overall architecture diagram of GOARG-IDGBC technique is shown in Figure 1. As demonstrated in Figure 1, the GOARG-IDGBC technique initially takes medical dataset as input. The medical dataset comprises of numerous numbers of patients information. Then, GOARG-IDGBC technique used Optimized Genetic Algorithm (OGA) for generating the association rules by using attributes from a medical dataset. After that the gradient boosting classifier is used in GOARG-IDGBC technique to classify the patient as normal or abnormal using generated association rules. The detailed process of GOARG-IDGBC technique is shown in forthcoming sections.

**Optimized Genetic Algorithm**

Optimized Genetic Algorithm (OGA) is designed in GOARG-IDGBC technique in order to generate accurate and reliable association rules. The generated association rules are used to construct the accurate classifier for diagnosing disease. Association rule is employed for finding interesting relations between attributes in large medical dataset. Association rules explains the attribute conditions that occur frequently together in a given medical dataset for efficient diagnosing disease based on a threshold called support and confidence value. The support value identifies the frequent item (i.e. attribute) sets in medical dataset. The confidence is conditional probability for disease diagnosis. The OGA for generating association rules based on population size, crossover, mutation, and fitness function and selection operation. The process involved in Optimized Genetic Algorithm (OGA) to generate the association rule for diagnosing disease is shown in below Figure 2.

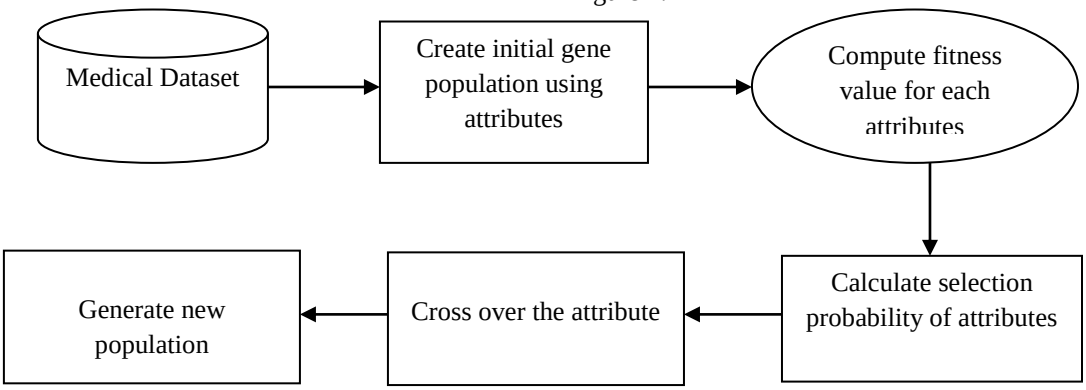


**Figure 2.** Process of Optimized Genetic Algorithm for Association Rule Generation

As demonstrated in Figure 2, process of OGA comprises of initialization, fitness evaluation, annealed selection, one point uniform crossover and Gaussian mutation process for generating optimal association rules. The input to OGA is a medical data set. At first OGA performs initialization process where attributes from the dataset gets initialized for association rule generation. Then, fitness value is determined for each attributes in a given medical data set with help of support and confidence value. After the fitness evaluation, annealed selection process is carried out to select the attribute with minimum support and confidence value. Then, one point uniform crossover process is carried out by combining low and high confidence attributes to form new attribute condition for association rule generation. Next, Gaussian Mutation randomly changes the bits of chromosome in order to generate optimal association rules.

***Initialization:***

An OGA at first randomly initializes the gene population with help of set of attributes in a given medical dataset.

***Annealed Selection Operation:***

In OGA, Selection process selects the gene population from initial population. During the selection process, attributes with higher fitness values is selected in order to generate association rule for disease diagnosis. Annealed Selection approach is employed in GOARG-IDGBC technique to select the individuals (i.e. attributes) from the population and creating the next generation. Selection probability of each individual is determined based on the fitness value. When the population generation changes, the fitness value and selection probability of each individual is also changed. Thus, the correlation between two individuals X and Y is evaluated as,

$$P(\mathrm{X.Y}) = \frac{\sum_{N=1}^{N} XY - \sum(X)(Y)}{\sqrt{\sum_{N=1}^{N} X^2 - \sum(X^2)[\sum_{N=1}^{N} Y^2 - \sum(Y^2)]}} \quad (1)$$

From equation (1), '$P$ (X.Y)' represents the Pearson correlation coefficient, N is the number of attributes in each data set. The value of the Pearson correlation coefficient varies between +1 and -1. If the value of the correlation coefficient is 1, then the two attributes has higher correlation whereas, the correlation coefficient '0'indicates lower correlation.

Besides, the value of higher correlation attribute is called as strong attribute which is selected for further processing whereas '0' indicates weak attribute. Annealed selection is gradually reduces the increasing number generation to remove

the weak individual (attribute). Besides, the population generation is sorted according to the selection probability where each individual (attribute) is allocated for selection probability. Minimum set of individuals (attributes), which is close enough to represent the original medical data is selected. The selected features form the smallest size of attribute selection to enable an efficient result.

***One Point Uniform Crossover :***

A crossover operation is performed in OGA to change chromosomes of attributes from one generation to the next generation to form association rule by combining low and high confidence attributes. The OGA designed in GOARG-IDGBC technique used One Point Uniform Crossover to get more than one parent chromosomes and generating child chromosomes from them. One point uniform crossover used OGA which presents the uniformity in combining the bits of both parents. One point uniform crossover is carried out through swapping bits in the parents by selecting a random real number $r$ (between 0 to 1). One point uniform crossover chooses the two parents and then constructs two offspring of $n$ genes uniformly. The random real number decides whether the first child choose the $i^{th}$ genes from first or second parent. The process of one point uniform crossover in OGA is shown in below.

As shown in Figure 3, one point uniform crossover is performed by changing chromosomes of two attributes to produce association rule for diagnosing disease.

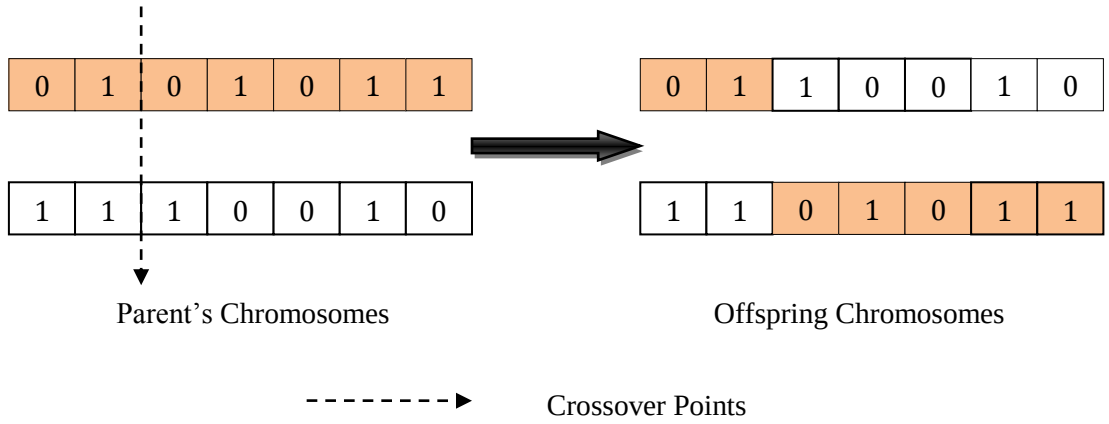


**Figure 3.** Process of One Point Uniform Crossover Operation

***Gaussian Mutation Operation:***

After the crossover operation, the mutation is carried out in OGA to randomly change the value of each bit of chromosome of attribute along with the probability. The OGA developed in GOARG-IDGBC technique employs Gaussian Mutation to randomly change the points of chromosome for generating association rules. Gaussian Mutation is carried out to preserve genetic diversity from one generation to the next generation. Gaussian mutation selects a random point from each population of an individual's vector to make a new offspring chromosome for association rule generation. Thus, Gaussian density function of population '$y$' is expressed as,

$$f_{Gaussian(0,\sigma^2)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} \quad (2)$$

From equation (2), $\sigma^2$ represents the variance whereas $y$ is a selected gene population. Let $y \in [a, b]$ is a real variable. The Gaussian mutation operator '$M_{Ga}$' change the population '$y$' into the next generation by using below mathematical representation,

$$M_{Ga}(y) := \min(\max(N(y, \sigma), a), b) \quad (3)$$

From equation (3), Gaussian Mutation is carried out for producing the new population for association rule generation. The Gaussian mutation process is shown in below.

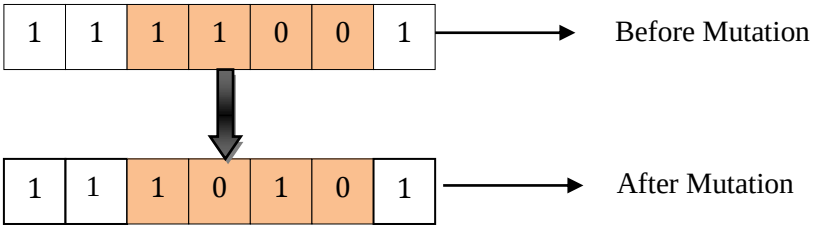


**Figure 4.** Process of Gaussian Mutation

As shown in Figure 4, Gaussian mutation is performed by interchanging the bit for formulating the new population in order to generate the association rule for constructing effective classifier to predict the diseases with higher classification accuracy.

***Fitness Function Calculation:***

In OGA, fitness value is measured for each attribute in a given medical data set in order to select the best attributes for generating association rules to efficient disease diagnosis. The fitness functions of an attribute based on the support and confidence values. Consider '$D$' is a medical data set that consists of numerous number of patients medical data $D = \{P_1, P_2, P_3, \dots P_N\}$ where each patient data contains a set of attributes $\{A_1, A_2, A_3, \dots A_M\}$. Generally, association rule is in the form of $X \rightarrow Y$ in which $X$ and $Y$ are disjoint attributes in medical dataset. The support measures probability of transaction containing both attributes value condition $X$ and $Y$. The confidence determines conditional probability that a transaction contains attributes value condition $Y$, given that it contains $X$. Thus, the support and confidence value of attribute is determined using below mathematical formula,

$$Sup(X \rightarrow Y) = \left(\frac{Support\ count\ of\ X \cup Y}{Total\ number\ of\ attributes\ in\ medical\ dataset}\right) \quad (4)$$

From equation (4), the support value of attributes is measured using the number of patient's records that contain $X \cup Y$ to the total number of patients records in a medical dataset. On the other hand, the confidence value is determined in such a way that the transactions which contain $X$ also contain $Y$ which is formulated as,

$$Conf(X \rightarrow Y) = \left(\frac{Support\ count\ of\ X \cup Y}{Support\ count\ of\ X}\right) \quad (5)$$

By using (4) and (5), support and confidence value of attributes in a medical dataset is calculated. Thus, the fitness value of attribute is measured as,

$$Fitness_A = sup + conf \quad (6)$$

From equation (6), fitness value of attributes in a medical dataset is determined with respect to support and confidence values. The attributes with minimum support and confidence values have a higher fitness value than other attributes in a medical dataset. Therefore, the attributes with higher fitness value are selected to generate the association rules for disease diagnosis.

The algorithmic process of optimized genetic algorithm for association rule generation is shown in below,

**// Optimized Genetic Algorithm based Association Rule Generation**

**Input:** Medical Data set (i.e. Collection of Attributes)

**Output:** Number of Association Rules

**Step 1: Begin**

**Step 2:** Initialize Gene Population from attributes in a medical dataset

**Step 3:** Calculate fitness value for each attribute using (6)

**Step 4:** **If** ($Fitness_A < 1.5\ Min_{th}$) **then**

**Step 5:** Select attribute to generate association rule

**Step 5:** **else**

**Step 5:** Perform annealed selection operation using (1)

**Step 6:** Perform one point uniform crossover

**Step 7:** Perform Gaussian Mutation using (2) and (3)

**Step 7:** Go to **Step 3**

**Step 8:** **End**

**End**

**Algorithm 1 Process of Optimized Genetic Algorithm**

Algorithm 1 shows the process of optimized Genetic Algorithm (OGA) to select the attributes in a medical dataset for generating optimal association rules for disease diagnosis. As shown in algorithm, initially OGA initialize gene population with help of attributes in a medical dataset. Then, fitness value is measured for each attribute in a medical dataset using support and confidence values. If fitness value of attribute is lesser than a minimum threshold $Min_{th}$ of 1.5, then attribute is selected to generate the association rule for diagnosing disease. Otherwise, annealed selection, one point uniform crossover, Gaussian mutation is carried out to form a new attribute value conditions for generating the association rules. This process is continual until an optimal solution is attained.

**Integral Derivative Gradient Boost Technique**

After generating the association rules, Integral Derivative Gradient Boost Classifier (GBC) is used in GOARG-IDGBC technique for diagnosing the disease with classification accuracy and time. IDGBC is a machine learning technique for classification problem. The IDGBC is in the form of ensemble of weak classifier (i.e. a decision tree). Besides, IDGBC is a machine learning technique used for building predictive tree-based models. Gradient boosting determines the residuals or errors of prior models and then combined together to predict the final results. Boosting is an ensemble technique in which new models are added to correct the errors made by base classifier model. In order to efficiently perform disease diagnosis, IDGBC algorithm employs the optimal rules generated from genetic approach. This in turn helps for

IDGBC algorithm to construct the accurate classifier for improving the performance of diagnosing the disease with minimum false positive rate.

The input of each training patient data in dataset is fixed where the base-learner model is just one Decision Tree (DT) and the loss function is the standard squared error. IDGBC algorithm also maximizes the correlation between generated rules in DT classifier to construct strong classifier. Predictions are performed by higher value of the weak learners' predictions, weighted by their individual accuracy. The main objective IDGBC is to reduce the loss of the model with the aid of adding weak learners using a gradient descent like procedure. Here, decision trees are considered as the weak learner in gradient boosting.

A gradient descent procedure is employed to reduce the loss (i.e. follow the gradient) when adding trees. Then the parameters of the tree are altered and move in the right way by (minimizing the residual error). Generally, gradient descent is applied to reduce a set of attributes in a network. After calculating error, the weights are updated to reduce that error for constructing strong classifier. Finally, the strong classifier effectively classifies training patient data as normal or abnormal classes. Therefore, IDGBC provides very higher classification accuracy with minimum time. The IDGBC designed in GOARG-IDGBC technique classify the patients as normal or abnormal based on attribute information of patients in a medical dataset by using generated association rules. The process of IDGBC algorithm for disease diagnosis is shown in below Figure 5.

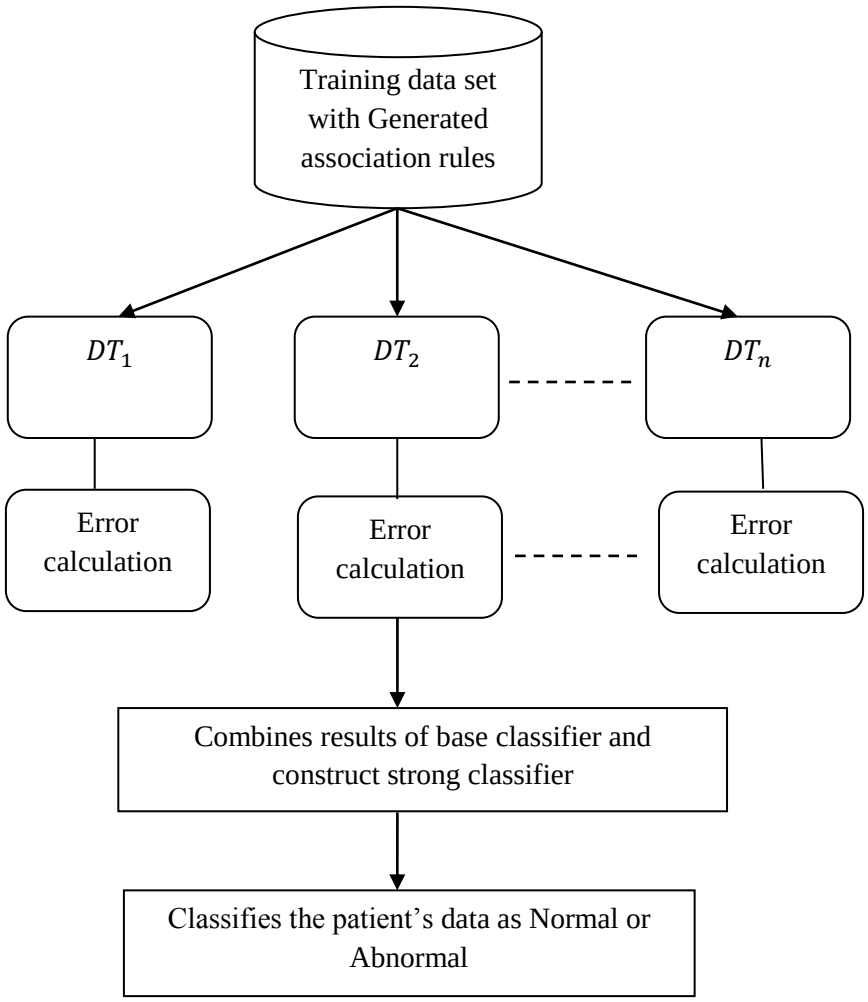


**Figure 5.** Process of Integral Derivative Gradient Boost Classifier for Disease Diagnosis

Figure 5 shows the process IDGBC algorithm for predicting the disease with higher classification accuracy. As demonstrated in figure, IDGBC algorithm takes the training data set and generated association rules as input. Initially, IDGBC algorithm performs decision tree classification for each training patient data with the help of generated association rules. After that, IDGBC algorithm determines the error rate of decision trees. Subsequently the weight value and vector parameter estimation is done to construct base decision tree classifier into a strong classifier in IDGBC. Finally, a strong classifier efficiently classifies the patients data as normal or abnormal which process resulting in improved classification accuracy with minimum time complexity.

In IDGBC algorithm, a base decision tree classification is carried out with help of generated association rules for achieving very high classification accuracy for disease diagnosis. IDGBC algorithm is the development of Gradient Boosting technique. The IDGBC uses the prediction models and loss function of decision tree classification for efficient disease diagnosis. Consider a set of N training dataset is in the form of$(q,r)_{i=1}^{N}$, which the task is to fit a model to loss function where $q$ tends to $q_N$ (i.e.,$q \rightarrow q_N$) denotes the patient information and $r$ tends to $r_N$ (i.e.,$r \rightarrow r_N$) denotes classification outputs of IDGBC algorithm The prediction model is used to reduce the new loss function of $\Psi(q,f(x))$ where $x = (x_1,..x_N)$. Therefore, the prediction output of the IDGBC is formulated as,

$$r_i = \Psi(q,f(x)) + h(q_i) \quad (7)$$

From equation (7), $\Psi(q,f(x))$ denotes predicted new loss function and $h(q_N)$ denotes the base DT classification result and $q_N$ denotes the number of patient information. In IDGBC, the new loss function (i.e. error of DT) is calculated as the difference between actual and predicted value. The loss function is,

$$\Psi(q,f(x)) = -\sum_{i=1}^{N} q_i log\ p_i(x) \quad (8)$$

From (8) $q$ represented as 1- of $N$ vector, where $N$ is the number of training data, $f(x)$ the overall boosting classifier function calculation, and

$$p_i(x) = \frac{\exp(f_i(x))}{\sum_{k=1}^{N}\exp(f_k(x))} \quad (9)$$

Then IDGBC generates N different gradient classifier trees $(h_1(q), h_1(q), \dots h_i(q))$ from training dataset. Therefore, the weight value is obtained as the integration of all decision tree classifier is computed as,

$$\int h(q_i) = \int_{i=1}^{N} h_i(q_i) \quad (10)$$

In order to minimize the new loss function $\Psi(q,f(x))$, the weighted sum of gradient classifier is require to be iterated from n = 1 to n = N. The negative gradient $g_t(x)$ is calculated as

$$g_t(x) = \left[\frac{\partial\Psi\,(q_n,f(x_n))}{\partial f(x_n)}\right]_{f(x)=f_{t-1}(x)}, n = 1,\dots,N \quad (11)$$

Substituting (9) into (8) and taking the first derivatives,

$$-g_t(x) = -\left[\frac{\partial\Psi\,(q_n,f(x_n))}{\partial f(x_n)}\right]_{f(x)=f_{t-1}(x)} = y - p(x) \quad (12)$$

By using (12), the residual errors are computed. Subsequently, the base decision tree classifier trains on the remaining errors. The error is calculated in each iteration. Then the weight of $h(q_i)$ and the vector parameter of $v_m$ for the m[th] iteration is computed as follows,

$$h((q_i), v_m) = arg\ min_{q,m}\ \sum_{i=1}^{N}\Psi\,(q_n, f_{m-1}(x_n) + h(x_n; v_m)) \quad (13)$$

Form (13), the input of $q_i$ is modified to the classification model $g_t(x)$ and obtains the calculation of $v_m$ of $h(q_i, v)$. Therefore parameter of $v_m$ is obtained as,

$$v_m = arg\ min_v \sum_{i=1}^{N}\frac{1}{1+e^{g_t(x)h(x_n;v_m)}} \quad (14)$$

Then compute parameter of $h(q_i)$ is obtained by reducing the new loss function $\Psi(q,f(x))$ as follows,

$$\Psi(q,f(x)) = \min\ \ \Psi\,(q_n, f_{m-1}(x_n) + h(x_n; v_m)) \quad (15)$$

At last, IDGBC updates the model to classify the patient information as normal or abnormal using the below mathematical representation, given below

$$\Psi(q,f(x)) = \Psi\,(q_n, f_{m-1}(x_n) + h(x_n; v_m)) \quad (16)$$

From equation (16), $\Psi(q,f(x))$ provides the strong classifier output. The strong classifier output '1' indicates the patient is in normal condition whereas '$-1$' denotes the patient is suffering from disease. The algorithmic process of IDGBC for disease diagnosis is shown in below,

**// Gradient Boost Classifier Algorithm**

**Input** : patient data$(q,r)_{i=1}^{N} = (q_1,r_1),(q_2,r_2)\ldots.(q_i,r_i)$,

**Output :** Improved classification accuracy with minimum time

**Step 1**: **Begin**

**Step 2**: **For** n=1 to N do

**Step 3**: Construct decision trees

**Step 4**: Measure the loss function using (8)

**Step 5**: Compute negative gradient value using (11)

**Step 6**: Determine weight and vector parameter using (13)

**Step 7:** Obtain the estimation of $h(q_i)$by minimizing new loss function using (15)

**Step 8**: Update the model as strong classifier using (16)

**Step 9:** **If** ($\Psi(q,f(x))$ results = '1' ) **then**

**Step 10:** The patient is classified as normal

**Step 11**: **else**

**Step 12:** The patient is classified as abnormal

**Step 13**: **End if**

**Step 14:** **End for**

**Step 15**: **End**

**Algorithm 2.** Integral Derivative Gradient Boost Classifier for Disease Diagnosis

Algorithm 2 shows the algorithmic process IDGBC for efficient disease diagnosis. For each training patient data in dataset, initially IDGBC algorithm performs decision tree classification with help of generated association rules from OGA. Subsequently, the IDGBC algorithm measures error

rate of base decision tree classifier in order to increase classification performance for disease diagnosis. Next, the negative gradient value is calculated to reduce the overall error or loss base decision tree classifier. After that, input is adapted to the gradient value and obtains weight and vector parameter estimation with minimum loss. At last strong classifier significantly classifies the patient as normal or abnormal. This helps for GOARG-IDGBC technique to enhance the classification performance with minimum false positive rate. Therefore, a GOARG-IDGBC technique attains higher classification accuracy with minimum time.

## EXPERIMENTAL SETTINGS

In order to analyze the performance, Gene Optimized Association Rule Generation based Gradient Boost Classification (GOARG-IDGBC) technique is implemented in MATLAB using Diabetes 130-US hospitals for years 1999-2008 Data Set. This dataset contains 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Diabetes 130-US hospitals for years 1999-2008 Data Set contains 50 attributes representing patient and hospital outcomes for diabetic disease diagnosis.

The Diabetes 130-US hospitals for years 1999-2008 Data Set includes following attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc. From that, GOARG-IDGBC technique selects attributes with minimum support and confidence value in order to generate the association rules by using optimized genetic algorithm.

By using the generated association rules, then GOARG-IDGBC technique constructs the gradient boost classifier in order to improve the classification performance of diabetic disease diagnosis. The GOARG-IDGBC technique considers 100 patients data from Diabetes 130-US hospitals for years 1999-2008 Data Set for conducting the experimental work. The experimental is conducted for many instances with respect to diverse number of patient data and averagely ten results are shown in below tables and graphs for analyzing the proposed performance. The effective of GOARG-IDGBC technique is measured in terms of classification accuracy, disease diagnosis time and false positive rate. The performance of GOARG-IDGBC technique is compared against with Neighborhood Rough Set based Classification (NRSC) algorithm [1] and hybrid decision support system based on Rough Set Theory (RST) and Bat optimization Algorithm (BA) called RST-Bat Miner [2] respectively.

## RESULT AND DISCUSSIONS

In this section, the result of GOARG-IDGBC technique is evaluated. The performance of GOARG-IDGBC technique is compared against with existing Neighborhood Rough Set based Classification (NRSC) algorithm [1] and hybrid decision support system based on Rough Set Theory (RST) and Bat optimization Algorithm (BA) called RST-Bat Miner [2]. The efficacy of GOARG-IDGBC technique is evaluated along with the following metrics with the assist of tables and graphs.

### Measure of Classification Accuracy

In GOARG-IDGBC technique, Classification Accuracy ($CA$) is determined as the ratio of number of patients that are correctly classified to the total number of patients. The classification accuracy is measured in terms of Percentages (%) and formulated as,

$$CA = \left(\frac{Number\ of\ patients\ correctly\ classified}{Total\ number\ of\ patients}\right) * 100 \qquad (17)$$

From equation (17), classification accuracy is evaluated with respect to different number of patient's data. While classification accuracy is higher, the method is said be more effectual.

**Table 1.** Tabulation for Classification Accuracy

| Number of Patients | Classification Accuracy (%) | | |
|---|---|---|---|
| | NRSC | RST-BatMiner | GOARG-IDGBC technique |
| 10 | 59.30 | 68.11 | 95.15 |
| 20 | 60.15 | 71.55 | 95.70 |
| 30 | 63.80 | 72.80 | 96.05 |
| 40 | 65.42 | 74.25 | 96.50 |
| 50 | 66.85 | 75.50 | 96.95 |
| 60 | 69.71 | 76.75 | 97.30 |
| 70 | 71.69 | 79.10 | 97.75 |
| 80 | 74.35 | 81.53 | 97.98 |
| 90 | 75.25 | 82.95 | 98.35 |
| 100 | 77.75 | 85.35 | 98.85 |

Table 1 portrays the performance of classification accuracy for predicting disease with respect to diverse number of patients using three methods. GOARG-IDGBC technique considers framework with different number of patients in the range of 10-100 for performing the simulation process using MATLAB. While considering the 70 number of patient data for diagnosing diabetic disease, proposed GOARG-IDGBC technique achieves 97.75 % classification accuracy whereas existing NRSC algorithm [1] and RST-BatMiner [2] achieves 71.69 % and 79.10 % respectively. Thus, classification accuracy for diabetic diagnosing disease using proposed GOARG-IDGBC technique is higher when compared to other existing [1], [2].

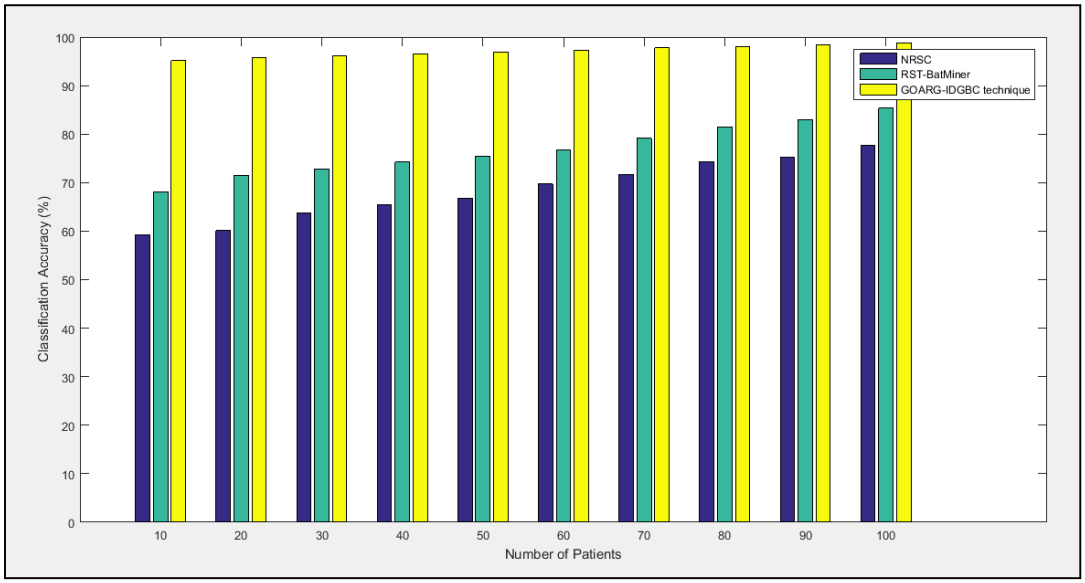


**Figure 6.** Measurement of Classification Accuracy versus Different Number of Patients

Figure 6 depicts the impact of classification accuracy for diagnosing diabetic disease using three methods based on dissimilar number of patients in the range of 10-100. As demonstrated in figure, proposed GOARG-IDGBC technique provides higher classification accuracy for predicting the diabetic disease as compared to existing NRSC algorithm [1] and RST-BatMiner [2]. In addition, while increasing the number of patients data for experimental work, the classification accuracy is also gets increased using all three methods. But comparatively, classification accuracy using GOARG-IDGBC technique is higher. This is because of application of OGA and IDGBC in GOARG-IDGBC technique. The OGA used in GOARG-IDGBC technique generates the optimal number of association rules for diabetic disease diagnosis. Then, IDGBC employed in GOARG-IDGBC technique classifies the patients as normal or abnormal with aid of generated association rules. In IDGBC, the base decision tree classification is performed based on generated association rules. Besides, IDGBC combines results of all base decision tree classifier into a strong classifier in order to improve the classification performance of diabetic disease diagnosis. This in turn helps for enhancing the classification accuracy for predicting diabetic diseases in an effective manner. As a result, proposed GOARG-IDGBC technique increases the classification accuracy by 43 % and 27 % when compared to existing NRSC algorithm [1] and RST-Bat Miner [2] respectively.

**Measure of Disease Diagnosing Time**

In GOARG-IDGBC technique, Disease Diagnosing Time ($DDT$) measures the amount of time taken for classifying the patients as normal or abnormal. The disease diagnosing time is measured in terms of milliseconds (ms) and mathematically represented as,

$$DDT = N * Time\ (classify\ the\ patients) \quad (18)$$

From equation (18), the disease diagnosing time is determined with respect to diverse number of patients (N). While disease diagnosing time is lower, the method is said be more effective.

**Table 2.** Tabulation for Disease Diagnosing Time

| Number of Patients | Disease Diagnosing Time (ms) | | |
|---|---|---|---|
| | NRSC | RST-BatMiner | GOARG-IDGBC technique |
| 10 | 24 | 21 | 13 |
| 20 | 31 | 29 | 18 |
| 30 | 37 | 34 | 25 |
| 40 | 42 | 38 | 29 |
| 50 | 51 | 44 | 35 |
| 60 | 59 | 50 | 42 |
| 70 | 63 | 57 | 50 |
| 80 | 68 | 62 | 56 |
| 90 | 73 | 70 | 61 |
| 100 | 81 | 77 | 68 |

Table 2 depicts the comparative result analysis of time taken for diagnosing diabetic disease based on various numbers of patients in the range of 10-100 using three methods. While considering the 80 number of patient data for conducting experimental work, proposed GOARG-IDGBC technique takes 56 ms disease diagnosing time whereas existing NRSC algorithm [1] and RST-BatMiner [2] achieves 68 ms and 62 ms respectively. Therefore, disease diagnosing time using proposed GOARG-IDGBC technique is lower when compared to other existing [1], [2].

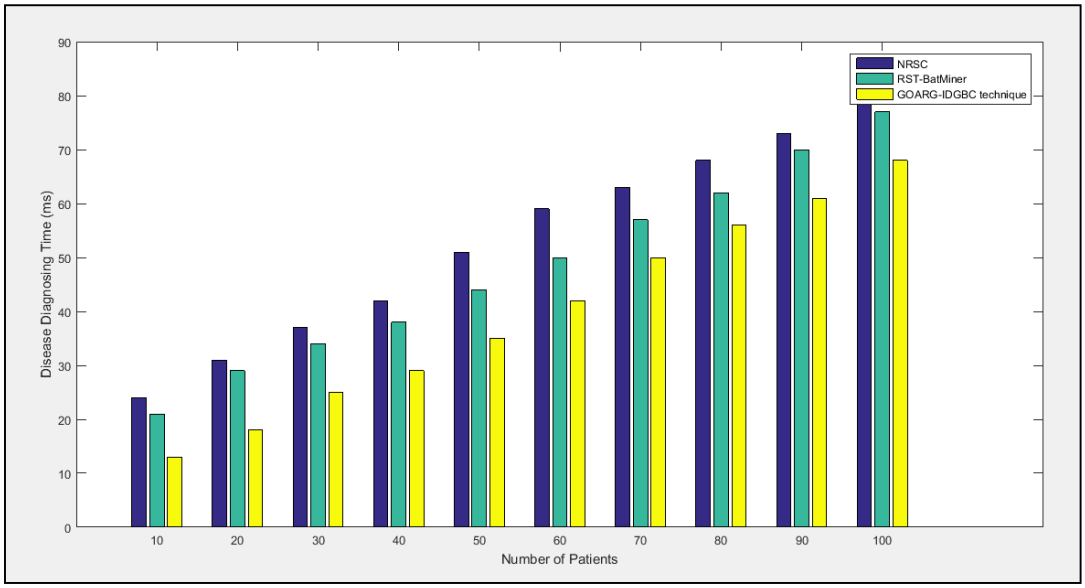


**Figure 7.** Measurement of Disease Diagnosing Time versus Different Number of Patients

Figure 7 explains the impact of disease diagnosing time for predicting diabetic disease using three methods with respect to different number of patients in the range of 10-100. As shown in figure, proposed GOARG-IDGBC technique provides minimum disease diagnosing time for predicting the diabetic disease as compared to existing NRSC algorithm [1] and RST-BatMiner [2]. As well, while increasing the number of patients data for conducting experimental process, the disease diagnosing time is also gets increased using all three methods. But comparatively, disease diagnosing time using GOARG-IDGBC technique is lower. This is owing to application of IDGBC in GOARG-IDGBC technique. Then, IDGBC developed in GOARG-IDGBC technique efficiently constructs the strong classifier by combing the results of all base decision tree classifier with aid of generated association rules. This supports for GOARG-IDGBC technique precisely categorizes the patients as normal or abnormal with minimum time. This in turn assists for reducing the amount of time taken for predicting diabetic diseases in a significant manner. Thus, proposed GOARG-IDGBC technique minimizes the disease diagnosing time by 28 % and 21 % when compared to existing NRSC algorithm [1] and RST-BatMiner [2] respectively.

**Measurement of False Positive Rate**

In GOARG-IDGBC technique, False Positive Rate ($FPR$) is defined as the ratio of number of incorrectly classified patients as a sick to the total number of patients. The false positive rate is measured in terms of percentage (%) and represented as below,

$$FPR = \frac{Number\ of\ incorrectly\ classified\ patients}{total\ number\ of\ patients} * 100 \quad (19)$$

From equation (19), false positive rate of classification for diagnosing disease is evaluated with respect to various numbers of patients (N). While false positive rate of classification is lower, the method is said be more efficient.

**Table 3.** Tabulation for False Positive Rate

| Number of Patients | False Positive Rate (%) | | |
|---|---|---|---|
| | NRSC | RST-BatMiner | GOARG-IDGBC technique |
| 10 | 43 | 35 | 21 |
| 20 | 45 | 37 | 24 |
| 30 | 49 | 40 | 26 |
| 40 | 50 | 41 | 27 |
| 50 | 52 | 43 | 29 |
| 60 | 53 | 44 | 30 |
| 70 | 55 | 47 | 31 |
| 80 | 58 | 48 | 33 |
| 90 | 59 | 50 | 34 |
| 100 | 61 | 52 | 37 |

Table 3 demonstrates the result analysis of false positive rate of classification for diabetic diagnosing disease with respect to dissimilar numbers of patients in the range of 10-100 using three methods. While considering the 90 number of patient data for performing experimental process, proposed GOARG-IDGBC technique obtains 34 % false positive rate whereas existing NRSC algorithm [1] and RST-BatMiner gets 59 % and 50 % respectively. As a result, false positive rate of classification for diabetic diagnosing disease using proposed GOARG-IDGBC technique is lower when compared to other existing [1], [2].

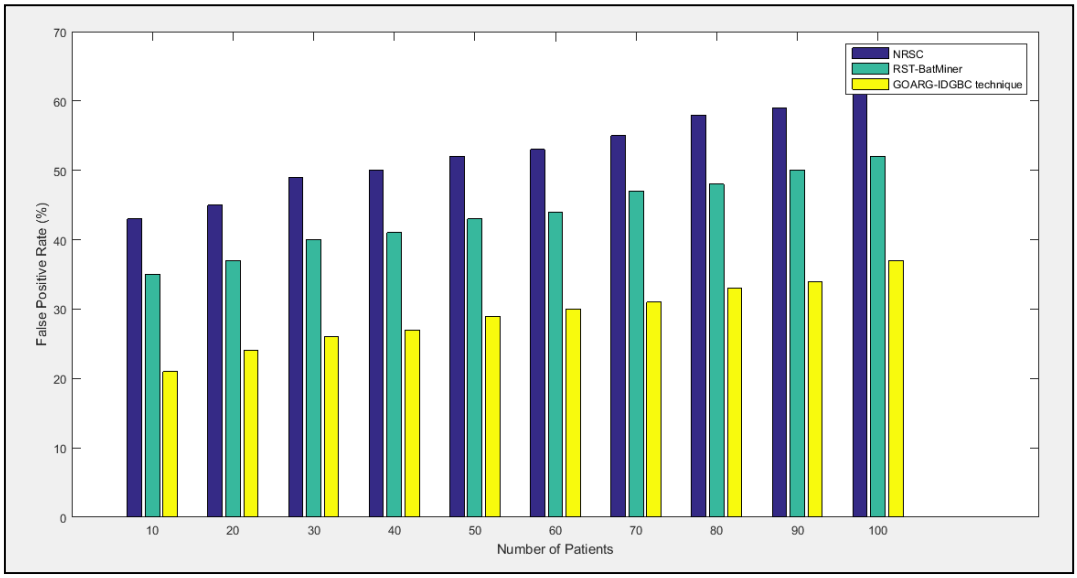


**Figure 8.** Measurement of False Positive Rate versus Different Number of Patients

Figure 8 describes the impact of false positive rate of classification for predicting diabetic disease using three methods based on various numbers of patients in the range of 10-100. As exposed in figure, proposed GOARG-IDGBC technique provides minimum false positive rate for predicting the diabetic disease as compared to existing NRSC algorithm [1] and RST-BatMiner [2]. Besides, while increasing the number of patients data for diagnosing disease, the false positive rate of classification is also gets increased using all three methods. But comparatively, false positive rate of classification using GOARG-IDGBC technique is lower. This is due to application of OGA and IDGBC in GOARG-IDGBC technique. The OGA generates the association rules using attributes in a medical dataset for constructing efficient classifier for diabetic disease diagnosis. With assists of generated association rules, then IDGBC creates strong classifier to significantly classify the patients as normal or abnormal. This helps for GOARG-IDGBC technique to increase the classification performance with minimum misclassification error. This process resulting in reduced false positive rate of classification for predicting diabetic disease in an effective manner. Hence, proposed GOARG-IDGBC technique minimizes the false positive rate by 45 % and 34 % when compared to existing NRSC algorithm [1] and RST-Bat Miner [2] respectively.

## RELATED WORKS

An efficient mining of association rules was intended in [11] for the early diagnosis of Alzheimer's disease through improving the classification accuracy. But, number of association rules was large. Hybrid disease diagnosis was presented in [12] with application of multi objective optimization and evolutionary parameter optimization to improve prediction accuracy of diseases. However, performance of disease prediction was not effectual.

Classification Rule Mining was intended in [13] for diagnosis of risk factors of Liver Disorders. But, classification rule mining performance was not efficient for early diagnosis. A hybrid classification system was developed in [14] with aid of Relief and Rough Set (RFRS) method to diagnosis of heart disease. However, classification using association rule was remained unsolved.

A rough-fuzzy classifier was used in [15] by combining with rough set theory and fuzzy set for heart disease diagnosing. But, disease diagnosing performance was not effective. A support vector machine (SVM) classifier was presented in [16] with application of logistic regression for achieving the high accuracy for diagnosis of Parkinson's disease (PD). However, the amount time taken for diagnosis was higher.

A Nearest neighbor (KNN) classifier was employed in [17] with objective of classifying heart disease. But, the prediction accuracy of disease was not at required. A hybrid intelligent system was presented in [18] with aiming at diagnosing breast cancer by using rough set theory and k-nearest neighbor algorithm. However, other medical disease diagnosis was remained unsolved.

A hybrid Meta heuristic technique was designed in [19] using ant colony optimization (ACO) phase and a genetic algorithm (GA) phase for classifying the medical data. But, false positive rate for diagnosing the disease was higher. In [20], a hierarchical learning algorithm was used in order to classifying patient records. However, hierarchical learning algorithm does not employ association rule for constructing classifier which resulting poor disease diagnosis performance.

## CONCLUSION

An effective Optimized Association Rule Generation based Gradient Boost Classification (GOARG-IDGBC) technique is developed to improve the performance of diagnosing the disease with higher classification accuracy. The key objective of the proposed GOARG-IDGBC technique is to improve the

classification performance of disease diagnosis with minimum false positive rates and time by using Optimized Genetic Algorithm (OGA) and Integral Derivative Gradient Boost Classifier (IDGBC). At first GOARG-IDGBC technique generate the association rules from the attributes in a medical dataset with application of OGA. Then, GOARG-IDGBC technique classifies the patients as normal or abnormal with help of IDGBC. This in turn helps for GOARG-IDGBC technique to attain higher classification accuracy for disease diagnosis. The effectiveness of GOARG-IDGBC technique is measured in terms of classification accuracy, disease diagnosis time, and false positive rate in comparison with state of the art works. The experimental result shows that GOARG-IDGBC technique provides better performance with an improvement of classification accuracy and reduction of disease diagnosis time as compared to state-of-the-art works.

## REFERENCES

[1] S. Udhaya kumar, H. Hannah Inbarani, "A Novel Neighborhood Rough set Based Classification Approach for Medical Diagnosis", Procedia Computer Science, Elsevier, Volume 47, Pages 351 – 359, 2015

[2] Ramalingaswamy Cheruku, Damodar Reddy Edla, Venkatanareshbabu Kuppili, Ramesh Dharavath, "RST-BatMiner: A Fuzzy Rule Miner Integrating Rough Set Feature Selection and Bat Optimization for Detection of Diabetes Disease", Applied Soft Computing, Elsevier, Pages 1-50, 2017

[3] Li-Yu Hu, Ya-Han Hu, Chih-Fong Tsai, Jian-Shian Wang, and Min-Wei Huang, "Building an associative classifier with multiple minimum supports", Springer plus, Volume 5, Pages 1-19, 2016

[4] Seyed Abbas Mahmoodi, Kamal Mirzaie and Seyed Mostafa Mahmoudi, "A new algorithm to extract hidden rules of gastric cancer data based on ontology", Springer Plus, Volume 5, Pages 1-21, 2016

[5] M Nandhini and S N Sivanandam, "An improved predictive association rule based classifier using gain ratio and T-test for health care data diagnosis", Sadhana, Springer, Volume 40, Issue 6, Pages 1683–1699, September 2015

[6] R.Chaves, J.Ramírez, J.M.Górriz, "Integrating discretization and association rule-based classification for Alzheimer's disease diagnosis", Expert Systems with Applications, Elsevier, Volume 40, Pages 1571–1578, 2013

[7] Jose Antonio Sanz, Mikel Galar, Aranzazu Jurio, Antonio Brugos, Miguel Pagola, Humberto Bustincea, "Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system", Applied Soft Computing, Elsevier, Volume 20, Pages 103-111, July 2014

[8] AichaBoutorh, AhmedGuessoum, "Complex diseases SNP selection and classification by hybrid Association Rule Mining and Artificial Neural Network—based Evolutionary Algorithms", Engineering Applications of Artificial Intelligence, Elsevier, Volume 51, Pages 58-70, May 2016

[9] Mostafa Fathi Ganji, Mohammad Saniee Abadeh, "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis", Expert Systems with Applications, Volume 38, Issue 12, Pages 14650-14659, November–December 2011

[10] Francis Bbosa, Ronald Wesonga, Peter Jehopio, "Clinical malaria diagnosis: rule-based classification statistical prototype", SpringerPlus, Volume 5, Pages 1-14, 2016

[11] R Chaves, J M Górriz, J Ramírez, I A Illán, D Salas-Gonzalez and M Gómez-Río, "Efficient mining of association rules for the early diagnosis of Alzheimer's disease", Physics in Medicine & Biology, Volume 56, Issue 18, Pages 6047–6063, 2011

[12] Madhu Sudana Rao Nalluri, Kannan K, Manisha M, and Diptendu Sinha Roy, "Hybrid Disease Diagnosis Using Multiobjective Optimization with Evolutionary Parameter Optimization", Journal of Healthcare Engineering, Hindawi, Volume 2017, Article ID 5907264, Pages 1-27, 2017

[13] Sushruta Mishra, Hrudaya Kumar Tripathy, Brojo Mishra and Soumya Sahoo, "Implementation of Classification Rule Mining to minimize Liver Disorder risks", International Journal of Control Theory and Applications, Volume 10, Issue 18, Pages 17-124, 2017

[14] Xiao Liu, Xiaoli Wang, Qiang Su, Mo Zhang, Yanhong Zhu, Qiugen Wang, and Qian Wang, "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method", Hindawi Publishing Corporation, Computational and Mathematical Methods in Medicine, Volume 2017, Article ID 8272091, Pages 1-11, 2017

[15] K. Srinivas, G. Raghavendra Rao, A. Govardhan, "Rough-Fuzzy Classifier: A System to Predict the Heart Disease by Blending Two Different Set Theories", Arabian Journal for Science and Engineering, Springer, Volume 39, Issue 4, Pages 2857–2868, April 2014

[16] R. Prashanth, Sumantra Dutta Roy, Pravat K. Mandal, Shantanu Ghosh," Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging", Expert Systems with Applications, Elsevier, Volume 41, Issue 7, Pages 3333-3342, 2014

[17] Jabbar, M.A,Deekshatulu, B.Land Chandra,Priti, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, "Procedia Technology, Elsevier, Volume 10, Pages 85-94, 2013

[18] A. H. El-Baz, "Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis", Neural Computing and Applications, Springer, Volume 26, Issue 2, Pages 437–446, 2015

[19] Sarab AlMuhaideb and Mohamed El Bachir Menai, "A new hybrid metaheuristic for medical data classification", International Journal of Metaheuristics, Volume 3, Issue 1, Pages 59-80, 2014

[20] Kuizhi Mei, Jinye Peng, Ling Gao, Naiquan (Nigel) Zheng, Jianping Fan "Hierarchical Classification of Large-Scale Patient Records for Automatic Treatment Stratification", IEEE Journal of Biomedical and Health Informatics, Volume 19, Issue 4, Pages 1234 – 1245, July 2015

[21] Diabetes 130-US hospitals for years 1999-2008 Data Set: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008