ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2018.0226

INTEGRATION OF ROUGH SET THEORY AND GENETIC ALGORITHM FOR OPTIMAL FEATURE SUBSET SELECTION ON DIABETIC DIAGNOSIS

K. Thangadurai[1] and N. Nandhini[2]
[1]*Department of Computer Science, Karur Arts and Science College, India*
[2]*Department of Computer Science, Periyar University, India*

Abstract

Diabetic diagnosis is an important research in health care domain to analyze relevant microorganisms at an earlier stage. Due to large growth in world's population, feature subset selection model receives a great deal in any domain of research and also a reliable tool for diabetic diagnosis. Several data mining techniques have been developed to evaluate the significant causes of diabetes with least sets of risk factors. These minimum set is selected without considering the potential significance of the risk factors and optimal feature subset selection, hence it failed to diagnosis the pattern of diabetes accurately. In order to improve the feature subset selection, an Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is introduced. The main objective of the IFRST-OGA is to find optimal risk factors for efficient pattern recognition on diabetes healthcare data. Initially, feature selection is performed using Fuzzy Rough Set Theory (FRST) for diagnosing the diabetes. After that, the Optimized Genetic Algorithm (OGA) is applied which mainly searches for an optimal feature subset through the selection, crossover, and mutation operations to diagnose the disease at an earlier stage. This helps to identify the risk factor and diagnosing the diabetes disease efficiently. Experimental results show that the proposed IFRST-OGA increases the performance in terms of true positive rate, computation time and diabetes diagnosing accuracy.

Keywords:

Diabetic Diagnosis, Risk Factors Analysis, Rough Set Theory, Feature Selection, Optimized Genetic Algorithm, Selection, Crossover, Mutation, Optimal Feature Subset Selection

1. INTRODUCTION

Extracting knowledge and patterns for the diagnosis and treatment of disease from the medical database becomes more important to promote the development of telemedicine and community medicine. Data mining can be defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley et al. 1992). Data mining is used to extract hidden knowledge across large sets of data. After the extraction it can be used to enhance decision making in a number of areas like Marketing, Business, Medical, Education and all huge repository of information. Historical patient data of medical field is in need to determine correlations between demonstrated symptoms and disease.

The disease diagnosis of diabetes is an ongoing research area of interest to the healthcare community. While increasing world's present disease rate, accurate diagnosis of diabetes with the aid of medical data remains a significant task due to several risk factors. A risk factor is some of the characteristics of a human being that increases the probability of developing a disease. The several risk factors are associated with diabetes includes older age, overweight or obesity, physical inactivity, marital status, smoking, lower education and low income, regular medical doctor, high blood pressure, self-apparent health, Dietary factors, environmental factors and life stress.

Based on the analysis of these risk factors, the diabetes disease is diagnosed with optimal feature subset selection. Feature selection is a process for selecting a relevant features among the multiple features related with the disease. Rough set theory (RST) has been used to enable the discovery of data dependencies and finding the minimal reduction from a relational data table. Genetic algorithms are optimization methods which are based on the principles of evolution.

The aim of this research is to determine a minimal feature subset from a diabetic's domain to hold a high accuracy in representing the original features.

Contribution of the paper is described as follows,

- An Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is introduced for improving the accuracy of diabetic diagnosis. An integrated approach is to take advantage of the different performance characteristics of both Fuzzy rough set theory and optimized genetic algorithm. Initially, Fuzzy Rough Set Theory (FRST) is introduced to select the relevant features and remove the irrelevant features for dimensionality reduction. The rough set consist of the two fuzzy set namely Lower Approximation and Upper Approximation. The fuzzy rough set representation accuracy values are used to select the relevant feature based on the two approximations.
- The proposed Integration method is used to select an appropriate feature subset using Optimized Genetic Algorithm (OGA). An OGA uses three different operators namely selection, crossover and mutation. Initially, the fitness of the chromosome (i.e. feature subset) is measured based on the weight assigned for each chromosome which is verified with threshold value. After that, selection process is performed to select the chromosomes from the initial population with higher fitness value. The crossover is used to swap the two chromosomes which provides new adapted ones to find a near to optimal feature subset. Finally, Mutation is performed for randomly interchanging the bit to obtain the optimal feature subset for diabetic diagnosis.

The rest of this paper is ordered as follows: In section 2, the proposed Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is described with neat diagram. In section 3, Experimental settings are presented and the analysis of results is explained in section 4. Section 5 introduces the background and reviews the related works. Section 6 provides the conclusion of the research work.

2. LITERATURE REVIEW

The several data mining techniques were developed for feature selection and diabetes disease diagnosis as listed below.

In [1], K-Means and genetic algorithms (GA) was introduced for finding an optimal set of features. However, the resulting accuracy of diabetes diagnosis was not enhanced. A novel hybrid (GAPSO-FS and GAFOA-FS) method was introduced in [2] that use a genetic algorithm (GA) in combination with a swarm optimization algorithm (particle swarm optimization (PSO) or fruit fly optimization algorithm (FOA)) for medical diagnosis. However, it failed to extend the hybrid approach for multiclass problems with optimization technique.

A novel method was introduced in [3] for feature selection by combining improved electromagnetism-like mechanism (IEM) algorithm with the nearest neighbor classifier and opposite sign test (OST). However, it failed to optimize feature selection and parameters setting concurrently. In [4], the different factors related with diabetes, late diabetes diagnosis, and whether these factors were varied from males and females. But, certain risk factors appeared for the impacts of males and female was not analyzed effectively.

A systematic effort was accomplished in [5] by using machine learning and data mining approaches applied on Diabetes mellitus (DM). But an optimal feature selection remained unaddressed. An incremental mechanism was introduced in [6] to select the feature based on rough set model with minimum time. While varying the data values, feature selection using rough set model was not used to find the knowledge from dynamic data tables.

K-nearest neighbors algorithm with the fuzzy K-nearest neighbors was introduced in [7] for increasing the diabetic diseases diagnosis accuracy. However, it failed to perform the dimensionality reduction. A hybrid approach was developed in [8] uses Artificial Bee Colony (ABC) algorithm for feature selection. However, the time for feature selection was not at required level.

The feature selection and instance selection was performed in [9] using genetic algorithms. However, the relevant features were not accurately selected to diagnosis the disease. A Hybrid Feature Selection (HFS) technique was introduced in [10] based efficient disease diagnostic model for Breast Cancer, Hepatitis, and Diabetes. But, efficient disease diagnosis accuracy was not attained at a required level.

To decrease the feature dimensionality a feature extraction and feature selection method is used with the help of K_SVM method. But there is no optimization techniques is used [21].

In [22] attempts to identify an intelligent approach to assist disease diagnostic procedure using an optimal set of attributes instead of all attributes present in the clinical data set. But it consist only continuous attributes.

The issues are identified from above reviews such as difficult to diagnosis the diabetes disease diagnosis, lack of accuracy, higher true positive rate and difficult to find relevant features for disease prediction. In order to overcome such kind of issues, an Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is introduced.

Artificial Neural Networks was introduced in [11] for diagnosing the disease with minimum time. But an optimal feature subset selection was not performed. The proposed IFRST-OGA technique significantly performs the feature subset selection for diabetic diagnosis through the optimized genetic algorithm.

An intelligent naïve Bayes approach based system was introduced in [12] for diabetic disease diagnosis. However, it failed to distinguish the features and include different other parameters for disease diagnosis. The IFRST-OGA significantly performs the features selection and the performances of various parameters are analyzed to show the accuracy of diabetic diagnosis.

A multi-class genetic programming (GP) was designed in [13] for diagnosing the diabetes disease. But an optimal set of feature selection was not performed. The IFRST-OGA using optimized genetic algorithm for optimal feature subset selection for improving the disease diagnosing accuracy.

A multidimensional time series feature selection method was developed in [14] for obtaining the dimension reduction using combining a K-L information entropy estimation and Mutual Information. But still it was not performs dimension reduction for feature selection under the condition of improving the accuracy. The proposed IFRST-OGA performs dimensionality reduction through fuzzy rough set based feature selection.

The hybridization approach of SVM technique and K-means clustering algorithms was introduced in [15] to diabetes disease diagnose. However, it failed to use optimization techniques for improving the accuracy of diagnosing. The IFRST-OGA improves the diabetic disease diagnosing accuracy through optimized genetic algorithm.

The performance of logistic regression, artificial neural networks (ANNs) and decision tree models was explained in [16] for diagnosing the diabetes or pre-diabetes using general risk factors. But it failed select an optimal predictive models to reduce the occurrence of diabetes. The IFRST-OGA uses an Optimized genetic algorithm for feature subset selection to diagnosis the diabetes.

An Intelligent Approach with Bi-Level Dimensionality Reduction was developed in [17] for Diagnosing the Diabetes. But, the optimality during the diagnosis was not attained effectively. The IFRST-OGA improves the optimal feature subset selection through the selection, crossover and mutation operators.

In [18], for diabetes disease diagnosis the Support Vector Machines is combined with feature selection. However, higher diagnosing rate was not achieved. The IFRST-OGA performs efficient feature and optimal feature subset selection for improving the diabetes diagnosing accuracy.

Genetic Algorithm (GA) and Multilayer Perceptron Neural Network (MLP NN) was introduced in [19] for feature selection. However, early prediction and diagnosis of diabetes disease was not achieved. The proposed IFRST-OGA predicts the disease at an earlier stage through optimal feature subset selection.

A linguistic hedges neuro-fuzzy classifier with selected features (LHNFCSF) was introduced in [20] for dimensionality reduction and feature selection. However, it failed to estimate the medical diagnosis problems like gene selection. The IFRST-OGA performs efficient feature selection through genetic algorithm for diabetic disease diagnosis.

As a result, an Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is developed to find optimal feature subset for efficient pattern recognition on diabetes healthcare data.

3. INTEGRATION OF FUZZY ROUGH SET THEORY AND OPTIMIZED GENETIC ALGORITHM

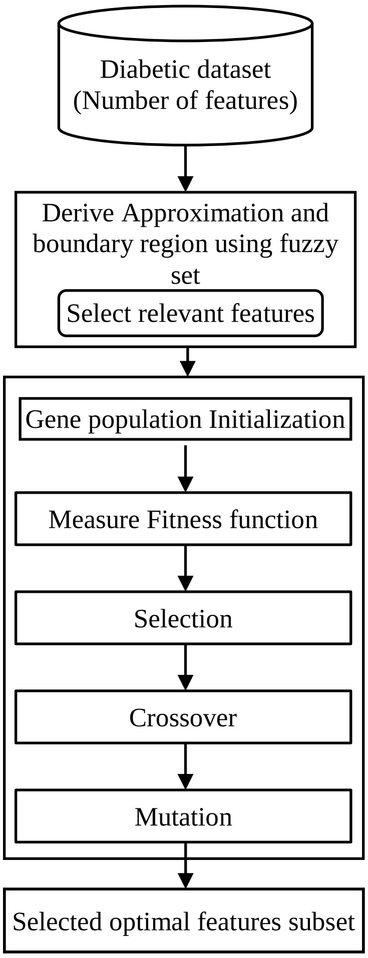

Fig.1. Flow processing diagram of the Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm

Feature subset selection is a process for pattern recognition and data mining. In real world applications, more number of features present in the training data set may increase the risk to diabetic diagnosis. This also contains the high dimensional feature space. Therefore, the feature selection is significant for dimensionality reduction. Dimensionality reduction is the process of reducing the number of irrelevant features in the dataset. The main benefits of feature selection are reducing the computation time and improving the performance of training sample sets. Rough set is used for feature selection to handle information inconsistency. In order to overcome the issues, an Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is developed to select optimal feature subset for diagnosing the disease on diabetic healthcare data. The proposed IFRST-OGA consists of two processing steps such as feature selection and feature subset selection.

The Fig.1 describes the flow processing diagram of the Integration of Fuzzy Rough Set Theory and Optimized Genetic Algorithm (IFRST-OGA). The block diagram illustrates a two-stage process for selecting an optimal feature subset. In the first stage, Rough Set theory based Feature Selection is carried out to select the disease features from the dataset. This helps to select the relevant feature and reduce the irrelevant feature for cause of diabetes without losing significant information. In second stage, Optimized Genetic Algorithm (OGA) is used in IFRST-OGA to find the feature subset for diagnosing the disease. The brief description of the IFRST-OGA is explained in forthcoming sections.

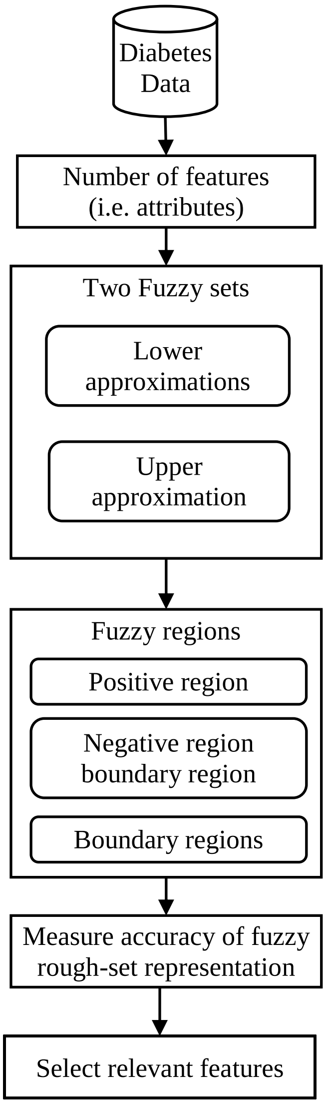

Fig.2. Fuzzy rough set theory based feature selection

3.1 FUZZY ROUGH SET THEORY FOR FEATURE SELECTION

The first step in the design of IFRST-OGA is the feature selection to find the risk factor for diagnosing the diabetes disease. A fuzzy Rough Set method is a data mining technique which is useful for feature selection and it is used for dimensionality reduction. Fuzzy Rough set-based analysis is applied in medicine application. It also reduces the complexity by minimizing the redundant disease features within the set of feature patterns. The feature selection using Fuzzy Rough Set model minimizes the redundant disease features by selecting relevant features for diabetic disease diagnosis.

As shown in Fig.2, fuzzy rough set theory (FRST) based feature selection is described. A fuzzy-rough set is a generalization of a rough set, derived from the approximation of a fuzzy set in a hard approximation. The dataset consists of several features (i.e. attributes), many of which may irrelevant or redundant. This causes a more time consumption task. In order to overcome such kind of issue, the FRST model is applied. As a result, this helps for Dimensionality reduction which diminishes the data size by removing the irrelevant features.

Fuzzy rough set theory in IFRST-OGA for feature selection whose feature has degree of membership function. A fuzzy rough set theory performs the evaluation of the membership of elements in a set; this is explained with the help of a membership function valued in the interval range [0, 1]. By applying the feature selection analysis, the related features are estimated to diagnosis the optimal risk features. A fuzzy set uses a membership functions which takes the values either 0 or 1. The fuzzy membership function values are $0 \leq \mu_x(A) \leq 1$. FRST is represented by two fuzzy sets such as lower and upper approximation. Fuzzy Rough set based feature Selection is efficiently used to reduce the discrete and real valued noisy risk features without any user information.

Let us consider A is a fuzzy set in 'X' hence $X \rightarrow [0,1]$, number of features in dataset $F = f_1, f_2, f_3, \ldots, f_n$. Among the multiple features, the lower and higher approximation set is used for selecting the features which is a member of an approximation. The fuzzy set approximation is illustrated in Fig.3.

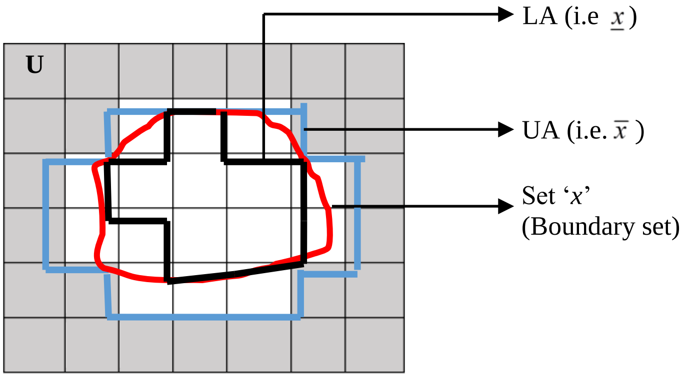

Fig.3. Fuzzy rough set approximation

The Fig.3 shows the fuzzy rough set approximation to identify the features which is relevant to analysis the risk on diabetes diagnosis. As shown in Fig.3, the rough set (x) contains the two fuzzy set namely Lower Approximation (LA) and Upper Approximation (UA). The lower approximation includes all features that definitely belong to the set, while the upper approximation includes all features that may be a member of the set. The difference between the two approximations symbolizes the rough set boundary region (x). The lower and upper approximation is defined as,

$$\mu_{\underline{x}}(A) = Inf_x I\left(\mu_F(A), \mu_x(A)\right) \quad (1)$$

$$\mu_{\overline{x}}(A) = Sup_x T\left(\mu_F(x), \mu_x(A)\right) \quad (2)$$

$$\text{Rough set boundary region } (x) = \mu_{\underline{x}}(A) - \mu_{\overline{x}}(A) \quad (3)$$

From Eq.(1), Eq.(2), $\mu_{\underline{x}}(A)$ and $\mu_{\overline{x}}(A)$ denotes a lower approximation and upper approximation of fuzzy set. Inf_x is the infimum (i.e) lower approximation set. sup_x denotes a supremum (i.e. upper approximation). I and T is the norm operator and implicatator respectively. $\mu_x(A)$ is a fuzzy membership in a set x and $\mu_F(x)$ is a feature set.

Let us consider $\langle \mu_{\underline{x}}(A), \mu_{\overline{x}}(A) \rangle$ is a lower and upper approximation which is called as a fuzzy rough set; hence, a rough set is comprises two fuzzy sets, one represents a lower boundary of the target set x and the other sets denotes an upper boundary of the target set x. Therefore, the accuracy of the fuzzy rough set representation is described as,

$$A = \frac{\left|\mu_{\underline{x}}(A)\right|}{\left|\mu_{\overline{x}}(A)\right|} \quad (4)$$

From Eq.(4), A denotes an accuracy of the rough set representation of set x. Accuracy of the rough set is a fuzzy member ship function which provides the value as 0 and 1which is expressed as follows,

$$A = \begin{cases} 1 & if \left|\mu_{\underline{x}}(A)\right| = \left|\mu_{\overline{x}}(A)\right| \\ 0 & otherwise \end{cases} \quad (5)$$

The Eq.(5), clearly shows that when the upper and lower approximations are equal (i.e., boundary region empty), then the accuracy value provides 1, and the approximation is perfect and the feature within this approximation is a relevant otherwise it is not a relevant feature. These relevant features are selected to perform the diabetic diagnosis. The fuzzy rough set based feature selection algorithm is described as follows,

Algorithm 1: Fuzzy rough set based feature selection algorithm

Input: Datasets, Number of features $F = f_1, f_2, f_3, \ldots, f_n$.
Output: Improve feature selection
Step 1: Begin
Step 2: For each feature in dataset
Step 3: Measure lower approximation to identify the feature within the boundary using Eq.(1)
Step 4: Measure upper approximation to identify the feature using Eq.(2)
Step 5: Measure the accuracy of fuzzy rough set representation using Eq.(4)
Step 6: if fuzzy rough set representation is 1 then
Step 7: Select the relevant feature
Step 8: else
Step 9: Not a relevant feature
Step 10: End if
Step 11: End for
Step 12: End

Algorithm 1 shows the fuzzy rough set based feature selection algorithm to identify the relevant feature on diabetes disease diagnosis. By applying the fuzzy rough set theory for each feature in the dataset, the two fuzzy set lower and upper approximations are measured for verifying the feature within the fuzzy set or not. Then the accuracy of the feature representation value is calculated. If the accuracy of feature representation value is '1', then the feature is said to be a relevant feature otherwise it is an irrelevant feature. As a result, the relevant features are chosen to find the risk factors to recognize the diabetes disease pattern. This helps to reduce the computation time and also used for dimensionality reduction.

3.2 OPTIMIZED GENETIC ALGORITHM BASED FEATURE SUBSET SELECTION

Once the relevant feature is selected, an optimal feature selection is performed using optimized genetic algorithm. Optimal feature selection is performed by calculating the multi relevance between the selected features. Multi relevance feature selection is a process of selecting the optimal features with large dependence between them to analyze diabetic risk factors. The Multi relevant feature selection is frequently used in IFRST-OGA technique to accurately identify resultant selected features (i.e. feature subset) and it is usually described with relevant feature selection. The feature subset selection is a process of selecting a set of optimal features with multi relevance from already selected features. The block diagram of the Optimized genetic algorithm is shown in Fig.4.

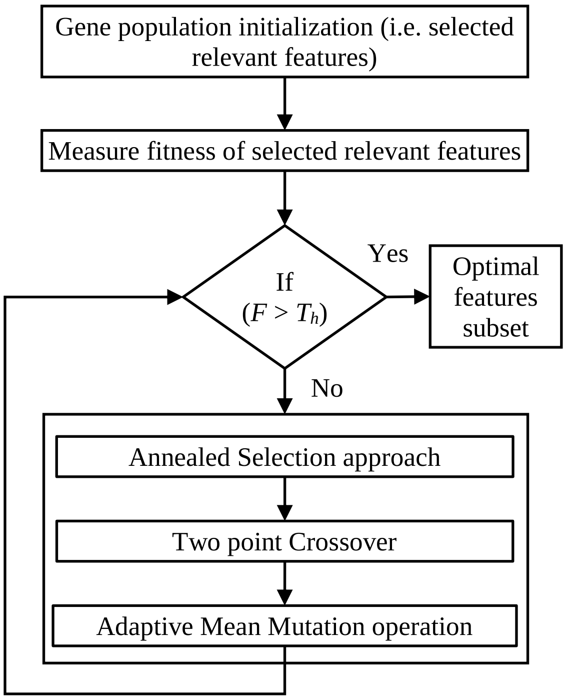

Fig.4. Block diagram of optimized genetic algorithm

The Fig.4 describes the Block diagram of optimized genetic algorithm. Selected relevant features are optimized using genetic algorithm to obtain final resultant features (i.e. subset) to analyze the diabetic risk factors. In OGA, each relevant selected feature is considered as a gene. The chromosome (i.e. optimal feature subset) contains the genetic information. The feature subset selection is a task of selecting the significant features subset to find diabetic risk factors. Genetic algorithms are generally used to generate high-quality results for optimization by using the unique operators such as Annealed Selection approach, two point crossover and Adaptive Mean Mutation. Initially, the gene population initialization is carried out for generating many individuals (i.e., relevant features selected from Fuzzy rough set theory) in random manner. The number of relevant features selected in an initial population is a major concern for improving the optimization. A large population suffers from slower convergence whereas a very small population uses minimum search space and it may converge to a local extreme.

In each population generation, the fitness of every relevant feature in the population is calculated. Generally, the fitness is the quality of being suitable to accomplish a particular task in the optimization. In OGA, the fitness function is measured based on the Multi relevance between the features. Let us considered the selected relevant features are f_1, $f_2 \ldots f_n$ given as input of the optimized genetic algorithm. In OGA, selected relevant features are considered as gene. The multi relevance of gene is measured for optimizing the risk factor. The Multi relevance between the relevant features is expressed as follows,

$$\text{Fitness Function } (F) = Max \sum D\left(f_1, f_2\right) \tag{6}$$

From Eq.(6), *Max* denotes a maximum function that provides more relevance and $D(f_1,f_2)$ is the dependence between the two features f_1 and f_2. From Eq.(6), fitness function (F) is measured based on the multi relevance between the two features to obtain the optimal subset features. This helps to identify the major risk factor for causing diabetes. Then the fitness is verified with the threshold value (T_h). If fitness value is greater than the threshold value (T_h), optimal feature is selected to identify the risk factor for diabetes. If the fitness value is less than the threshold value, then the genetic operator's such as Annealed Selection, two point crossover and adaptive mean mutation are carried out to select a final optimal resultant feature (i.e. subset). The processes of the genetic operators are explained as below subsections.

3.2.1 Annealed Selection Approach:

Annealed Selection approach in OGA is used to select the individual (i.e. selected relevant features) from the population for crating the next generation. The individual's selection is used to create consecutive generations in an optimized genetic algorithm. By applying Selection approach, fitness of each individual is measured. Selection probability of each individual is calculated based on the fitness value. When the population generation changes, the fitness value as well as selection probability of each individual gets varied. Therefore, the probability of the selecting the individuals is calculated as follows,

$$P\left(f_i\right) = \frac{f_i}{\sum_{i=1}^{n} f_i} \tag{7}$$

From Eq.(7), f_i denotes the Average Fitness of the population in all generation using Annealed Selection approach and n denotes the total number of individuals.

3.2.2 Two Point Crossover:

The new operator used in optimized genetic algorithm is a Two point Crossover. A Two point Crossover operator used to adjust chromosomes from one generation to the next generation. Crossover process considers a more than one parent chromosomes and producing an offspring from them. Crossover is applied on the two selected individuals (i.e. relevant features) along with a probability which is called the crossover probability. The Two-point crossover called as two points which is selected based on the parent individuals. Swapping of the chromosomes provides the two offspring's. Let us consider two selected relevant features for detecting the maximum relevance between them (i.e. f_1, f_2) f_1 = 1011101 and f_2 = 1101011. Each chromosome has a binary string values.

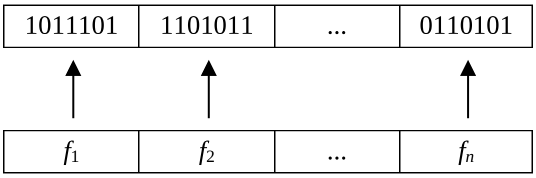

Fig.5. Representation of chromosomes with selected features

The Fig.5 clearly illustrates that the chromosome representation with selected relevant features. From the figure, f_1, $f_2 \ldots f_n$ are the selected relevant features and the length is varied along with the size of the number of total features and the number of selected features. But the length of the chromosome is equal size for each chromosome. After the representation, the crossover is performed between the two chromosomes to generate new offspring's. Therefore, the crossover point is selected randomly among the chromosomes' as shown in Fig.6.

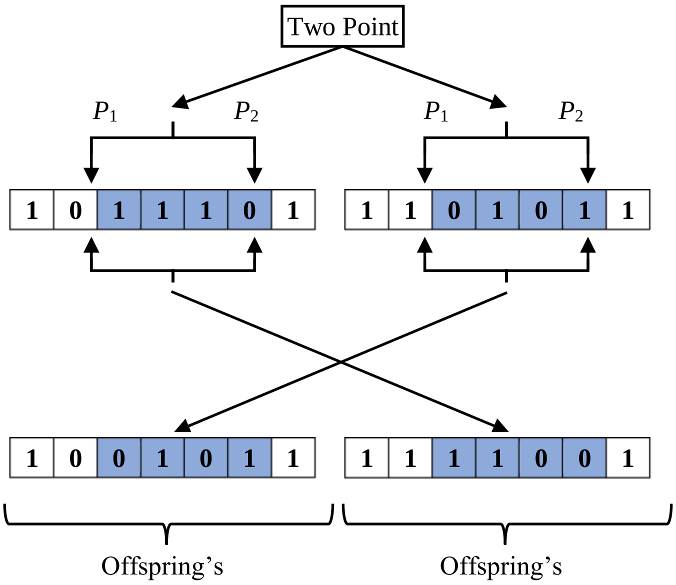

Fig.6. Two point Crossover

The Fig.6 illustrates that the two point crossover to swap the chromosomes for obtaining the best offspring's in order to select optimal features subset. The crossover is performed based on the combination of two well defined chromosomes provides the new adapted ones. Crossover is the major process which helps to the investigation of the feature space in order to find a near to optimal resultant feature. When all the individuals of the population are equal, the use of the crossover operator on two chromosomes generates the same chromosomes. This means that the two point crossover process is not able to create diversity within a population. Therefore, the diversity is maintained by using adaptive mutation operator.

3.2.3 Adaptive Mean Mutation Operator:

Once offspring's are generated using two point crossover operator, the adaptive mean mutation is performed to maintain the diversity within the population. The adaptive mutation operation consists randomly changing the value of each bit of the chromosome along with the probability which is called the mutation probability. Mutation takes place in evolution depending on probability function. The value of the mutation probability is used for maximizing the probability that the genetic algorithm discover the optimum relevant features of the objective function under simple assumptions. Adaptive Mean Mutation is carried out by generating offspring from parent for selecting the optimal multi relevant features. The new offspring is created as follows,

$$y_i' = y_i + \sigma_1 N(0,1) + \sigma_2 C(0,1) \tag{8}$$

From Eq.(8), y_i' newly created offspring value from the population and y_i represents the offspring created from crossover. σ_1 and σ_2 represents a standard deviation parameter. $N(0,1)$ denotes a Gaussian normally distributed random variable with mean 0 and variance 1 and $C(0,1)$ represents a cache operator. Generally, mutation probability consist very low value. While considering the binary chromosome value, the adaptive mean mutation inverting the value of a bit from 0 to 1, or vice versa. Mutation is a process for controlling the genetic diversity from one generation of chromosomes to the new generation. Let us taken as the offspring value taken from the two point crossover operation to perform mutation. The mutation process is described as follows,

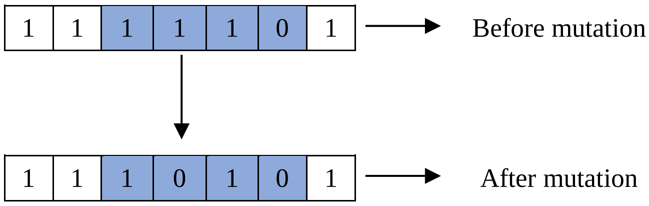

Fig.7. Adaptive Mean Mutation operation

The Fig.7 shows the Adaptive Mean Mutation operation, the offspring chromosome value of string '1' is altered randomly with the exact chromosome string of '0'. Mutation is in the population, a randomly interchanging the bit for creating the new offspring to obtain the optimal selected feature with multi relevance. The optimized genetic algorithm is described as follows.

Algorithm 2 Optimized genetic algorithm based feature subset selection

Input: Selected relevant features $f_1, f_2 \ldots f_n$
Output : Optimal selected relevant features (i.e. feature subset)
Step 1: Begin
Step 2: Initialize the gene population
Step 3: Calculate fitness for selected relevant features using Eq.(6)
Step 4: if $F > T_h$ then
Step 5: Arrive optimal selected features
Step 6: else
Step 7: Perform Annealed Selection approach using Eq.(7)
Step 8: Perform two point crossover generates new off spring chromosome
Step 9: Perform Adaptive Mean Mutation operation using Eq.(8) to select optimal relevant feature
Step 10: Terminate the algorithm until the specified condition is satisfied or else go to Step 3.
Step 11: End if
Step 12: End

Algorithm 2 describes the optimized Genetic Algorithm (OGA) to select the optimal feature selection to find the optimum diabetic risks factors. For each selected relevant feature, the population generation is carried out randomly. After that, the multi relevance between the selected features is measured to calculate the fitness of each selected feature. If the fitness value is greater than the threshold value, then the iteration gets stopped and arrive the optimal resultant features (i.e. feature subset). Otherwise, it performs genetic operators such as Annealed Selection approach, two point crossover and Adaptive Mean Mutation operation in order to select an optimal feature subset for identifying the diabetic risk factors on healthcare data. This process is repeated until an optimal solution is attained. This helps for IFRST-OGA efficiently discovering the diabetic risk factors with less number of selected optimal features.

4. EXPERIMENTAL SETTINGS

An Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is introduced is experimented using JAVA language with Weka tool for selecting the optimal feature subset for analyzing the risk on diabetes disease. Two dataset namely Diabetes Data Set and Pima Indians Diabetes Data Set are taken from the UCI machine learning repository to perform the experimental.

In Diabetes dataset, Diabetes patient records are attained from two basic resources namely automatic electronic recording device and paper records. An automatic device consist internal clock for timestamp events analysis. The paper record includes the logical time slot i.e. breakfast, lunch, dinner, bedtime. The fixed times are allocated to breakfast (08:00), lunch (12:00), dinner (18:00), and bedtime (22:00). Therefore, the paper record includes uniform recording times whereas electronic records have more practical time stamps. The diabetes patient information file consists of four files. Each file is separated by a record and each record is separated by a newline. The date file format is described as MM-DD-YYYY, Time in XX: YY format, code and Value. The Diabetes dataset contains 20 attributes.

The Pima Indians Diabetes Data Set consists of 8 attributes with class value. All the attributes are numeric valued. The dataset provided data from diabetes cases in which all patients were females as a minimum 21 years of age. The Pima Indians Diabetes dataset consists of 768 samples (i.e. instances). Out of 768, normal samples considered as 500 and 268 diabetic samples represented by 8 attributes. The proposed Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is compared with existing K-Means with Genetic Algorithms (*k*-means-GA) [1] and Hybrid method (i.e. GAPSO-FS and GAFOA-FS) [2]. The experimental evaluation is carried out in terms of true positive rate, computation time and diabetic diagnosis accuracy.

5. RESULT ANALYSES

Result analysis of Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is described in this section. The IFRST-OGA is compared against the existing K-Means with Genetic Algorithms (*k*-means-GA) [1] and Hybrid method [2]. The experiment is conducted on the factors such as true positive rate, computation time and diabetic diagnosis accuracy with four dataset Diabetes Data Set, Pima Indians Diabetes Data Set, UCI and Diabetics 130 US. Experimental results are compared and analyzed with the help of table and graph.

5.1 IMPACT OF TRUE POSITIVE RATE

True positive rate is defined as the ratio of the number of (i.e. no. of) relevant features are selected to the number of relevant features selected and irrelevant feature incorrectly rejected. It is measured in terms of percentage (%).

$$TPR = \frac{TP}{TP + FN} \times 100 \tag{7}$$

From Eq.(7), *TPR* is the true positive rate, *TP* is the true positive that the features are correctly selected and *FN* denotes a number of irrelevant feature which is incorrectly rejected.

Table.1. Tabulation for true positive rate

No. of features	True positive rate (%) for Diabetes dataset		
	IFRST-OGA	***k*-means-GA**	**Hybrid method**
3	67.36	52.68	35.24
6	82.68	65.35	40.12
9	88.12	70.12	43.65
12	92.36	75.42	50.12
15	96.78	80.12	58.68
18	98.35	86.36	65.36

The Table.1 shows the true positive rate using diabetes dataset with three different methods IFRST-OGA, *k*-means-GA [1] and Hybrid method [2]. As shown in Table.1, the true positive rate of the proposed IFRST-OGA is considerably increased while increasing the number of features than the existing methods.

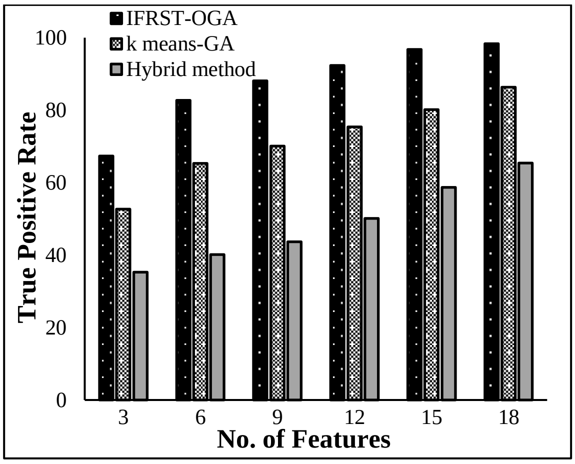

Fig.8. Measure of true positive rate using Diabetes database

As shown in Fig.8, performance analysis of true positive rate is measured using Diabetes dataset. While increasing the number of features (i.e. attributes), the true positive rate gets increased in all the three methods. But comparatively, the proposed IFRST-OGA achieves higher true positive rate. This is because; the Fuzzy Rough Set Theory (FRST) is used in IFRST-OGA for feature selection. By applying FRST based feature selection, the relevant features are selected to diagnosis the diabetes with optimal risk features. A fuzzy set uses a membership functions which takes the values either 0 or 1. The FRST contains the two fuzzy sets lower and upper approximation. By using this approximation, the features which are located inside and outside the set are identified. With the help of the approximation, the accuracy representation is identified to select the relevant features and reject the irrelevant features to identify the optimum risk on diabetes data. Therefore, the true positive rate is increased by 23% and 57% compared to existing k means-GA [1] and Hybrid method [2]. The Table.2 provides tabulation for true positive rate.

Table.2. True positive rate between different algorithms

No. of features	True positive rate (%) for Pima Indians Diabetes Data Set		
	IFRST-OGA	k-means-GA	Hybrid method
2	53.65	48.12	47.65
4	66.67	52.36	50.13
6	80.36	62.36	58.65
8	88.72	76.45	64.53

Using Pima, based on number of features, the features are varied from 2 to 8. The true positive rate during the feature selection is increased in proposed IFRST-OGA when compared to existing *k*-means-GA [1] and Hybrid method [2].

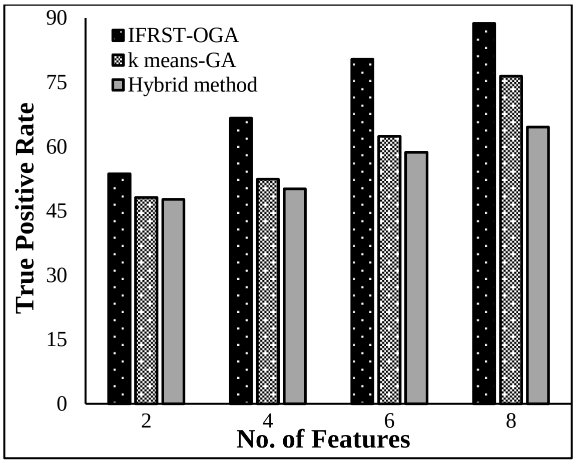

Fig.9. Measure of true positive rate using Pima Indians Diabetes Data Set

The Fig.9. clearly shows the true positive rate using Pima Indians Diabetes Data Set with number of features. By using Pima Indians Diabetes Data Set, the true positive rate of proposed IFRST-OGA is higher than the existing methods. By applying the fuzzy rough Set theory, the number of features is correctly selected as relevant and removes the irrelevant features to diagnosis the diabetic disease accurately. Therefore, the true positive rate is increased by 21% and 30% using IFRST-OGA compared to existing *k*-means-GA [1] and Hybrid method [2] respectively.

Table.3. True positive rate for Diabetics dataset

No. of features	True positive rate (%) for Diabetics 130 US dataset		
	IFRST-OGA	k-means-GA	Hybrid method
2	65.25	52.06	38.42
4	80.08	65.35	40.12
5	84.12	76.03	40.06
8	86.36	75.42	50.12
10	89.70	80.12	68.86
12	92.07	86.36	65.36

The Table.3 shows the true positive rate using Diabetics 130 US dataset with three different methods IFRST-OGA, *k*-means-GA [1] and Hybrid method [2]. As per the value in Table.3, the true positive rate of the proposed IFRST-OGA is considerably increased while increasing the number of features than the existing methods.

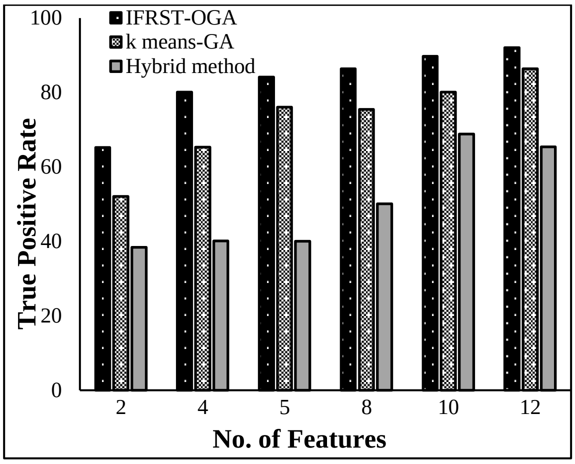

Fig.10. Measure of true positive rate using Diabetics 130 US

As shown in Fig.10, the performance analysis of true positive rate is measured using Diabetes dataset. While increasing the number of features (i.e. attributes), the true positive rate gets increased in all the three methods.

Table.4. True positive rate for UCI dataset

No. of features subset	Diabetic diagnosing accuracy (%)		
	UCI		
	IFRST OGA	K Means GA	Hybrid Method
2	75.08	56.12	72.04
5	84.05	72.34	74.23
8	81.02	67.47	79.3
12	78.45	75.23	72.34
14	71.23	60.36	52.31

The Table.4 shows the true positive rate using UCI dataset with three different methods IFRST-OGA, *k*-means-GA [1] and Hybrid method [2]. As per the value in Table.4, the true positive rate of the proposed IFRST-OGA is considerably increased than the existing methods.

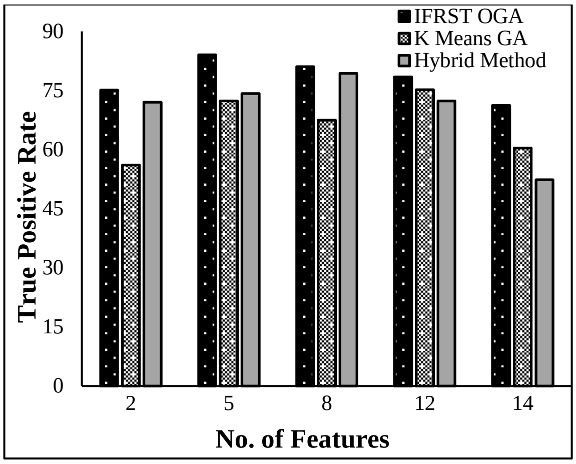

Fig.11. Measure of true positive rate using UCI

The Fig.11 clearly shows the true positive rate using UCI Diabetes Data Set with number of features. By using UCI Diabetes Data Set, the true positive rate of proposed IFRST-OGA is higher than the existing methods.

5.2 IMPACTS OF COMPUTATION TIME

Computation time is defined as the amount of time required to select an optimal feature subset for diabetic disease diagnosis. The formula for computation time is expressed as follows,

$$CT = No\ of\ \ features \times Time\left(select\ the\ features\ subset\right) \quad (8)$$

From Eq.(8), CT is the computation time which is measured in terms of millisecond (ms). Followed by this, the related features subsets are effectively selected.

Table.5. Tabulation for Computation time

No. of features	Computation time (ms) for Diabetes dataset		
	IFRST-OGA	***k*-means-GA**	**Hybrid method**
3	8.12	12.35	15.21
6	10.24	15.65	18.64
9	12.37	18.24	22.35
12	15.10	20.16	25.69
15	18.54	23.41	28.46
18	21.97	27.46	32.14

The Table.5 shows the measurement of computation time using diabetes dataset with different methods namely IFRST-OGA, *k*-means-GA [1] and Hybrid method [2]. As per the values of Table.5, it is clearly illustrates that the computation time of proposed IFRST-OGA is reduced when compared to the existing methods.

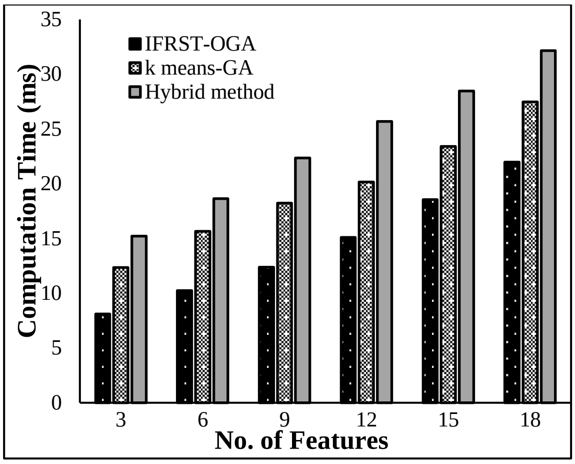

Fig.12. Measure of computation time using Diabetes dataset

The Fig.12 shows the measurement of computation time using diabetes dataset with respect to number of features. While considering the diabetes dataset, the number of relevant features and the optimal feature subset are selected to find the best possible risk factors for diabetes diagnosis. The fuzzy rough set theory is used in IFRST-OGA for selecting the relevant features. With the selected features, the features subset is identified and selected by optimized genetic algorithm (OGA). By applying the OGA, the weight value of feature subset (i.e. chromosomes) is calculated. Then the weight of the each chromosome is compared with the threshold value. If the weight value of chromosome is greater than the threshold value, then the optimal feature subset is selected. Otherwise, the genetic operator's such as crossover, mutation is carried out to create the new chromosomes. As a result, the features with more optimal subset are selected for maximizing the accuracy with minimum computation time. Therefore, the computation time is reduced by 28% and 41% when compared to existing k means-GA [1] and Hybrid method [2] respectively.

Table.6. Tabulation for Computation time using Pima Indians Diabetes Data Set

No. of features	Computation time (ms) for Pima Indians Diabetes Data Set		
	IFRST-OGA	**k means-GA**	**Hybrid method**
2	10.35	13.81	17.55
4	12.41	16.58	20.17
6	14.75	20.12	24.45
8	17.44	21.47	26.51

The Table.6 describes the computation time for feature subset selection using Pima Indians Diabetes Data Set with number of features. An optimal feature subset selection time is considerably reduced in proposed IFRST-OGA when compared to existing k-means-GA [1] and Hybrid method [2].

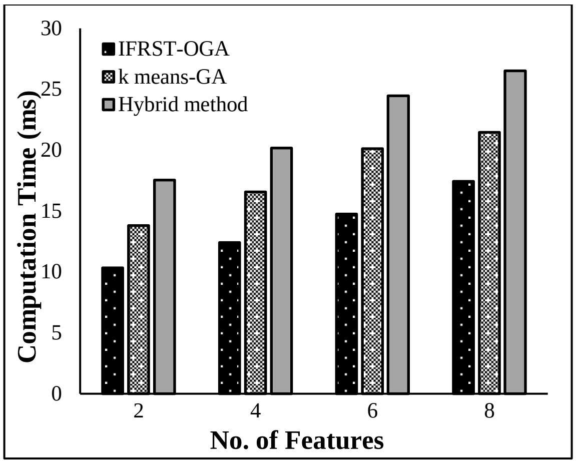

Fig.13. Measure of computation time using Pima Indians Diabetes Data Set

The Fig.13 shows the measure of computation time using Pima Indians Diabetes Data Set with respect to number of features as input. As per the values of Table.6, it is clearly illustrates that the computation time using proposed IFRST-OGA is reduced when compared to the other existing methods [1] [2]. This is due to an optimal feature subset is selected among the multiple features using optimized genetic algorithm. The optimal feature subset selection is a most significant approach for identifying and selecting the useful subset of features to find optimum risk patterns from a larger set. Therefore, the IFRST-OGA selects more optimal feature subset with minimum computation time. The computation time is reduced by 24% and 38% compared to existing k-means-GA [1] and Hybrid method [2] respectively.

Table.7. Tabulation for Computation time using Diabetes 130 US Data Set

No. of features	Computation time (ms) for Diabetes 130 Data Set		
	IFRST-OGA	k means-GA	Hybrid method
2	10.05	12.81	15.05
4	10.01	14.58	18.08
6	13.65	20.12	14.06
8	16.02	18.47	26.51

The Table.7 describes the computation time for feature subset selection using Diabetes 130 US Data Set with number of features. An optimal feature subset selection time is considerably reduced in proposed IFRST-OGA when compared to existing *k*-means-GA [1] and Hybrid method [2].

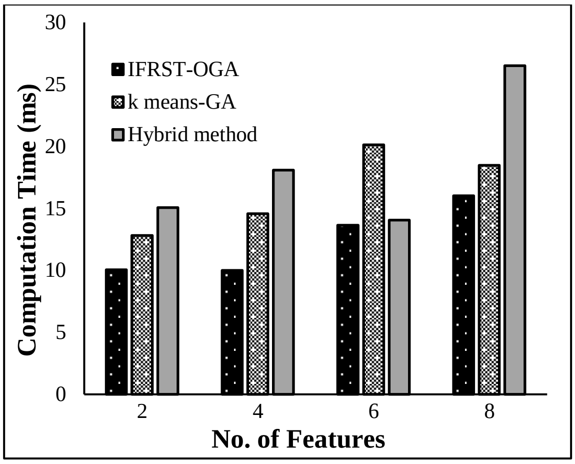

Fig.14. Measure of computation time using Diabetes 130 US Data Set

It is clearly illustrates that the computation time using proposed IFRST-OGA is reduced when compared to the other existing methods [1] [2]. This is due to an optimal feature subset is selected among the multiple features using optimized genetic algorithm.

Table.8. Tabulation for Computation time using UCI Data Set

No. of features	Computation time (ms) for UCI Data Set		
	IFRST-OGA	k means-GA	Hybrid method
2	11.18	15.35	13.76
4	12.15	16.85	14.18
6	10.65	12.34	14.06
8	15.45	18.47	17.45

The Table.8 describes the computation time for feature subset selection using UC I Data Set with number of features. An optimal feature subset selection time is considerably reduced in proposed IFRST-OGA when compared to existing *k*-means-GA [1] and Hybrid method [2].

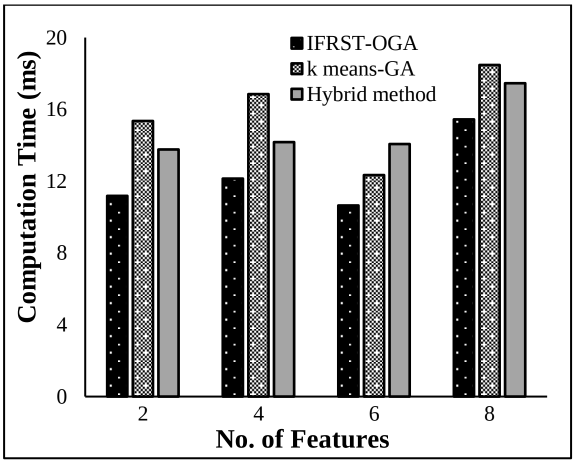

Fig.15. Measure of computation time using UCI Data Set

The computation time using proposed IFRST-OGA is reduced when compared to the other existing methods [1] [2]. This is due to an optimal feature subset is selected among the multiple features using optimized genetic algorithm.

5.3 IMPACT OF DIABETIC DIAGNOSING ACCURACY

Diabetes diagnosing accuracy is defined as the ratio of number of optimal feature subset are correctly selected and incorrectly selected for diagnosis to the total number of feature subset. The formula for diabetic diagnosing accuracy is expressed as follows,

$$DDA = \frac{FSCS + FSIS}{No.\ of\ \ features\ subsets} \times 100 \qquad (9)$$

From Eq.(9), DDA is the Diabetic Diagnosing Accuracy, FSCS is the number of optimal feature subset correctly selected for cause of diabetes and FSIS is the number of feature subset incorrectly selected for cause of diabetes. It is measured in terms of percentage (%). Higher diagnosing accuracy, the method is said to be more efficient.

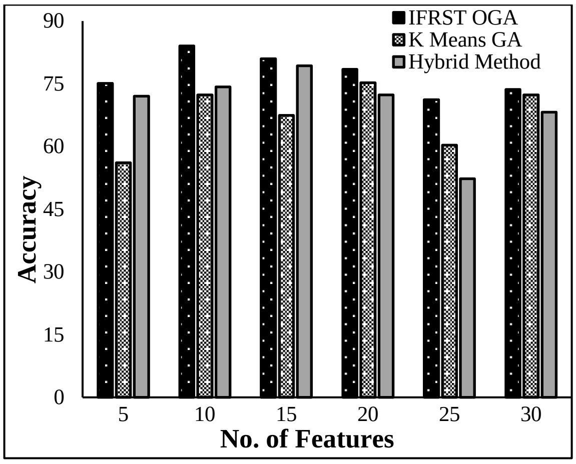

Fig.16. Measure of diabetic diagnosing accuracy using Diabetes Data Set

The Fig.16 illustrates the diabetic diagnosing accuracy using Diabetes Data Set. As shown in the figure, the diabetic diagnosing accuracy is considerably increased while increasing the number of features subset. This is because; the FRST used for selecting the relevant feature with the help of fuzzy sets namely lower and upper approximation. Based on the selected features, an optimal feature subset is selected by applying the optimized genetic algorithm. Each selected features are considered as gene. Each gene has a number of chromosomes. From which, the optimal chromosomes are selected for identifying the disease. An optimal feature subset is selected by performing the genetic operator's such as selection, crossover, and mutation. This helps to correctly select the optimal feature subset among the multiple feature subsets and reduce the incorrectly selected optimal feature subset in dataset. This helps to significantly improve the diabetic diagnosing accuracy. The accuracy of the proposed IFRST-OGA is increased by 24% and 35% compared to existing *k*-means-GA [1] and Hybrid method [2] respectively.

The Fig.17 illustrates the measurement of Diabetic diagnosing accuracy using Pima Indians Diabetes Data Set. The comparison is made with proposed IFRST-OGA and two different existing *k*-means-GA [1] and Hybrid method [2]. The experimental result shows that the diabetic diagnosing accuracy is increased in proposed IFRST-OGA. The optimized genetic algorithm provides optimal set of feature subset effective disease diagnosis. The final genetic operation namely mutation provides the more accurate results in the selection of optimal feature subset. Therefore, the Diabetic diagnosing accuracy is considerably increased by 24% and 38% compared to existing *k*-means-GA [1] and Hybrid method [2] respectively.

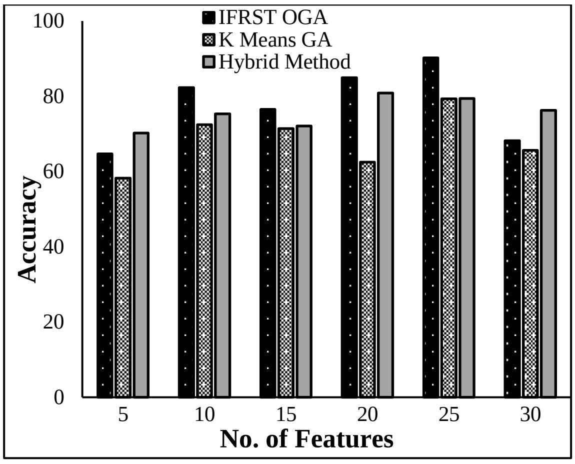

Fig.17. Measure of diabetic diagnosing accuracy using Pima Indians Diabetes Data Set

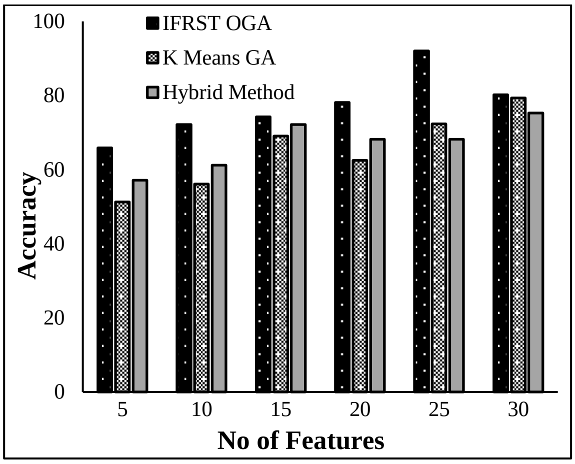

Fig.18. Measure of diabetic diagnosing accuracy using Diabetes US 130 Data Set

The comparison is made with proposed IFRST-OGA and two different existing *k*-means-GA [1] and Hybrid method [2]. The experimental result shows that the diabetic diagnosing accuracy is increased in proposed IFRST-OGA. Therefore, the Diabetic diagnosing accuracy is considerably increased by 25% and 35% compared to existing *k*-means-GA [1] and Hybrid method [2] respectively.

The Diabetic diagnosing accuracy is considerably increased by 25% and 35% compared to existing *k*-means-GA [1] and Hybrid method [2] respectively.

6. CONCLUSION

An efficient data mining technique called an Integration of Fuzzy Rough Set Theory and Optimized Genetic algorithm (IFRST-OGA) is introduced for optimal feature subset selection on diabetic data. Two processing steps are used in IFRST-OGA. Initially, Fuzzy rough set theory based feature selection is performed to select the relevant feature and remove the irrelevant feature. This helps to obtain the dimensionality reduction for efficient diabetic disease diagnosis. Secondly, optimized genetic algorithm (OGA) is applied to select an optimal feature subset for improving the diabetic diagnosing accuracy with minimum computation time. Experimental evaluation is performed with four different dataset namely Diabetes Data Set, Pima Indians Diabetes Data Set, UCI and Diabetes US 130 Data Set. The performance results shows that the proposed IFRST-OGA significantly improves the diabetic disease diagnosing accuracy, true positive rate and also reduces the computation time than the state-of-the-art methods.

REFERENCES

[1] T. Santhanam and M.S Padmavathi, "Application of K-Means and Genetic Algorithms Dimension Reduction by Integrating SVM for Diabetes Diagnosis", *Procedia Computer Science*, Vol. 47, pp. 76-83, 2015.

[2] Fei Ye, "Evolving the SVM Model based on a Hybrid Method using Swarm Optimization Techniques in Combination with a Genetic Algorithm for Medical Diagnosis", *Multimedia Tools and Applications*, pp. 1-30, 2016.

[3] Kung-Jeng Wang, Angelia Melani Adrian, Kun-Huang Chen , Kung-Min Wang, "An Improved Electromagnetism-like Mechanism Algorithm and its Application to the Prediction of Diabetes Mellitus", *Journal of Biomedical Informatics*, Vol. 54, pp. 220-229, 2015.

[4] Madonna M. Roche and Peizhong Peter Wang, "Factors Associated with a Diabetes Diagnosis and Late Diabetes Diagnosis for Males and Female", *Journal of Clinical and Translational Endocrinology*, Vol. 1, pp. 77-84, 2014.

[5] Ioannis Kavakiotis, Olga Tsave , Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas and Ioanna Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", *Computational and Structural Biotechnology Journal*, Vol. 15, pp. 104-116, 2017.

[6] Jiye Liang, Feng Wang, Chuangyin Dang and Yuhua Qian, "A Group Incremental Approach to Feature Selection Applying Rough Set Technique", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 2, pp. 294-308, 2014.

[7] Mohamed Amine Chikh, Meryem Saidi and Nesma Settouti, "Diagnosis of Diabetes Diseases using an Artificial Immune Recognition System with Fuzzy K-Nearest Neighbor", *Journal of Medical Systems*, Vol. 36, No. 5, pp. 2721-2729, 2012.

[8] Mustafa Serter Uzer, Nihat Yilmaz and Onur Inan, "Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification", *The Scientific World Journal*, Vol. 2013, pp. 1-10, 2013.

[9] Chih-Fong Tsai, William Eberle and Chi-Yuan Chu, "Genetic Algorithms in Feature and Instance Selection", *Knowledge-Based Systems*, Vol. 39, pp. 240-247, 2013.

[10] Divya Tomar and Sonali Agarwal, "Hybrid Feature Selection Based Weighted Least Squares Twin Support Vector Machine Approach for Diagnosing Breast Cancer, Hepatitis, and Diabetes", *Advances in Artificial Neural Systems*, Vol. 2015, pp. 1-10, 2015.

[11] Filippo Amato, Alberto Lopez, Eladia Maria Pena-Mendez, Petr Vanhara, Ales Hamp and Josef Havel, "Artificial Neural Networks in Medical Diagnosis", *Journal of Applied Biomedicine*, Vol. 11, pp. 47-58, 2013.

[12] Abid Sarwar and Vinod Sharma, "Intelligent Naive Bayes Approach to Diagnose Diabetes Type-2", *International Journal of Computer Application*, Vol. 3, pp. 14-16, 2012.

[13] A. Pradhan, G.R. Bamnote, Vinit Tribhuvan, Kiran Jadhav, Vijay Chabukswar, Vijay Dhobale, "A Genetic Programming Approach for Detection of Diabetes", *International Journal of Computational Engineering Research*, Vol. 2, No. 6, pp. 91-94, 2012.

[14] Liying Fang, Han Zhaoa, Pu Wanga, Mingwei Yud, Jianzhuo Yana, Wenshuai Cheng and Peiyu Chen, "Feature Selection Method based on Mutual Information and Class Separability for Dimension Reduction in Multidimensional Time Series for Clinical Data", *Biomedical Signal Processing and Control*, Vol. 21, pp. 82-89, 2015.

[15] Ahmed Hamza Osman and Hani Moetque Aljahdali, "Diabetes Disease Diagnosis Method based on Feature Extraction using K-SVM", *International Journal of Advanced Computer Science and Applications*, Vol. 8, No. 1, pp. 236-244, 2017.

[16] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang and Qing Liu, "Comparison of Three Data Mining Models for Predicting Diabetes or Prediabetes by Risk Factors", *Kaohsiung Journal of Medical Sciences*, Vol. 29, pp. 93-99, 2013.

[17] Razieh Sheikhpour and Mehdi Agha Sarram, "Diagnosis of Diabetes Using an Intelligent Approach Based on Bi-Level Dimensionality Reduction and Classification Algorithms", *Iranian Journal of Diabetes and Obesity*, Vol. 6, No. 2, pp. 74-84, 2014.

[18] Fatma Patlar Akbulut and Aydın Akan, "Support Vector Machines Combined with Feature Selection for Diabetes Diagnosis", *Istanbul University-Journal of Electrical and Electronics Engineering*, Vol. 17, No. 1, pp. 3219-3225, 2017.

[19] Dilip Kumar Choubey and Sanchita Paul, "GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis", *International Journal of Intelligent Systems and Applications*, Vol. 1, pp. 49-59, 2016.

[20] Ahmad Taher Azar and Aboul Ella Hassanien "Dimensionality Reduction of Medical Big Data using Neural-Fuzzy Classifier", *Soft Computing*, Vol. 19, No. 4, pp. 1115-1127, 2015.

[21] Ahmed Hamza Osman Hari "Diabetes Disease Diagonosis method based on Feature Extraction using KSVM", *International Journal of Advanced Computer Science and Application*, Vol. 8, No. 1, pp. 236-244, 2017.

[22] M. Sudha, "Evolutionary and Neural Computing Based Decision Support system for Disease Diagnosis from Clinical Data set in Medical Practice", *Journal of Medical Systems*, Vol. 41, No. 11, pp. 178-183, 2017.