Ovarian Cancer Disease Prediction and Categorization Its Level Using Hybrid Classification Approach

Guhan Thangavelu[1*] and Selvarajan S.[2]
1. Department of Computer Science and Engineering, Anna University, INDIA
2. Muthayammal Engineering College, Rasipuram, Namakkal, INDIA
***justforutsk@gmail.com**

Abstract
In recent days, ovarian cancer has been spread widely, which is required to identify the disease as soon as possible to eliminate the loss of human. More research works has been done by earlier researchers to identify this disease. Nevertheless, those work lag in identifying the disease at the early stage, but this issue is resolved in research work of Ovary Cancer Detection using Hybridized Bacterial Foraging with Particle Swarm and Multi Kernel SVM Approach (HBFMPSO-MKSVM). This will perfectly classify the levels of ovarian disease. However, this process is direct to have the more number of complexity and it makes the accurate result in the example of the presence of the large amount of gene expression data with more number of values and these values are missed. Therefore, in the proposed research work this trouble can be overcome which brought-in the new methodology is Hybrid Glow Worm Swarm Optimization with Fish Swarm Optimization and Fuzzy Support Vector Machine (HGSO-FSO-FSVM) methodology. This work will cluster the same gene together at the early stage by utilizing the Modified K means clustering algorithm and the missing value replacement happened by utilizing NLLS imputation method. Then feature reduction is happened by utilizing the Alpha Rough Set Theory (α-RST) approach. At last, the proposed research method HGSO-FSO-FSVM acquires optimal feature selection and classification. The proposed research methodology is executed and computed in the MATLAB simulation environment which distinguishes the earlier and the current wok to enhance the effectiveness.

Keywords: Optimal Feature Set Selection, disease Classification, Ovarian Cancer Detection, Similar Gene Grouping, Missing Value Replacement.

Introduction

Ovarian cancer is a cancerous growth which mainly affects the ovary. In the outer lining of the ovary (epithelium) this can be presented [1]. American cancer society shows that, this ovarian cancer is common between women and it is the most famous 8th cancer types and also in the United States too (except non-melanoma skin cancers) [2]. On the other hand, in women death it is the fifth general reason because of the cancer. Between the different gynaecologic cancers (uterine, cervical, and ovarian), the ovarian cancer is the most reason of the death. In the last year, most of the women are affected by the ovarian cancer in United States and among them 14,000 will most probably die. Truly, in the last five years the survival rate is 46% in the different countries. On the other hand, the National Cancer Institute, the ovarian cancer is identified in the earlier stage, earlier than the tumour spreads compared to the survival percent will be 94 [3].

Colonoscopy (bowel examination) has been done for some women to ensure that this symptom is not because of bowel issue [4]. In order to diagnose this, the doctor will insert a thin, flexible tube with a small camera and a light (endoscope) into the bowel. The result of this test and biopsies assist the doctor to define the level to which the cancer has been spread, which is known as staging. This process assists the team of health care to confirm which treatment is is required to cure this. If this is not understandable, the following stage will explain it through simple terms [5].

- **Stage 1:** Cancer is affected in one ovarian or both.
- **Stage 2:** In the pelvis the cancer is distributed to the other organs.
- **Stage 3:** The Cancer is distributed outside the pelvis to the lining of the abdomen, the bowel or lymph nodes in the abdomen or pelvis.
- **Stage 4:** The cancer has distributed further, to the bladder or rectum, **throughout** the abdomen or to other body parts.

Furthermore, to the initial stage, as like other cancers, ovarian cancer will also be graded. The histological grade of a tumour computes the abnormality or diseased of the appearance of the cells through microscope. The four grades point the possibility of the occurrence of the cancer to wide spread and the higher grade. Grade 0 I s utilized to explain the non-invasive tumours. Grade 0 cancers are also termed as borderline tumours [6].Grade 1 tumours have well comprehended cells (looks exact like normal tissues) and those cells will be with good prognosis. Grade 2 tumours are also termed as well-differentiated to some extent and they are made up of cells that simulate the normal tissue. Grade 3 tumours have the worst prognosis and their cells will be abnormal, which is termed as poorly discriminate [7].

Through this step identification of the ovarian cancer disease level is done. This is completed by bringing-in the novel approach like Hybrid Glow worm Swarm Optimization with Fish Swarm Optimization and Fuzzy Support Vector

Machine (HGSO-FSO-FSVM) method. This work will cluster the same gene together at the early stage by utilizing the Modified K means clustering algorithm and the missing value replacement happened by utilizing NLLS imputation method. Then feature reduction is happened by utilizing the Alpha Rough Set Theory (α-RST) approach. At last, optimal feature selection and classification is acquired with the help of the proposed research method HGSO-FSO-FSVM.

The proposed work is arranged as like: In this section, the view of the occurrence of the ovarian cancer and their symptoms were discussed. In section 2, different research work has been explained which is related to this work. The proposed research methodology overview is given in section 3 with the suitable samples. The simulation environment and the comparison between the preceding and the proposed work are explained in section 4. At last, the research work conclusion is explained in section 5 based on the simulation results.

Related Works

Here, various research work which has been chosen for the optimal features and classifying the data set to estimate the disease level. In [8] Lei Yu and Huan Liu proposed a novel method, which is referred as most important correlation, and he proposed a fast filter method which can detect the suitable features as well as duplication among the suitable features without any analysis of correlation pair-wise.

J. Shreve, H. Schneider, O. Soysal [9] proposed a methodology for distinguishing the classification methods via the evaluation of the resistance of the model and effectiveness in the variable selection. The systematic model is provided in this research for individual performance of six classification methodology by making use of the Monte Carlo simulation and describes the process of variable by differentiating the methodologies to confirm the minimal bias, enhanced resistance and improving the performance.

Sudha et al. [10] proposed the classification algorithm namely Naive Bayes, Decision tree and Neural Network for estimating the stroke diseases. The classification algorithm like decision trees, Bayesian classifier and back propagation neural network were chosen for this research work. The records with inappropriate data will be eliminated from the data warehouse before initiating the extracting process. Data mining classification technology has classification model and evaluation model.

Tinghua, Wanga, Houkuan Huang, Shengfeng Tian, Jianfeng Xu [11] proposing the SVMs feature selection process is completed through the optimization of kernel polarization with Gaussian ARD kernels. This work concentrates mainly on the choosing the features effectively for Support Vector Machine (SVM). From the conventional like quick and/or urination symptoms (very often). But these symptoms creates some other serious conditions also. If in searching methodologies, the feature selection is altered into the model selection of SVM.

Syed Umar Amin et al. [12] implemented the genetic neural network hybrid system, which creates the global optimization advantages of genetic algorithm for the neural network weights. A back propagation algorithm guides the networks to enhance the initialization of synaptic weights.

Ranjana Raut et al. [13] established and confirmed the dimensionally minimized MLP Neural Network method to be a fastest network. They noticed that that MLP NN is simple in design and synthesis, lowest average MSE, highest accuracy and ROC analysis is better. Experiments have been done with the Switzerland heart disease database on doing the comparison among the available and unavailable.

Venkatesan P. et al. and Rajan [14, 16] investigated and distinguished the effective and three famous classification algorithms such as namely C4.5, ID3 and CART to classify Tuberculosis dataset. These are highly helpful in medical research work to build the algorithms for predicting and classifying the disease. The notification result provides that C4.5 and CART performs better with respect to accuracy.

L. Shen And E.C.Tan [15] proposed the penalized logistic regression for cancer classification. The penalized logistic regression joined with the two-dimension reduction methods through this way; the accuracy classification and computational speed were enhanced. This process is known as Recursive Feature Elimination (RFE) which is utilizing the repetitive gene selection, which attempts to select the gene subset and it is suitable for this cancer [17]. Based on this performance is computed, seven data sets are selected such as breast cancer, central nervous system, colon tumour, Acute Leukaemia, Lung cancer, ovarian cancer and Prostate cancer. Linear SVM is utilized to distinguish the regression methods. Two software packages one was MATLAB by Schwaighofer and the other one is by Gunn [18], were utilized for executing SVM. The perfect performance will be accomplished by combining the Penalized logistic regression and PLS.

Predicting Early Stages of Ovarian Cancer

Mostly 20% of ovarian cancers were identified at the initial stage. If in case it is identified at the earlier stage, 94% of the patients can survive more than 5 years after analysis. Many studies were still in progress to study the best solution to identify the ovarian cancer as soon as possible. At the initial stage this cancer won't show any symptoms. When it causes symptoms, this will be caused by other things. The symptoms include abdominal swelling or bloating (because of a mass or a build-up of fluid), pelvic pressure or abdominal pain, complexity in eating or feeling

case these symptoms were caused by ovarian cancer, it will create sever problems, but this case isn't true always.

Gathering DNA spots which is attached with the solid surface which creates a Micro array gene expression data sets. From the entire spot the signal is collected in the micro array experiments and it is used to estimate the expression level of a gene. There are more than thousands of DNA is spotted in the micro array; each gene is wrapped-up in a genome. These data were utilized, to oversee the level of the expression for different genomes, which in turn differs from the existence of the disease. To estimate the occurrence of the disease, gathered Micro array gene expression data were examined. Nevertheless, controlling and identifying the different gene expression is a complex task which can't be done perfectly. The perfect technique is required to brought-in for the effective identification of the existence of the cancer disease.

The intention of this work is to give a knowledge based system to separate the ovarian cancer and normal case and this occurs by bringing-in the novel approach such as Hybrid Glow worm Swarm Optimization with Fish Swarm Optimization and Fuzzy Support Vector Machine (HGSO-FSO-FSVM) method. This work will cluster the same gene together at the early stage by utilizing the Modified K means clustering algorithm and the missing value replacement happened by utilizing NLLS imputation method. Then feature reduction is happened by utilizing the Alpha Rough Set Theory (α-RST) approach. At last, optimal feature selection and classification is acquired with the help of the proposed research method HGSO-FSO-FSVM. To identify the ovarian cancer following steps were followed:

- Modified K-Means clustering algorithm for Preprocessing the input dataset
- Missing value Handling and Normalization using NLLS imputation methods
- Feature Selection of the dataset using Alpha rough set theory (α-RST) approach
- Optimal feature selection using Hybrid Glowworm Swarm Optimization with Fish Swarm Optimization (HGSO-FSO) method
- Classification using fuzzy based SVM

The overall framework of the research work is given in the following figure 1.

Ovarian Cancer Dataset

The goal of these experiments is to recognize the proteomic patterns in the serum and it can be discriminate the ovarian cancer from other. This research is very important for the women who have been affected by this disease. The mass spectroscopy is creating the proteomic spectra and the given data set is 6-19-02, which includes 91 controls (Normal) and 162 ovarian cancers [20]. In the entire sample the raw spectral data has the amplitude of the intensity at the entire molecular mass/ charge (M/Z) identity. There are 15154 M/Z identities are available in this research. Based on the equation the intensity values are normalized: NV= (V-Min)/(Max-Min). In this case the normalized value is represented as NV, the raw value is represented as V, minimum intensity and the maximum intensity is represented as Min and Max, respectively. By using the 253 samples the normalization process is done for all the 15154 M/Z identities. The entire intensity value is reduced within the region of 0 to 1 after the normalization process.

Preprocessing using modified K means algorithm

The appropriate features are included in both the pre-processing and feature selection, and do the appropriate pre-processing and drawing-out features on the data items to estimate the values for the feature set. This process is very suitable to choose the subset of the entire features, to decrease the dimensionality of the trouble space. This step is frequently request the better deal of domain and the data analysis. The clustering trouble is solved by K-Means algorithm.

The algorithm is divided into objects to pre-defined clusters, which is given by the user. The random cluster centres is chosen for the entire cluster. These centres were chosen to be attainable from one another. The starting point controls the clustering process and outputs. Here the Centroid initialization works desirably in defining the cluster assignment in an efficient way.

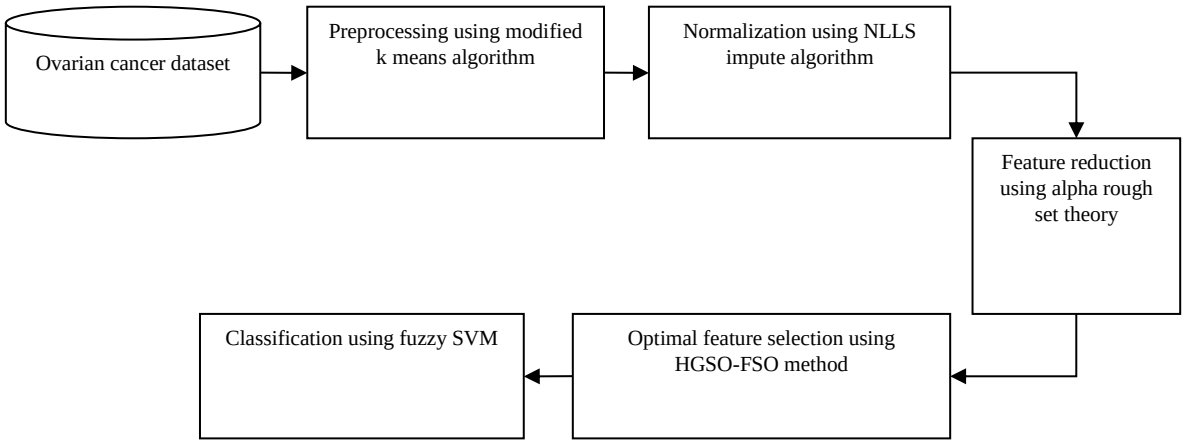

Figure 1: Ovarian Cancer Detection Approach

The convergence nature of clustering depends on the allotted values of the initial centroid. This work concentrates in the allotment of the cluster centroid selection so as to enhance the performance of the clustering process with the help of the K-Means clustering algorithm. This research utilizes the Initial Cluster Centres Derived from Data Partitioning with the Data Axis with the greatest variance to assign the values from the cluster centroids.

In this research work the enhanced clustering methodologies are explained, both the phases of the real k-means algorithm were altered to optimize the accuracy and efficiency. The enhanced methodology is described as follows.

1. **Phase 1:** Verify the primary centroids of the clusters with the help of algorithm 1.
2. **Phase 2:** Allot every data point to the appropriate clusters by utilizing Algorithm 2.

The centroids are defined in sequence order in the first phase so as to generate the clusters with good accuracy. The data points were alloted to the proper clusters in the second phase. The primary clusters are created according to the distance of the entire data-point from the initial centroids. These clusters were consistently fine-tuned with a help of the heuristic approach, thereby enhancing the efficiency.

Algorithm 1: Finding the initial centroids
Input: D = {d1, d2,.....,dn} // set of n data items
k // Number of desired clusters
Output: A set of k initial centroids.
Steps:
1. Set m = 1;
2. Calculate the distance among the entire data point and other data- points in the set D;
3. Identify the closest pair of data points from the set D and form a data-point set Am (1<= m <= k) which contains these two data- points, Remove these two data points from the set D;
4. Identify the data point in D that is nearer to the data point set Am, Add it to Am and remove it from D;
5. Continue step 4 until the number of data points in Am reaches 0.75*(n/k);
6. If m<k, then m = m+1, identify another pair of data points from D among which the distance is very small, form another data-point set Am and remove them from D, Go to step 4;
7. for each data-point set Am (1<=m<=k) identify the arithmetic mean of the vectors of data points in Am, these means will be the primary centroids.

Algorithm 1 explains an approach for identifying the clusters' initial centroids. Fundamentally, required to calculate the distance between the entire data point and the remaining data points in the data points set. After that the closet pair of the data points are identified and make the set A1 which has these two data points, and remove these points from the data point set D. after that the data points is defined and it is nearer to the set A1, and these value is included in A1 and remove them from the D point. This algorithm is iterated till the elements in the set A1 obtains the threshold. Here this process is again going to the second step and create the data-point set A2. Iterate this process till acquiring the k data points was obtained. Finally, the primary centroids are used to describe the nearest data point to the cluster centroids. The distance between the one vector X=(x1, x2… xn) and another vector Y=(y1,y2,…yn) is acquired as:

$$d(X,Y) = \sqrt{(x1-y1)^2 + (x2-y2)^2 + \cdots \ldots + (xn-yn)^2}$$

The distance among the data point X and the set D is described as:

$$d(X,D) = \min(d(X,Y), \quad \text{where } Y \epsilon D)$$

The primary centroids of the cluster were given as input to the second phase, for accrediting the data-points to the proper clusters. This phase's steps were outlined as Algorithm 2.

Algorithm 2: Assigning data-points to clusters
Input: D = {d1, d2,.....,dn} // set of n data-points.
C = {c1, c2,.....,ck} // set of k centroids
Output:
A set of k clusters
Steps:
1. Calculate the distance of the entire data-point di (1<=i<=n) to the every centroids cj (1<=j<=k) as d(di, cj);
2. **For** each data-point di, identify the closest centroid cj and assign di to cluster j.
3. Assign ClusterId[i]=j; // j:Id of the closest cluster
4. Assign Nearest_Dist[i]= d(di, cj);
5. For every cluster j (1<=j<=k), recalculate the centroids;
6. **Repeat**
7. **For** every data-point di,
7.1. Calculate its distance from the centroid of the present nearest cluster;
7.2. **If** this distance is less than or equal to the present nearest distance, the data-point available in the cluster;
Else
7.2.1 For every centroid cj (1<=j<=k)
Compute the distance d(di, cj);
Endfor;
7.2.2 Allocate the data-point di to the cluster with the nearest centroid cj
7.2.3 Assign ClusterId[i]=j;
7.2.4 Assign Nearest_Dist[i]= d(di, cj);
Endfor;
8. For each cluster j (1<=j<=k), recalculate the centroids;
Until the convergence criteria is met.

In Phase 2 the initial step to compute the distance between the entire data-point and the initial centroids of the entire clusters. After that the data points are assigned to the clusters and it has the nearest centroids. At first this result collects

the data points. For the entire data-point, the cluster is indicated as Cluster Id and the distance from the centroids of the adjacent cluster as Nearest_Dist. In various types of cluster the data points are direct to modify the values of the cluster centroids. For the entire cluster, the centroids are recalculated by considering the mean values of its data-points. In this process, the algorithm is alike the real k-means procedure but the primary centroids will be estimated analytically.

Missing value replacement using NLLS impute algorithm

The micro array data the gene expressions were utilized for pre-processing the large number of genes to select the suitable genes, which is utilized to increase the capability of the classification and the accuracy rate from the high dimensional data set. For different conditional reasons the microarray data has a missing value. Moreover, the classification, the network design and clustering gene expression data analysis algorithms label a matrix of gene array for the duration of analysis. The missing values are required earlier than the micro array data in data analysis procedure because it is significant. The Local Least Square imputation based algorithm (NLLS impute) is executed in this proposed work. Both the local least square imputation based algorithm and the k nearest neighbour imputation algorithm is same. The k-nearest gene algorithm is chosen by this proposal work from the entire gene as a substitute for the entire genes and also to utilize the imputed information again.

Through this method, huge amount of details which are available for predicting the missing values in the target genes. In this sequence this higher missing rate will be imputed in the target gene with the presented detail. This means the imputed information acquired from the genes, whose missing values have been supposed before.

The process of New Local Least Square impute algorithm is listed as follows:

Algorithm: NLLS impute
Input: From the micro array gene expression matrix T the input is considered as m and n. the number of genes are represented as m and the number of samples is represented as n. in this matrix the artificial missing entries are also available.

Output: The missing entries are calculated.

1. In the initial stage from the entire gene expression matrix X to change the missing positions of the provided gene expression matrix.
2. Based on the missing rate and imputation one after another the m number of genes in X is sorted in ascending order.
3. For every target gene gt in X do:
   a) For every missing position p in target gene gt do:
      i. The Pearson Correlation coefficient is used to choose the k number of candidate genes adjacent to the target gene. For the period of the computation of Pearson Correlation coefficient the missing positions filled with row average were skipped [9].
      ii. ii) The commencement of local least squares is imputing the missing position p in the target gene under deliberation [9].
      iii. iii) The output of imputed values is located in the missing positions of the considered target gene.
4. End

Feature reduction using alpha rough set theory approach

Once after gathering the genes of same patterns, in-appropriate gene removal is done. This will minimize the computation overhead and accuracy degradation which might happen because of processing highly inappropriate gene data which is available in the micro array gene expression data set. In this research work Alpha Rough Set Theory is adjusted to identify the appropriate genes which is belongs to the issues of the specification. This is done through estimation of indiscernibility relation value. The earlier indiscernibility determination isn't enough while conceiving the generalized information system since it doesn't consider the likelihood degrees which are correlate with the values of attributes in R. To rectify this issues, the parameterized relation is introduced, which is represented as IND.R; a. or, in short, Ra. Then consider the two elements are inappreciable if and only if they have the similar values for the entire attribute and if their likelihood level is higher than the given similarity threshold, which is represented as a:

$$\forall x, y \in U \; xR_{\alpha}y \Leftrightarrow x \,\widehat{R} y \text{ and } \mu_{\bar{s}}(\pi x, \pi y$$

The relation R α is determined on both equivalence relation $\widehat{R}$ and a similarity relation$\bar{s}$. The relation $\widehat{R}$ is similar to the classical indiscernibility relation R, description depends on the same attribute values. The approximation definition does not hold while the generalized information systems, certainly $[\hat{x}]_{R\alpha}$and X are both fuzzy sets. This is extended to calculate the approximations of fuzzy sets which are also fuzzy. Consider X is a fuzzy subset of U and x ∈ U. The object x belongs to the smallest approximation of X iff $[\hat{x}]_{R\alpha}$is added in, which entails that belongs to X ∩ $[\hat{x}]_{R\alpha}$. Consequently, the level to which x owes to the smallest approximation which is adequate to the decreased degree α1, which it owes to X and degree α2 to which it owes to $[\hat{x}]_{R\alpha}$: Furthermore,if x belongs to the upper approximation, then x owes to $[\hat{x}]_{R\alpha}$or to X, thus it belongs to the union of X and $[\hat{x}]_{R\alpha}$. In this case the degree of x is equal to the higher α 1 and α 2. As a consequence, the generalized approximations were determined as follows:

$$R_\alpha X = \{(x, \Delta_x) \in U \times [0,1] | [\hat{x}]_{R\alpha} \subseteq X \text{ and } \Delta_x = \min(\mu_{|\hat{x}|_{R\alpha}}(\hat{x}), \mu_{(\hat{x})})\}$$

For instance, the idea is not clear. It means the concept without a sharp boundary, for the reason that, there are a small number of sits for which it can't be completed either they are beautiful sight or not. This complex set of methodology is ruled by a logic which allows an unclear idea of two values which are true or false. In Alpha rough set, the restriction of the fuzzy set is also a fuzzy set and an instance has the level of membership in a boundary set A with a boundary isn't only rough but it is also fuzzy.

Optimal feature selection using HGSO-FSO method

The earlier section utilizes the alpha rough set theory to avoid the inappropriate genes which exist in the database and the output is obtained with appropriate genes which denote the ovarian cancer disease. Nevertheless, managing these genes leads to computation overhead it will be eliminated by choosing the optimal genes between them which represents response to the ovarian cancer. This is done with the help of the HGSO-FSO. The proposed technique is utilized to improve the quality of global optima of multimodal functions.

In the normal GSO algorithm, every glow worm only in concurrent with Lucifer in values of glow worms in its neighbour set, choose the glow worm by a specific likelihood and navigate it near. Nevertheless, if the search space of an issue is huge or irregular, the neighbour sets of some glow worms may be empty, which tends to maintain these glow worms in repetitive process. To eliminate this case and to assure that every glow worm is in motion. We need to bring-in the predatory behaviour of FSA into GSO and propose a Hybrid GS with FSO (HGSO-FSO) algorithm. The concept of IGSO is as follows: the glow worms whose neighbour sets are empty and it is preceded for predatory behaviour in their dynamic decision domains. Consider that N denotes population size, $x_i(t) = \left[x_i^{(1)}(t), x_i^{(2)}(t), \ldots\ldots, x_i^{(d)}(t)\right]$represents the position of the ith glow worm at the t-th iteration. The procedure of HGSO-FSO can be described as follows:

Step 1: let $l_i(0) = l_0$, $r_d^i(t) = r0$, t = 0; here, t represents the GSO iteration' counts. Randomly initialize the position $x_i(t)(i = 1, 2, \ldots, N)$ of every glow worm in the search space. Estimate the fitness value $f(x_i)$ of each glow worm. Initialize the current optimal position x^* and the existing optimal value f_x^*based on the fitness values.

Step 2: Update the Lucifer in value $l_i(t)$ of every glow worm based on the following equation.

$$l_i(t+1) = (1-\rho)l_i(t) + \gamma f(x_i(t+1))$$

Step 3: Compute Ni(t) and Pij(t) for each glow worm based on the following equation.

Step 4: For every glow worm, if $N_i(t)$ isn't empty, then based on $P_{ij}(t)$ and roulette method, choose the j-th glow worm in Ni(t) and proceed toward it, compute $x_i(t+1)$ according to (8), Or else, execute predatory behaviour in $r_d^i(t)$ and get $x_i(t + 1)$. If $x_i^j(t + 1) < a_j$, then $x_i^j(t +1) = a_j$; If $x_i^j(t + 1) > b_j$, then $x_i^j(t + 1) = b_j$, where j = 1,2, …, d.

Step 5: Compute the current fitness value $f(x_i(t))$ of every glow worm, if the enhanced position and enhanced value of the current population are better than x^* and f_x^*, then update x^* and f_x^*, or else, don't update.

Step 6: If the maximum iteration's count is fulfilled, then terminate and output x^* andf_x^*; or else, compute $r_d^i(t+1)$ according to (9) and let t = t + 1, return Step 2.

Classification using fuzzy based SVM

Here, classification of gene data set is completed to estimate whether the ovarian cancer is available or not. This is finished with the help of the Fuzzy SVM approach. In this research work, Multiclass SVM is utilized for estimation, where this decision is made by utilizing the fuzzy decision making system [21]. Thus SVM classification is completed according to the fuzzy decision making process. The new model differs from the conventional SVM classifiers, because it considers the fuzzy theory. And the fuzzy decision making function is also created to change the sign function in the prediction level of classification process. In the prediction segment, the methodology is introduced to make the decision value as the independent variable of fuzzy decision making function to categorize the test data set into the different classes. The decision making model is very suitable to the properties of real-world conditions, where the interactions and noise influence available in and around the restriction of the various types of clusters.

Fuzzy Decision-making SVM processing is a theory which depends on examining the database, SVM algorithm and a mass of experiments. It is lengthening the conventional decision-making theory which is utilized to conceive the dataset as a crisp one. But as addressed earlier, in various real world applications, every input may not be classified perfectly and at the time of SVM classifier entire input points won't be divided exactly. Alternatively, few characteristic input points were simple to be divided but others have some complexity. Need to compute the degree of every input point, whether they can be divided exactly or it is misclassified. The decision value is extracted from the traditional decision-making function of SVM, which is a sign function. Due to these correlated values, various input points refers to the adjacent grades among these points and the optimal separate hyper plane, they can be utilized as an individual variable in fuzzy decision-making function which is proposed in the work.

Modelling the basic structure of fuzzy decision-making SVM processing segregates it into three main stages: SVM straining, decision value prediction, and at last, fuzzy decision-making processing. In the fuzzy decision making processing segment, the fuzzy model is created to re-construct the decision making function of SVM by confirming the clusters as fuzzy sets.

Fuzzy decision making model

In traditional methods, sign function is usually utilized to separate the test detains to various classes, so zero is considered to be one significant value as a threshold. But in various real conditions, the relationship among the classes isn't easy. In both classes, specifically around the limitations among them, communication occurs generally. Closer to other classes, the points were distributed by the noise. Because of the effect from every other, the assessment one particular decision value will be created as an interpretable and here, zero isn't significant. Various classification methods weren't sufficient and few training data sets aren't usually utilized for estimation, various misclassified cases happen in the neighbour threshold. To alter this incapable division, we construct a fuzzy limitation among two fuzzy sets with binary labels. From the previous stage, we can obtain the decision value of every input vector, which is distributed on both sides of zero.

Consider that these two fuzzy sets which is known as A (the value in this set are regarded to be estimated as -1) and B (the values in this set are regarded to be estimated as +1), and they utilize decision value as an individual variable, through this a fuzzy decision model will be built. In fuzzy theory, there is no apt limitation among set A and set B but a gray-zone will alternate it. Decision values in this zone weren't separated to set A or set B certainly. In our method, boundary functions were constructed by utilizing these decision values as an individual variable, and the values of the function represents their dependability which manages the concept of belief issues (belief degree among 1 (completely believable) and 0 (completely false)). To face the real-world cases very well, a suitable model is required. If the decision value is represented by v, and converting the fuzzy decision value to the range from 0 to 1 exactly, the limitations of set A and set B is determined as follows:

$$f_A(v) = \frac{\arctan(-v.s - d.s)}{\pi} + 0.5$$

$$f_B(v) = \frac{\arctan(-v.s - d.s)}{\pi} + 0.5$$

The above procedure will be utilized to estimate the existence of ovarian cancer disease presence accurately and early. In this research work the computation process is performed with the help of the MATLAB simulation environment which is described in the following subsections.

Experimental Results

In these sections the numerical calculations are described, with the terms of various performance attributes to investigate the improvement of the performance of the proposed and the existing methodologies. MATLAB simulation environment is utilized to execute the proposed research methodology. The performances measures are conceived in this work were given as follows: Accuracy, Sensitivity, Specificity, Precision, Recall, F-Measure and G-Mean".

The above measures were estimated for both the proposed methodology such as HGSO-FSO-FVM, HBFMPSO-MKSVM and the current methodologies such as SVM-REF, ANN which is explained graphically from figure 2 to 8. These plots exactly indicate that the HGSO-FSO-FSVM is the fastest convergence at the time of training which is followed by the HBFMPSO-MKSVM, SVM-REF and ANN based models.

The proposed research method chooses 35 best M/Z identities from ovarian sample dataset which is conceived in this work. The M/Z identities that were chosen to describe the following table 1.

Accuracy: The accuracy is determined as the exact estimation point of the stock index value from the differing stock index which varies at times. The accuracy of the proposed system should be more than the other current methodologies like CSO, ABFO and BFO. The accuracy value is computed in terms of the stock index value prediction system's true positive, false positive, true negative and false negative values. The accuracy is estimated as like:

$$\text{Accuracy} = \frac{T_p}{(T_p + F_p + F_n)}$$

Table 1
M/Z identities selected by proposed work

MZ7929.999	MZ1270.2829	MZ15578.142	MZ9256.079	MZ2.4304278
MZ16.332996	MZ1053.5652	MZ7325.3416	MZ38.753108	MZ7258.4005
MZ548.72469	MZ2.8234234	MZ681.38131	MZ545.23249	MZ12246.622
MZ200.70672	MZ2935.9679	MZ1057.8103	MZ1209.8184	MZ7911.7263
MZ5809.2516	MZ3401.0871	MZ2460.3929	MZ7933.3236	MZ635.41141
MZ9828.5219	MZ793.83523	MZ8707.7758	MZ407.11886	MZ197.28389
MZ4996.2722	MZ636.82367	MZ126.89376	MZ2045.3401	MZ1215.0177

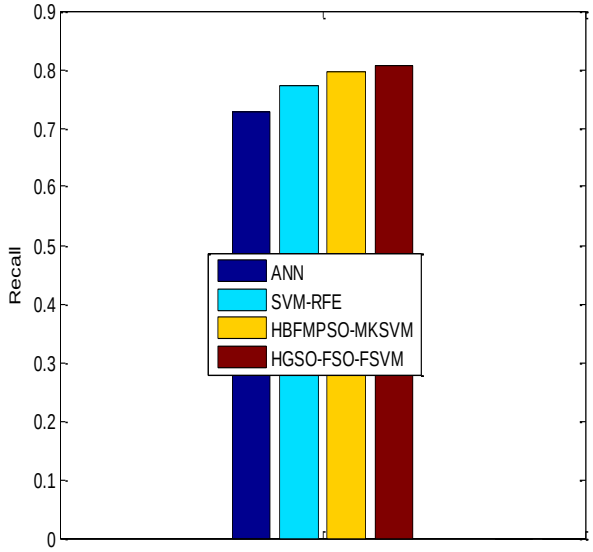

Figure 6: Recall comparison

The figure 6 explains the recall metric graphically. It shows that the suggested research method is better than the current research methods. HGSO-FSO-FSVM is 3% better than HBFMPSO-MKSVM, 4% better than SVM-REF and 8% better than ANN.

F-Measure: F measure is a group of the precision and recall and is the harmonic mean of precision and recall, the conventional F-measure or equalized F-score:

$$\text{F-Measure} = 2.\frac{\text{Precision.recall}}{\text{Precision+Recall}}$$

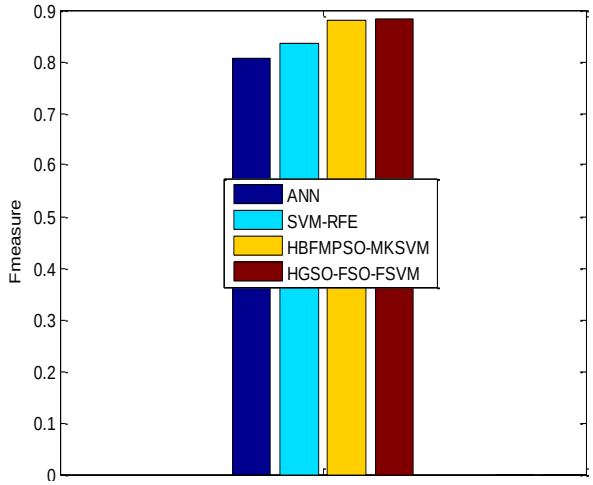

Figure 7: F-measure comparison

The figure 7 explains the F-Measure metric graphically. The graph shows that the suggested research method is better than the current research methods. HGSO-FSO-FSVM is 1% better than HBFMPSO-MKSVM, 6% better than SVM-REF and 4% better than ANN.

Gmean: The geometric mean is an average which is useful for set of positive numbers which is translated according to the product and not on their sum (as is the case with the arithmetic mean) e.g. rates of growth.

$$\bar{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

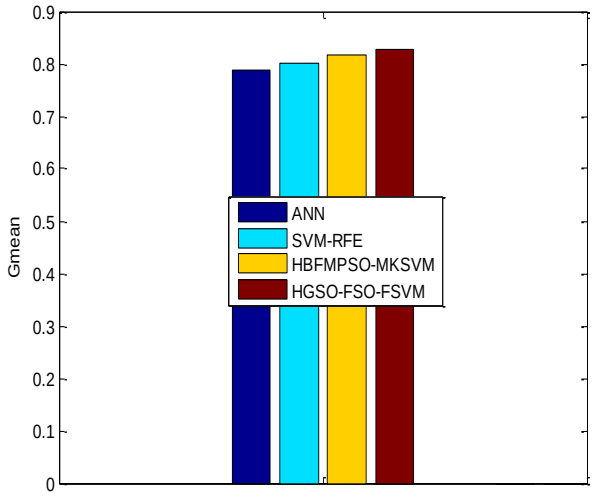

Figure 8: G-Mean comparison

The figure 8 explains the F-Measure metric graphically. It shows that the suggested research method is better than the current research methods. HGSO-FSO-FSVM is 2% better than HBFMPSO-MKSVM, 1% better than SVM-REF and 3% better than ANN.

Conclusion

Ovarian cancer identification is the very common consideration for research in the medical which affects mostly for women. Identification of this disease as soon as possible will avoid the effect on the health of the women. In this research work, Hybrid Glow worm Swarm Optimization with Fish Swarm Optimization and Fuzzy Support Vector Machine (HGSO-FSO-FSVM) method is proposed for the better identification of the ovarian cancer disease. This proposed research method basically clusters the similar genes collected with the help of the Modified K means clustering algorithm and the missing value replacement is done by NLLS imputation method. Then feature reduction is completed by utilizing the Alpha Rough Set Theory (α-RST) approach. At last, optimal feature selection and classification is done by using he suggested research method HGSO-FSO-FSVM. The proposed research methodology is executed and computed in the MATLAB simulation environment which distinguishes the earlier and current research work to confirm the efficiency.

References

1. Marcus, C. S., Maxwell, G. L., Darcy, K. M., Hamilton, C. A., & McGuire, W. P. (2014). Current approaches and challenges in managing and monitoring treatment response in ovarian cancer. Journal of Cancer, 5(1), 25-30.

2. Smith, R. A., Manassaram Baptiste, D., Brooks, D., Doroshenk, M., Fedewa, S., Saslow, D., & Wender, R. (2015). Cancer screening in the United States, 2015: a review of current American cancer society guidelines and current issues in cancer screening. CA: a cancer journal for clinicians, 65(1), 30-54.

3. Wender, R., Fontham, E. T., Barrera, E., Colditz, G. A., Church, T. R., Ettinger, D. S., & LaMonte, S. J. (2013). American Cancer Society lung cancer screening guidelines. CA: a cancer journal for clinicians, 63(2), 106-117.

4. Rex, D. K., Schoenfeld, P. S., Cohen, J., Pike, I. M., Adler, D. G., Fennerty, M. B., & Shaheen, N. J. (2015). Quality indicators for colonoscopy. Gastrointestinal endoscopy, 81(1), 31-53.

5. Riester, M., Wei, W., Waldron, L., Culhane, A. C., Trippa, L., Oliva, E., & Birrer, M. J. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. JNCI: Journal of the National Cancer Institute, 106(5).

6. Mutch, D. G., & Prat, J. (2014). FIGO staging for ovarian, fallopian tube and peritoneal cancer. Gynecologic oncology, 133(3), 401-404.

7. Jayson, G. C., Kohn, E. C., Kitchener, H. C., & Ledermann, J. A. (2014). Ovarian cancer. The Lancet, 384(9951), 1376-1388.

8. Li, W., Han, J., &Pei J., CMAR: accurate and efficient classification based on multiple association rules. Proceedings of 2001 International Conference on Data Mining, 2001, pp. 369-376.

9. Shreve, J., Schneider, H., &Soysal, O. (2011). A methodology for comparing classification methods through the assessment of model stability and validity in variable selection. Decision Support Systems, 52(1), 247–257.

10. Sudha, A., Gayathiri, P., &Jaisankar, N. (2012). Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Applications, 43(14), 26–31.

11. Wanga, T, Huang, H., Tian, S., &Xu, J. (2010). Feature selection for SVM. via optimization of kernel polarization with Gaussian, ARD kernels. Expert Systems with Applications, 37(9), 6663–6668.

12. Amin, S. U., Agarwal, K., Beg, R. (2013). Genetic neural network based data mining in prediction of heart disease using risk factor. Proceeding of IEEE Conference on Information and Communication Technologies (ICT), pp. 1227– 1231.

13. Raut, R., Dudul, S. V. (2009). Maximum heart rate resting blood pressure scatter plot for the prominent features abnormal normal design and performance analysis of MLP NN based binary classifier for heart diseases. Indian Journal of Science and Technology, 2(8), 43–48.

14. Venkatesan, P., Yamuna, N. R. (2013). Treatment response classification in randomized clinical trials: a decision tree approach. Indian Journal of Science and Technology, 6(1), 3912–3917.

15. Shen, L., & Tan, E.C. (2005). Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data. IEEE/ACM Trans. Computational Biology and Bioinformatics, 2(2), 166-175.

16. Rajan, C, Geetha, K, (2017),"Heuristic Classifier for Observe Accuracy of Cancer Polyp Using Video Capsule Endoscopy", Asian Pacific Journal of Cancer Prevention, vol. 18, no.6, pp. 1681-1688.

17. Li, J., & Liu, H., Kent Ridge Biomedical Data Set Repository, http://sdmc-lit.org.sg/GEDatasets, 2002.

18. Gunn, S., SVM MATLAB Toolbox. http:// www .isis. ecs. soto n.ac.uk/resources/svminfo/, 2001.

19. Rajan, C, Geetha, K, (2016), "Automatic Colorectal Polyp Detection in Colonoscopy Video Frames", Asian Pacific Journal of Cancer Prevention, vol. 17, no 11, pp. 4869-4873.

20. Sorace, J. M., & Zhan, M. (2003). A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC bioinformatics, 4(1), 24,1-13.

21. Mardani, A., Jusoh, A., & Zavadskas, E. K. (2015). Fuzzy multiple criteria decision-making techniques and applications–Two decades review from 1994 to 2014. Expert Systems with Applications, 42(8), 4126-4148.