IJCTA, 9(36), 2016, pp. 167-186

Improved Minimum Redundancy Maximum Relevance with Hybrid Swarm Intelligence Based Gene Selection Algorithm for Cancer Classification of Microarray Gene Expression Data

N. Kanchana*

Abstract : In recent times, microarray based gene expression profiling has turned out to be most vital and promising dataset for the purpose of cancer classification that are used for effective diagnosis and prognosis. It is extremely vital to determine the most informative and defective genes in order to improve premature cancer diagnosis and to provide effective chemotherapy processes. In addition, in order to find perfect gene selection methods that considerably reduce the dimensionality and choose informative genes is extremely noteworthy issue in the field of cancer classification. Here, in this work, at first preprocessing process is done with the assistance of Probabilistic Principle Component Analysis (PPCA) in order to discover the Mutual Information detection on Micro array dataset and to effectively diminish the noise included in the dataset. Then, by using the preprocessed dataset an improved minimum Redundancy Maximum Relevance with Glowworm Swarm Optimization (ImRMR-GSO) algorithm is proposed for the purpose of selecting the predictive genes from the cancer microarray gene expression profile. Subsequently, these genes are classified with the assistance of a hybrid classification method which utilizes Random forest, SVM classifier and boosting ensemble learning method. Experimental results demonstrate that the proposed ImRMR-GSO algorithm achieves most accurate classification performance with small amount of predictive genes when tested using both datasets and compared against previously suggested schemes. This proves that the proposed ImRMR-GSO is a promising approach for effectively solving gene selection and cancer classification problems.

Keywords : Cancer classification, microarray, gene selection, gene expression, swarm optimization.

1. INTRODUCTION

Gene expression profiles that are acquired from specific microarray experiments have been extensively utilized for the purpose of cancer classification to construct an effective scheme. This scheme can effectively distinguish normal or different cancerous states with the assistance of selected informative genes [1]. On the other hand, studying microarray dataset in relation to their gene expression profiles poses a challenging process. The complexity of the problem increases from the enormous amount of features that contribute to a profile as compared against the extremely low number of samples normally existing in microarray analysis. An additional challenge is the existence of noise (biological or technical) in the dataset, which additionally disturbs the accuracy of the experimental results.

* Assistant Professor, School of IT & Science, Dr.G.R. Damodaran College of Science, Coimbatore, India.

Microarrays, recognized as DNA chips or some time regarded as gene chips, are chips that are hybridized to a labeled indefinite molecular extracted from a specific tissue of interest. Hence, it is possible to measure instantaneously the expression level in a cell or tissue sample for each gene represented on the chip [2][3]. DNA microarrays can be utilized for the purpose of determining which genes are being expressed in a particular cell category at a specific time and under precise conditions. This permits to relate the gene expression in two different cell categories or tissue samples, where it can effectively determine the additional informative genes that are accountable for causing a specific disease or cancer [4].

In recent times, microarray technologies have opened up several windows of opportunity to explore cancer diseases by means of gene expressions. The principal task of a microarray data investigation is to control a computational model from the particular microarray data that can predict the class of the particular unknown samples. The accuracy, quality, and robustness are extremely vital components of microarray analysis. The accuracy of microarray dataset analysis completely based on both the quality of the provided microarray data and the applied analysis scheme or objective. On the other hand, the curse of dimensionality, the insignificant number of samples, and the level of inappropriate and noise genes make the classification task of a test sample extremely challenging [5][6]. Those inappropriate genes not only introduce certain unnecessary noise to gene expression data analysis, however also increase the dimensionality of the gene expression matrix. This results in the growth of the computational complexity in several consequent research objectives like classification and clustering [7].

Bioinspired evolutionary schemes are more appropriate and precise than the wrapper gene selection scheme [8] since they have the capability for searching and discovering the optimal or near-optimal solutions on high-dimensional solution spaces. In addition, they permit searching the solution space by means of considering more than one attribute simultaneously [8]. However, as other evolutionary schemes, the ABC has certain challenging issues, particularly in computational efficiency, when it is processed on complex and high-dimensional data like microarray datasets. As a result, to effectively enhance the performance of the ABC algorithm in high-dimensional datasets, here proposed the idea of adding a feature selection algorithm, minimum Redundancy Maximum Relevance (mRMR), as a preprocessing stage. At this point, it is combined with the ABC algorithm, mRMR-ABC, with the intention of choosing informative genes from cancer microarray profiles. This hybrid gene selection provides a better balance between filters and wrapper gene selection schemes, being more computationally effective, as in filter schemes, and model feature dependencies as in wrapper schemes [8].

The existing mRMR is effectively utilized for Mutual Information is taken as the fundamental criterion to discover the feature relevance and redundancy. The mutual information among a feature and class labels states the consequence of that particular feature. Another time, the mutual information among different features states the correlation i.e., the redundancy among those specific features. Moreover, the ABC algorithm's drawbacks like poor response in case of local search ability and the process of ABC gets slow down when it is utilized in sequential processing. As a result, with the aim of overcoming these complications, Improved Minimum Redundancy Maximum Relevance (I- mRMR) combined with GSO approach is proposed for the purpose of reducing noisy and irrelevant genes. Furthermore, a hybrid classifier based on random forest, SVM and boosting schemes is used for the purpose of classifying the gene expression dataset. It provides higher values of accuracy, specificity, positive predictive value and negative predictive value. The major reason for more effective results in the case of hybrid classification methodology used in this paper is it effectively makes use of the advantages of each of the traditional SVM, RF classification schemes.

2. RELATED WORKS

Yifeng Li &Ngom (2012) [9] formulated a novel Kernel NMF (KNMF) scheme for the purpose of effective feature extraction and classification of microarray data. This scheme is also generalized to kernel High-Order NMF (HONMF). Broad experiments on eight microarray datasets demonstrate that this scheme generally outperforms the conventional NMF and existing KNMFs.

Lingyan Sheng et al (2009) [10] improved a Block Diagonal Linear Discriminant Analysis (BDLDA) and effectively applied to gene expression data. BDLDA is a kind of classification tool with embedded feature selection that has exhibited better performance on simulated data. On the other hand, with the use of cross validation in training, BDLDA is extremely time consuming, as a result, it is not an appropriate scheme for gene expression data, which has a huge number of features and comparatively small number of samples. In this scheme, estimated error rate is utilized as a measure to select the best model. The algorithm is optimized by continuously repeating the model construction procedure with previously selected features removed, which leads to increased classification robustness.

Herold et al (2008) [11] effectively compared the unsupervised and supervised gene selection schemes. Recent mechanism learning schemes completely depend on matrix disintegration schemes which are more resembling Independent Component Analysis (ICA) offer innovative and well-organized investigation tools which are explored for the purpose of evaluating gene expression outline. These tentative characteristic extraction schemes gave instructive expression modes which offered indication of fundamental regulatory techniques. The gene which exhibited the strong behaviour was taken for the purpose of classification of the tissue samples under inspection. In order to assess this result, it was compared against supervised gene selection schemes which completely depended on numerical scores or support vector. This scheme was used in macrophages loaded/de-loaded with chemically customized low density lipids.

Daliri & Mohammad Reza (2012) [12] formulated a scheme, which is completely based on combination of Genetic Algorithm (GA) for the purpose of feature selection and newly proposed scheme, namely the Extreme Learning Machines (ELM) for the purpose of classification of lung cancer data. The dimension of the feature space is effectively reduced with the assistance of GA and the effective features are chosen in this manner. The data are subsequently fed to a Fuzzy Inference System (FIS) which is trained with the help of the fuzzy extreme learning machine scheme.

Saraswathi et al (2011) [13] assessed the performance of ICGA-PSO-ELM and compared this scheme results with existing schemes in the literature. An investigation into the functions of the chosen genes, by means of a systems biology approach, revealed that most of the recognized genes are involved in cell signaling and proliferation. An examination of these gene sets shows a larger representation of genes that encode secreted proteins than found in arbitrarily chosen gene sets. Secreted proteins constitute a major means through which cells intermingle with their adjacent cells. Mounting biological evidence has recognized the tumor microenvironment as a serious factor that regulates tumor survival and development. As a result, the genes identified by this investigation that encode secreted proteins might provide significant insights to the nature of the critical biological characteristics in the microenvironment of each tumor type that permit these cells to increase and proliferate.

Subbulakshmi & Deepa (2015) [14] formulated a hybrid scheme in accordance with the machine learning paradigm. This paradigm integrated the effective exploration scheme called self-regulated learning capability of the Particle Swarm Optimization (PSO) algorithm with the ELM classifier. With the recent off-line learning scheme, ELM is a single-hidden layer Feed Forward Neural Network (FFNN), proved to be an effective classifier with huge amount of hidden layer neurons. In this scheme, PSO is effectively utilized to determine the optimum collection of parameters for the ELM, as a result reducing the amount of hidden layer neurons, and it further enhances the network generalization performance.

Rong et al (2009) [15] formulated an Online Sequential Fuzzy Extreme Learning Machine (OS-Fuzzy-ELM) for the purpose of function approximation and classification problems. The equivalence of a Takagi Sugeno Kang (TSK) Fuzzy Inference System (FIS) to a generalized single hidden-layer feed forward network is shown first, which is subsequently used to develop the OS-Fuzzy-ELM algorithm. This results in a FIS that can handle any bounded non constant piecewise continuous membership function. In addition, the learning in OS-Fuzzy-ELM can be done with the input data coming in a one-by-one mode or a chunk-by-chunk (a block of data) mode with fixed or varying chunk size. In case of OS-Fuzzy-ELM, the entire the antecedent parameters of membership functions are randomly assigned first,

and subsequently, the equivalent consequent parameters are determined in analytic manner. Performance comparisons of OS-Fuzzy-ELM with other existing schemes are presented using real-world benchmark problems in the areas of nonlinear system identification, regression, and classification.

Mohanasundaram and Periasamy [30] formulated a hybrid optimization approach for identifying the optimal data storage position in WSN. The work provided significant results through the hybrid genetic algorithm and particle swarm optimization approach.

3. PROPOSED METHODOLOGY

In this section, the novel ImRMR-GSO algorithm is proposed for selecting the predictive genes from the cancer microarray gene expression profile. The goal of this algorithm is the selection of the more informative gene for the purpose of improving the RFSVM classifier accuracy performance through the pre-selection of the relative and informative genes employing the ImRMR technique. Thereafter the estimation of the best predictive genes is done by using the GSO algorithm in the form of a wrapper gene selection strategy along with the RFSVM classifier. It is evident that, in this new ImRMR-GSO algorithm, the genes are selected to make a small dataset (ImRMR dataset) which comprises of the informative genes. As a result, the optimization process will be enhanced, and the comparison made with the actual GSO algorithm which did the selection of the genes for the initial microarray dataset directly. The system flow diagram is illustrated in the figure 1.

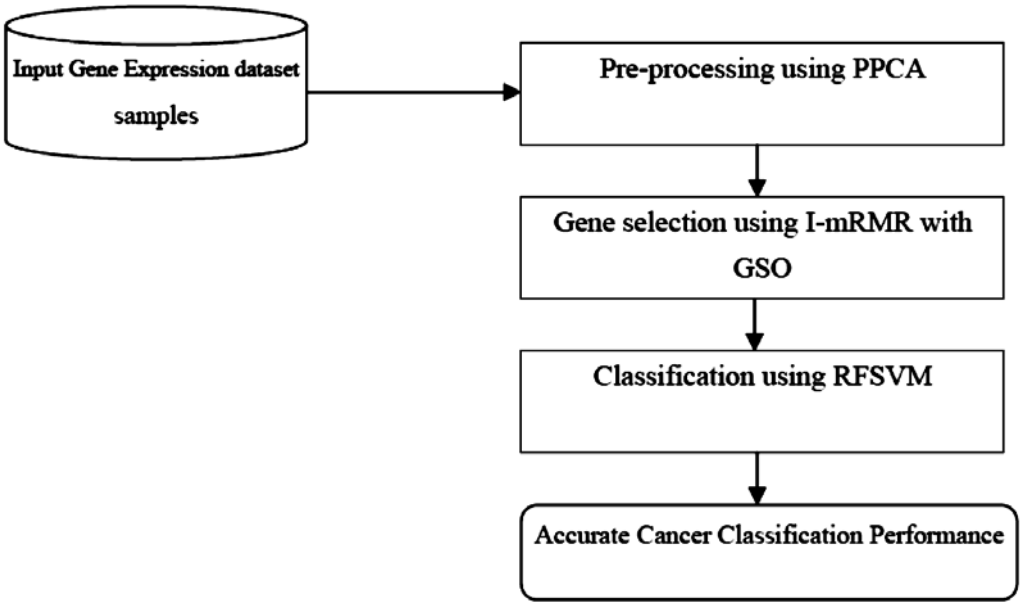

Figure 1: Architecture view representation of the contribution

3.1. Microarray Gene Expression Data

The recent progress in the microarray gene expression [21,22] data has rendered the measurement and analysis of the high dimensional gene expression data feasible. Also it improves the area of genetic research. Factually, microarray gene expression analysis has a significant role to play in the area of molecular classification of cancer, in the recent times. In the primitive level, it can be regarded as a sample against the gene two dimensional matrix along with an extra column that represents the respective classes of samples. Sometimes, the rows of the microarray data has the experimental conditions in place of samples. In almost all the cases, the unrefined data contains noise or missing values. Furthermore, the tremendous size of the dataset leads to an increase in the difficulty level for the researcher. Still again, few genes are not important to the respective class labels and even though they tend to make the data size larger. Hence, prior to the application of the microarray data, it has to beundergo preprocessing with some scheme.

3.2. Pre-processing using PPCA

PCA is only a replacement for an n-dimensional data space and the selection of m dimensions present in the turn or rotated space to be the new m dimensional linear subspace. With the condition that the data present in the actual space is Gaussian, then the data present in the rotated subspace also tends to be Gaussian. Hence, PPCA is a Gaussian modeler which describes the relation between the Gaussians in the distinct space and the subspace. The model is expressed as

$$t = Wx + \mu + \in$$

the connection between these two Gaussians is stated, where *t* (*n*-dimensional) indicates the data vector, *x* (m-dimensional) refers to the subspace vector, W indicates the m leading eigenvectors, μ stands the data mean, and ∈ refers to a noise model that is tacit isotropic Gaussian *i.e.*, $\in \sim N(0, \sigma 1)$) which has similarity with the average of the minor eigen values. This allows the next subsequent definitions of probability distributions over *t*-space and *x*-space:

$$p(t) = (2\pi)^{-\frac{n}{2}} |C|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(t-\mu)^T C^{-1}(t-\mu)\right)$$

$$p(t|x) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sigma^2 \|t - Wx - \mu\|^2\right)$$

$$p(x) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}x^T x\right)$$

The Gaussian formula along with the covariance matrix, in general is expressed as:

$$C = \sigma^2 i + WW^T$$

Therefore the preprocessed gene dataset is got here. Here, the preprocessed data undergoes filtering and the gene selection process that enhances the classification performance.

3.3. Mutual Information detection on Micro array dataset

It is very essential to discover the features that possess the maximum information content. With regard to microarray gene expression data, the aim of any type of relevant gene selection process is the identification of genes that have the maximum information corresponding to the class labels of the samples. For the identification of these genes, feature entropy is an appropriate metric. The initial uncertainty of the output class that is called as the entropy is defined below in Equation 1:

$$H(C) = -\sum_{x=1}^{N_x} P_x(x) \log (P_x(x)) ;$$

Where $P_x(x)$, $x = 1, 2, \ldots, N_x$ indicate the probabilities for the different classes, like the P_x which indicates the probability density for class *x*.

Afterwards, the average uncertainty with regard to the input feature vector is computed as the conditional entropy defined in Equation 2:

$$E(X|S) = -\sum_{v=1}^{N_v} P(v) \left(\sum_{x=1}^{N_x} P_x(x|v) \log P_x(x|v)); \right)$$

Here the *s* refers to the input feature vector that has Ns samples and $P_x(x|v)$ indicates the conditional probability for the class *x* obtained from the input vector *v*. Typically, the conditional entropy will be lesser than or will equal the initial entropy. When there exists complete independence between the feature and output class, then the conditional entropy equals the initial entropy. Hence, the mutual information is defined by the quantity of uncertainty which is reduced. The mutual information I(X; S) between the variables *x* and *s* can be expressed as:

$$I(X ; S) = H(X) = H(X|S)$$

The Equation 3 above can be rewritten as:

$$I(X ; S) = I(S ; X) = \sum_{x,v} P(x, v) \log \frac{P(x, v)}{P(x)P(v)}$$

Since the function of mutual information has symmetry with regard to X and S hence I(X; S) equals to I(S ; X).

3.4. Improved Minimum Redundancy Maximum Relevance (mRMR) Filter Algorithm (ImRMR)

The gene selection procedure is highly necessary for the accurate classification prediction and the mRMR technique can considerably enhance the classification accuracy [16]. In the case of a high-dimensional microarray dataset, since there are thousands of genes present, it is not effective to follow an evolutionary algorithm like the artificial bee colony in a microarray dataset directly. Moreover, it is hard for a classifier to be trained with accuracy.

Alternate methods has to be adopted efficiently to resolve this issue. Hence, in the form of a first step, mRMR is utilized for reducing the noisy and unnecessary genes. The mRMR approach was introduced by Peng et al. in 2005 [17]. It is a heuristic method which can be utilized for continuous and discrete datasets for measuring the relevancy and redundancy of features and decide over the features that are promising. The experimentation results show that ImRMR is an efficient technique for improving the performance feature selection. Features that are chosen by ImRMR have more predictive capability and accomplish accurate classification results compared to those chosen by mRMR and MaxRel.

mRMR Approach

This section deals with the well-known minimum redundancy maximum relevance with MIQ and MID scheme [20] which is explained in a detailed manner. The genes having considerably varied expressions in two diverse classes (normal and tumor or two diverse subtypes of cancer) are referred to as the differentially expressed genes [18]. The relevance of a gene is known to be the degree of differentially expression of that gene. Then the relevance of gene can be computed by mutual information [19]. In case the expression of a gene isstochastically or uniformly distributed in diverse classes, then its mutual information with these classes tends to be zero. When a gene is highly differentially expressed for diverse classes, it must possess enormous mutual information. Consider a mutual information for the case of discrete variable only. For the case of discrete variables, the mutual information I of two variables X and S is expressed in Equation 4. The fundamental concept of minimum redundancy is the selection of the genes in such a manner that they are mutually maximally dissimilar to the other genes. Let s represent the subset of genes which are seen. The average minimum redundancy is defined in Equation 5:

$$\text{Minimum } W = \frac{1}{|v|^2} \sum_{i, j \in v} I(i, j),$$

Maximum relevance and minimum relevance redundancy feature selection

Input : Discretized data *d*, class *c*, number of output features *n*, number of features in *d* is *g*.

Output : Output feature set F.

1. Idle $f_t = [1 : g]$
2. For $i = 1: g$ do
3. Relevance (i) = mutual–info ($d(:,i),c$);
4. End for
5. [R, id] = Max (relevance);
6. F[1] = *id*;

7. Idle f_t = idle f_t – F;
8. For i = 2: n do
9. Obj 1 = relevance (idle f_t);
10. For j = 1: |idel f_t| do
11. Sum = $\sum_{k=1}^{F}$(mutual – info(d(:, k), d(:, idel f_t))) ;
12. Redun(j) = sum/ |F|;
13. End for
14. Obj 2= relevance (idle f_t/(redun + 0.0001);
15. [newid, obj2] = Non dominated-Feature Selection (obi1, obj2, idle f_t)
16. [R, id] = Max(obj2);
17. F[i] = id;
18. Idle
19. End for

where indicates the mutual information between i-th gene and j-th gene and represents the number of genes in S. In order to choose the differentially expressed gene, again the mutual information can be utilized. The discriminant capabilities of a genes by the mutual information I(h, g_i)is computed in Equation 6. This indicates that the mutual information between the targeted classesh = h_1, h_2, ..., h_k and the gene expression gives the measure of relevance of that particular gene.This way, the maximum relevance condition is the maximization of the average relevance of every gene in s is given in Equation 6:

$$\text{Maximum} = \frac{1}{|v|}\sum_{i \in v} \mathrm{I}(h, i)$$

Hence, the gene redundancy must be reduced and the gene relevance has to be increased. These two conditions are merged into one single criterion function in mRMR. Since the two conditions have equal importance, then the two simplest combined criteria become: Max(V − W), and Max(V/W). Here the mRMR for discrete variable are described in the form of mRMR mutual information difference (mRMR MID) and mRMR mutual information quotient (mRMR MIQ). The mRMR with MID Scheme is expressed as Equation 7 and mRMR with MIQ Scheme is defined in the Equation 8.

$$m\mathrm{RMR}_{\mathrm{MID}} = \max_{i \in \Omega_v}[\mathrm{I}(i, h) - \frac{1}{|v|}\sum_{j \in v} \mathrm{I}(i, j)],$$

$$m\mathrm{RMR}_{\mathrm{MIQ}} = \max_{i \in \Omega_v}[\mathrm{I}(i, h)/[\frac{1}{|v|}\sum_{j \in v} \mathrm{I}(i, j)]$$

Proposed ImRMR Method

The data matrix will be preprocessed and then discretized corresponding to the mean of every gene's expression (column). The number of the output features (genes) consider n is provided by the user. The data matrix with the classes c = {1, 2, ⋯ ,C} act as the inputs. At the start, the first objective (obj1) *i.e*., the relevance of every gene is computed bythemutual information according to the Equation 6. The highest scorer gene id is extracted from the relevance score and then added in the last solution set. Afterwards, a looping is conducted for the rest of the output features. Now the redundancy observed between the output feature and the rest of the features (idle f t) is computed according to the Equation 5. In case the output feature set has more than one feature, then the mean is treated to be the redundancy score as expressed in Equation 9.

$$\text{mean} - \text{redundancy}\,(i) \;=\; \sum_{k=1}^{F}(\text{mutual} - \text{info}\,[x_k, x_i]))/|F|)$$

where F refers to output feature set, indicates the output feature vector and stands for the ith feature vector. Thereafter the second objective (obj2) can be seen modeled to be the ratio of relevance to the redundancy and it has to get maximized. After the calculation of the two objectives for every feature, the identification of non-dominated features is done. A reference feature is known as the non-dominated feature when it meets the following conditions: 1) when the obj1 of the reference feature is larger than or equals all the other features' obj1 and the obj2 of the reference feature is larger than or equals all the other features' obj2 2) else if the obj1 of the reference feature is larger compared to the remaining other features' obj1 and the obj2 of the reference feature is lesser than every other features' obj2 and vice-versa. Later, the feature having maximum obj2 gets integrated in the output feature set from the non-dominated features. This way an incremental procedure is following for getting the rest of the output features. The newly introduced algorithm is given in Algorithm 1 and the algorithm of the ruling features or non-dominated feature selection is shown in Algorithm 2.

Non-dominated Feature Selection

Input : The feature *id* idle f_t, first objective obj1, second objective obj2,

Output : Non-dominated Feature *id* id_{ns}, the second objective $obje2_{ns}$ of non-dominated features.

1. $k = 1$
2. for $i = 1$: |idle f_t| do
3. $t = 0$;
4. for $j = 1$: |idle f_t| do
5. if then (i! = j)
6. if then (obj1(i) < = obj2(i) < = obj2(j));
7. else if then (obj1(i)< obj1(i) < & obj2(i) > obj2(j) || (obj1(i) > obj1(j) & obj2(i) < obj2(j)
8. else
9. $t = 1$;
10. break;
11. end if
12. end if
13. end for
14. if then(t == 0 & j ==|idle f_t|)
15. id_{ns} (k) = 1
16. $obj2_{ns}$ (k) = obj2(i);
17. $k = k + 1$;
18. end if
19. end for

The filtering of the initial microarray gene expression profiling is done making use of the ImRMR gene selection technique. Every gene gets assessed and then ordered making use of the ImRMR mutual information MI operations. The greatest relevant genes which yield 100% classification accuracy along with an RFSVM classifier are then identified to create a new subset referred to as the ImRMR dataset, as indicated in Figure 2. The mRMR dataset represents the more relevant and lesser redundant genes as chosen by the mRMR technique. The mRMR is used for filtering the genes which are irrelevant and noisy and minimizes the load of computation for the GSO algorithm and RFSVM classifier

3.5. Glowworm Swarm Optimization (GSO) for gene selection

This is used for selecting the genes having the most information and predictive capability from an ImRMR dataset which render the greatest classification accuracy along with an RFSVM classifier. Every solution is indicated in the form of group of genes indices which are chosen form the ImRMR dataset. In a gene selection issue, every solution (*i.e.*, subset of the selected genes) is linked with the fitness value that is the classification accuracy making use of an RFSVM classifier.

Simultaneously capturing the multiple optima of multimodal functions is carried out by using the Glowworm Swarm Optimization algorithm. The algorithm utilizes a set of agents, for the purpose of scanning the search space and then having information exchanged about a fitness of their present position. Degree of a luminescent quantity known as the luciferin indicates the fitness. The increased value of the luciferin will be depicted by the agent when traversing in the direction of randomly selected neighbor. But, it is unlucky that the agent does not traverse at all with no neighbors present and hence tends to become an unnecessary feature diminishing the performance of the algorithm. Also, this attribute can result in imbalanced loads in the case of parallel processing. This gives rise to simple modifications to the actual algorithm that increases the performance of the algorithm by a constraining condition of having no agent relocated.

In GSO a swarm is composed of N agents called glowworms. A state of a glowworm i at time t can be described by the following set of variables: a position in the search space ($x^i(t)$), a luciferin level ($I^i(t)$) and a neighbourhood range ($N_i(t)$).

Luciferin-update phase : The luciferin update depends on the function value at the glowworm position. During the luciferin-update phase, each glowworm adds, to its previous luciferin level, a luciferin quantity proportional to the fitness of its current location in the objective function domain. Also, a fraction of the luciferin value is subtracted to simulate the decay in luciferin with time. The luciferin update rule is given by:

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

where $l_i(t)$ represents the luciferin level associated with glowworm at time t, ρ is the luciferin decay constant ($0 \leq \rho \leq 1$), γ is the luciferin enhancement constant, and $f(x_i(t))$ represents the value of the objective function at agents location at time .

Movement phase : During the time of movement phase, every 'glowworm determines, making use of a probabilistic mechanism, to traverse toward a neighbor which contains a luciferin value greater than its own self. It means that, glowworms get attracted to neighbors which glow brighter. The set consisting of the neighbors of glowworm i at time t is computed as below:

$$N_i(t) = \{j : ||x_j(t) - x_i(t)|| < r_i^d(t) ; l_i(t) < l_i(t)\}$$

where the $||x||$ refers to the Euclidean norm of x, and r_i^d indicates the variable neighborhood range which is interrelated with the glowworm at any time t that is restrained above by a circular sensor range $r_s(0 < r_i^d(t) < r_s)$. For every glowworm i, the probability of moving towards a neighbor $j \in N_i(t)$ is expressed by:

$$p_{ij}(t) = \frac{l_j(t) - l_i(t)}{\sum_{k \in N_i(t)} l_k(t) - l_i(t)}$$

Suppose glowworm i choose a glowworm $j \in N_i(t)$ with $p_{ij}(t)$as given in (5.11). Afterwards, the discrete-time model of the glowworm movements can be given as:

$$xi(t + 1) = x_i(t) + s\left(\frac{x_j(t) - x_i(t)}{||x_j(t) - x_i(t)||}\right)$$

where $x_i(t) \in R^d$ stands for the location of glowworm i, at any time t, in the d-dimensional real space R_d, and $s >$ refers to the step size.

Neighborhood range update rule: Associate every agent i with a neighborhood the radial range $r_i^d(t)$ of whose is dynamic by nature. Assume r_0 refer to the initial neighborhood range of everyglowworm (that is, $r_i^d(0) = r_0, \forall i$). In order to update the neighborhood range of every glowworm adaptively, the rule is as below:

$$r_i^d(t+1) = \min\{r_s, \max\{0, r_i^d(t) + \beta(n_i - |N_i(t)|\}\}$$

where β refers to a constant parameter and n_i indicates a parameter employed for controlling the number of neighbors.

GSO Algorithm

The GSO algorithm is as given below :

Assume that,

1. Luciferin decay constant (ρ) = (0 to 1) = 0.4
2. The luciferin enhancement constant (γ) = (0 to 1) = 0.6
3. The step size (s) = (0 to 0.1)
4. The sensor range (r_s) = 0.5
5. Constant Value (β) = (0. to 0.1)
6. $n = 50$
7. $r_d^0 = 5$
8. $n_t = 5$

Procedure

1. Set parameters: n, $l0$, $r0$, ρ, γ, β, s, rs, nt
2. Arbitrarily generate the population of glow-worms x_i = (4,2,2) (2,2) = {4R, 2R, 2R, R, R} = (4, 2, 2,) (2, 2)
3. for i = 1 to n do
4. Initialize luciferin $l_i(1) = l_1 = 5$
5. Initialize neighborhood range $r_d^i = 5$
6. end for
7. $t = 1$
8. while stop condition not met do
9. for each glow-worm i do {update luciferin}

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

10. Local-decision range update:

$$r_d^i(t+1) = \min\{r_s, \max\{0, r_d^i(t) + \beta(n_t - |N_i(t)|)\}\} ;$$
$$r_d^0 = 5$$

11. The number of glow in local-decision range:

$$N_i(t) = \{J : ||x_j(t) - x_i(t)|| < r_d^i ; l_i(t) < l_j(t)\}$$
$$x_j(t) = \{2, 1\}$$

12. If one iteration is finished, move into the next subsequent iteration, judges if the termination condition is to be satisfied, if the withdrawal circulation is to be satisfied, the record result, else again repeat the same process by choosing another neighbour.

3.6. Classification using Modified Support Vector Machine (RFSVM)

In this work, microarray gene data set are divided making use of a hybrid classification technique that exploits the Random forest, SVM classifier and boosting ensemble learning technique. In the hybrid method, the input data set is subdivided into subsets randomly. Every data item in all of the subsets contains a weight factor that is associated with it. The data items present in the subsets get classified by SVM classifier. In case a misclassification has happened, then the weight factor of the data items is raised else it gets decreased. The data subsets are then rearranged and once more the SVM classifier is employed for conducting the classification at every subset. The weights are updated again based on if it is a right classification or a misclassification. These steps are then iteratively repeated until all of the weights are updated to a very lesser value. The output from the input data set is then computed by using a voting strategy to every random subset classification outputs [21]. The algorithm for the new hybrid method is provided in the sample code as follows:

Algorithm 1 Hybrid classification using RF and SVM supplemented by boosting

Input : D Training Instances Intermediate Output: Osvm, Classification output at every feature subset Output: O, Classification Output for the hybrid methodology

Step 1. Begin

Step 2. Initialize the weight *wi* for each data vectort *i* ε D.

Step 3. Generate a new data feature subset Di from D using random replacement method.

Step 4. Begin

Step 5. Forevery random feature subset Di do Step

Step 6. Begin

Step 7. Apply SVM to each feature subset

Step 8. Generate Osvm, the classification output from

Step 9. End

Step 10. Update the weights of every data vector in the training set based on the classification outcome. If an example was misclassified then its weight is increased, else the weight is reduced.

Step 11. Repeat the steps 2 to 10 by regenerating the random subsets until every input data vector is suitably classified or apply iteration constraint.

Step 12. Calculate output O of the entire data set by using majority voting technique among the final outputs of every Random feature subset Di of The original set D is obtained after Step 11.

Step 13. Return O

Step 14. End

4. RESULTS AND DISCUSSION

In this section, evaluate the overall performance of gene selection methods using six popular binary and multiclass microarray cancer datasets, which were downloaded from http://www.gems-system.org/. These datasets have been widely used to benchmark the performance of gene selection methods in bioinformatics field. The binary-class microarray datasets are colon [22], leukemia [22, 23], and lung [24] while the multiclass microarray datasets are SRBCT [25], lymphoma [26], and leukemia [27]. In Table 1, a detailed description of these six benchmark microarray gene expression datasets with respect to the number of classes, number of samples, number of genes, and a brief description of each dataset construction.

Table 1
Statistics of microarray cancer datasets

Microarray datasets	*Number of classes*	*Number of samples*	*Number of genes*	*Description*
Colon [22]	2	62	2000	40 cancer samples and 22 normal samples
Leukemia1 [23]	2	72	7129	25 AML samples and 47 ALL samples
Lung [24]	2	96	7129	86 cancer samples and 10 normal samples
SRBCT [25]	4	83	2308	29 EWS samples, 18 NB samples, 11 BL samples, and 25 RMS samples
Lymphoma [26]	3	62	4026	42 DLBCL samples, 9 FL samples, and 11 B-CLL samples
Leukemia2 [27]	3	72	7129	28 AML sample, 24 ALL sample, and 20 MLL samples

Table 2 shows the control parameters for the ImRMR-GSO algorithm that was used in our experiments. The first control parameter is the bee colony size or population, with a value of 80. The second control parameter is the maximum cycle, which is equal to the maximum number of generations. A value of 100 is used for this parameter. Another control parameter is the number of runs, which was used as stopping criterion, and used a value of 30 in our experiments, which has been shown to be acceptable. A value of 5 iterations is used for this parameter.

Table 2
ImRMR-GSO control parameters

Parameter	*Value*
Population Size	80
Max cycle	100
Number of runs	30
Limit	5

In this study, the performance of the proposed ImRMR-GSO algorithm is tested by comparing it with other standard bioinspired algorithms, including ABC, GA, and PSO. Compare the performance of each gene selection approach based on two parameters: the classification accuracy and the number of predictive genes that have been used for cancer classification. Classification accuracy is the overall correctness of the classifier and is calculated as the sum of correct cancer classifications divided by the total number of classifications. It is computed by the expression shown below:

$$\text{Classification Accuracy} = \frac{CC}{N} \times 100$$

where N is the total number of the instances in the initial microarray dataset. And, CC refers to correctly classified instances.

Apply leave-one-out cross validation (LOOCV) [28] in order to evaluate the performance of our proposed algorithm and the existing methods in the literature. LOOCV is very suitable to our problem because it has the ability to prevent the "overfitting" problem [28]. It also provides an unbiased estimate of the generalization error for stable classifiers such as the SVM classifier. In LOOCV, one sample from the original dataset is considered testing dataset, and the remaining samples are considered training dataset. This is repeated such that each sample in the microarray dataset is used once as the testing dataset. Implement GA, PSO algorithm, and SVM using the Waikato Environment for Knowledge Analysis (WEKA version

3.6.10), an open source data mining tool [29]. Furthermore, in order to make experiments more statistically valid, conduct each experiment 30 times on each dataset. In addition, best, worst, and average results of the classification accuracies of the 30 independent runs are calculated in order to evaluate the performance of the proposed algorithm.

Performance Evaluation

In this section, analyze the results that are obtained by the proposed algorithm. As a first step, employ the ImRMR method to identify the top relevant genes that give 100% accuracy with an RFSVM classifier. From Table 3 and Figure 2, can see that the top150 genes in the leukemia1 dataset generate 100% classification accuracy while in the colon dataset, can get 100% accuracy using 350 genes. For the lung dataset, achieved 100% accuracy using 200 genes and 250 genes to get the same classification accuracy for the SRBCT dataset. In addition, using 150 high relevant genes from the lymphoma dataset and 250 genes from the leukemia2 dataset, achieved 100% classification accuracy. Then used these high relevant genes as input in the proposed GSO algorithm to determine the most predictive and informative genes.

Table 3
The classification accuracy performance of the mRMR method with an RFSVM classifier for all microarray datasets

Number of genes	*Colon*	*Leukemia1*	*Lung*	*SRBCT*	*Lymphoma*	*Leukemia2*
50	91.94%	91.66%	89.56%	62.65%	93.93%	77.77%
100	93.55%	97.22%	95.83%	91.44%	98.48%	86.11%
150	95.16%	100%	98.95%	96.39%	100%	98.61%
200	96.77%	100%	100%	97.59%	100%	100%
250	98.38%	100%	100%	100%	100%	100%
300	98.38%	100%	100%	100%	100%	100%
350	100%	100%	100%	100%	100%	100%
400	100%	100%	100%	100%	100%	100%

Compare the performance of the proposed ImRMR-GSO algorithm and the existing mRMR-ABC, when using RFSVM as a classifier with the same number of selected genes for all six benchmark microarray datasets.

Table 4
Comparison between ImRMR-GSO, mRMR-ABC classification performance when applied with the RFSVM classifier for colon dataset

Classification Accuracy in (%)		
Number of genes	*mRMR-ABC*	*Proposed ImRMR-GSO*
3	87.50	88.00
4	88.27	89.90
5	89.50	90.00
6	90.12	90.80
7	91.64	92.00
8	91.80	92.20
9	92.11	92.75
10	92.74	93.10
15	93.60	94.00
20	94.17	94.80

The comparison results for the binary-class microarray datasets: colon, leukemia1, and lung are shown in Tables 4, 5, and 6, respectively while Tables 7, 8, and 9, respectively, present the comparison result for multiclass microarray datasets: SRBCT, lymphoma, and leukemia2. From these tables, it is clear that our proposed ImRMR-GSO algorithm performs better than the original ABC algorithm in every single case (*i.e.*, all datasets using a different number of selected genes).

Table 5
Comparison between ImRMR-GSO, mRMR-ABC classification performance when applied with the RFSVM classifier for leukemia1 dataset

Classification Accuracy in (%)		
Number of genes	*mRMR-ABC*	*Proposed ImRMR-GSO*
2	89.63	90
3	90.37	91
4	91.29	92
5	92.82	93
6	92.82	93
7	93.10	93.50
10	94.44	95
13	94.93	95
14	95.83	96

Table 6
Comparison between ImRMR-GSO, mRMR-ABC classification performance when applied with the RFSVM classifier for Lungdataset

Classification Accuracy in (%)		
Number of genes	*mRMR-ABC*	*Proposed ImRMR-GSO*
2	95.83	96
3	96.31	97
4	97.91	98
5	97.98	99
6	98.27	98.60
7	98.53	98.85
8	98.95	99

Table 7
Comparison between ImRMR-GSO, mRMR-ABC classification performance when applied with the RFSVM classifier for SRBCT dataset

Classification Accuracy in (%)		
Number of genes	***mRMR-ABC***	***Proposed ImRMR-GSO***
2	71.08	71.60
3	79.51	80.00
4	84.33	84.90
5	86.74	87.00
6	91.56	92.00
7	94.05	94.50
8	96.30	96.90

Table 8
Comparison between ImRMR-GSO, mRMR-ABC classification performance when applied with the RFSVM classifier for lymphoma dataset

Classification Accuracy in (%)		
Number of genes	*mRMR-ABC*	*Proposed ImRMR-GSO*
2	86.36	86.90
3	90.90	91.20
4	92.42	92.80
5	96.96	97.10

Table 9
Comparison between ImRMR-GSO, mRMR-ABC classification performance when applied with the RFSVM classifier for Leukemia 2 dataset.

Classification Accuracy in (%)		
Number of genes	*mRMR-ABC*	*Proposed ImRMR-GSO*
2	84.72	85.03
3	86.11	86.50
4	87.5	87.90
5	88.88	89.00
6	90.27	90.65
7	89.49	89.90
8	91.66	92.05
9	92.38	92.70
10	91.66	92.10
15	94.44	94.85
18	95.67	96.00
20	96.12	96.50

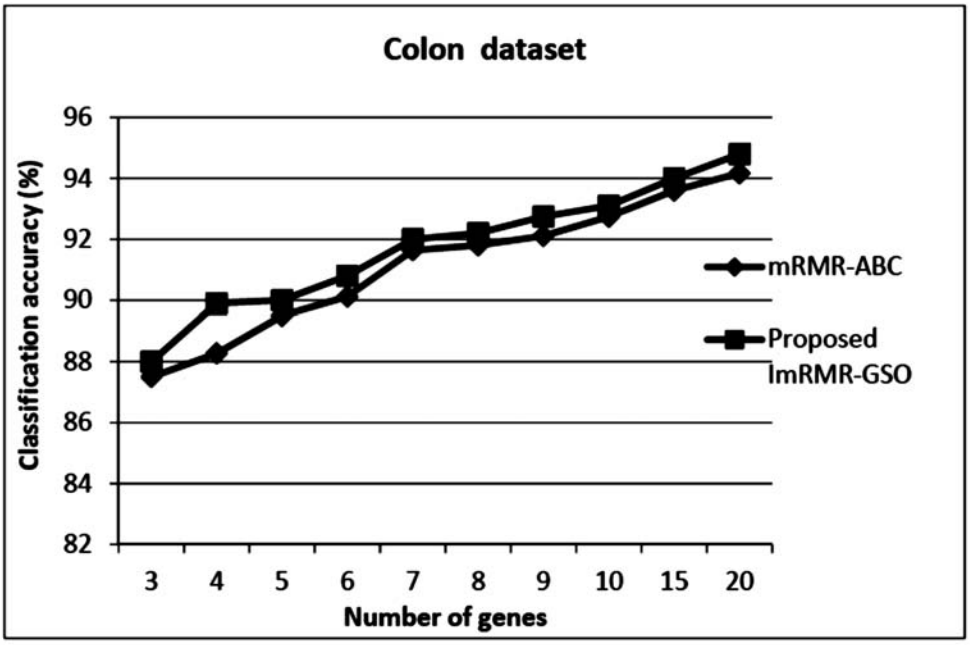

Figure 2: Feature selection results comparison for colon dataset

The comparison results for the binary-class microarray datasets: colon, leukemia1, and lung are shown in Figure 2,3, and 4, respectively while Figures 5, 6, and 7, respectively, present the comparison result for multiclass microarray datasets: SRBCT, lymphoma, and leukemia2. From these tables, it is clear that our proposed ImRMR-GSO algorithm performs better than the original ABC algorithm in every single case (*i.e.*, all datasets using a different number of selected genes).

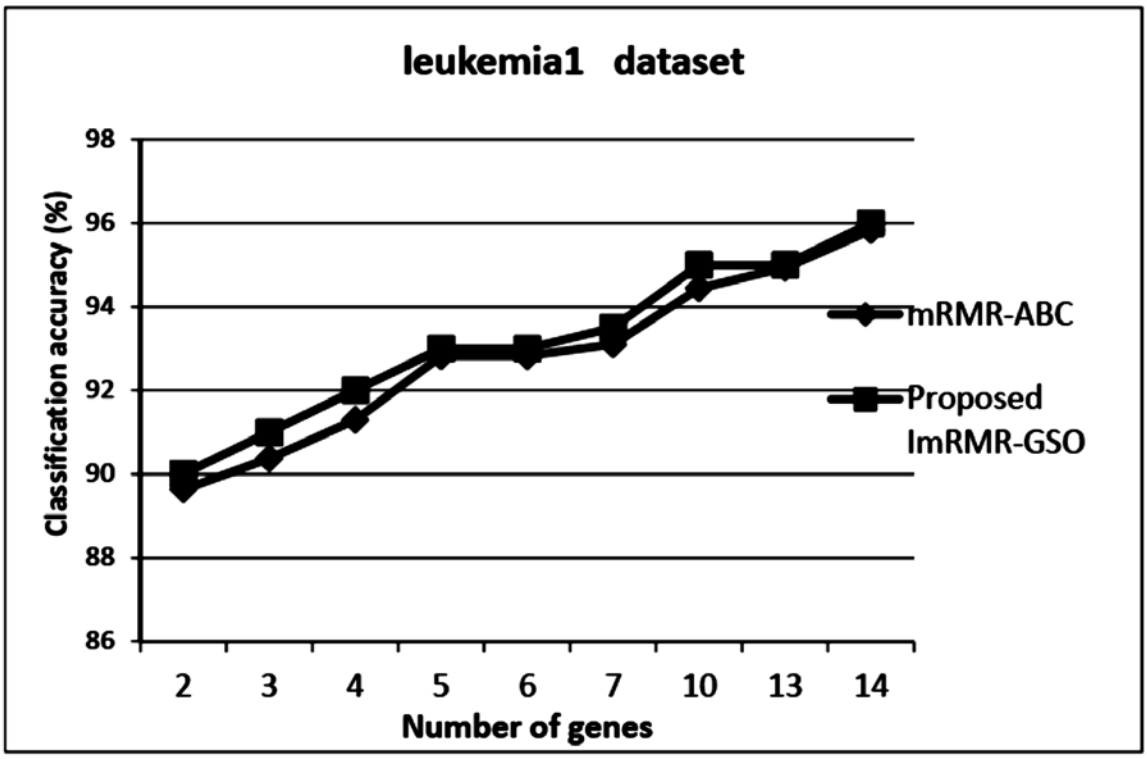

Figure 3: Feature selection results comparison for leukemia1 dataset

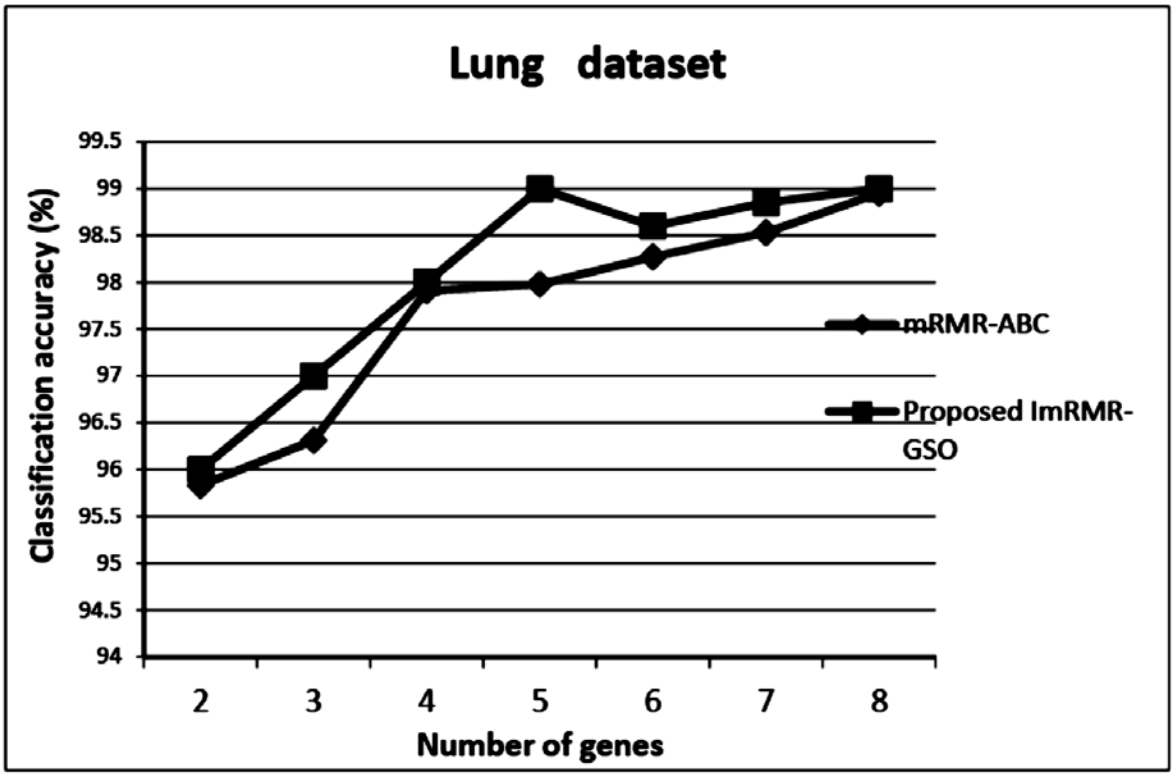

Figure 4: Feature selection results comparison for Lung dataset

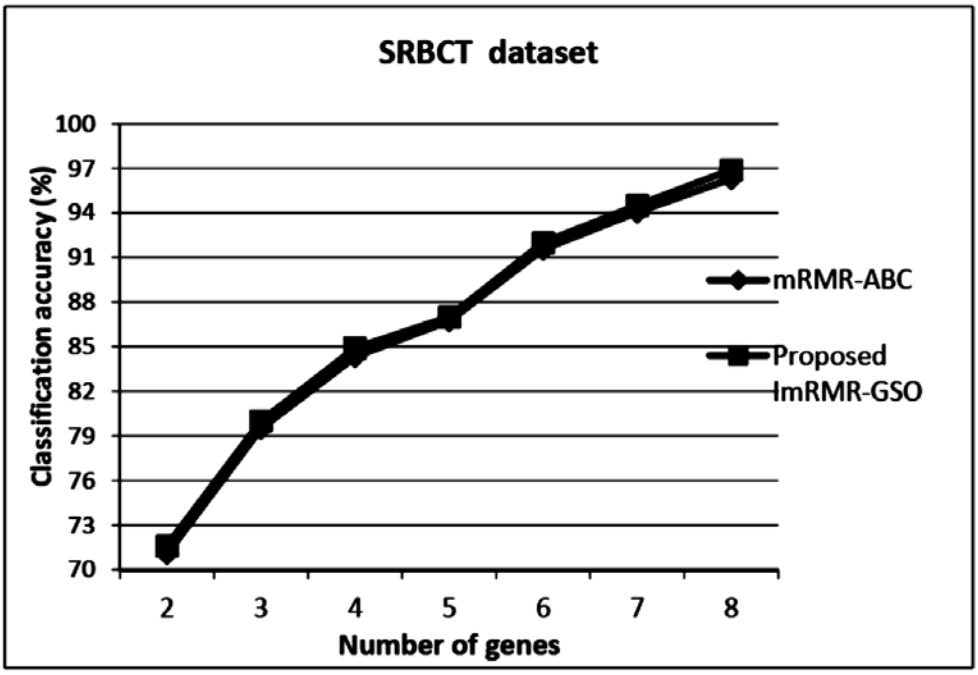

Figure 5: Feature selection results comparison for SRBCT dataset

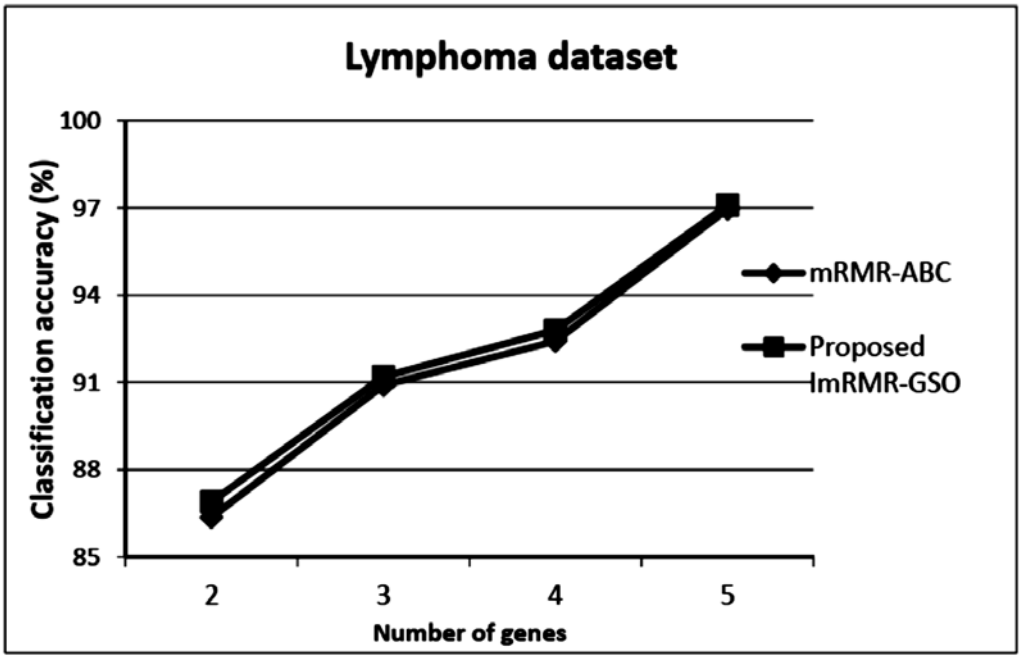

Figure 6: Feature selection results comparison for Lymphoma dataset

The explanation of the best predictive and highly frequent genes that give highest classification accuracy for all microarray datasets using ImRMR-GSO algorithm has been reported in Table 10. It is worth mentioning that the accuracy of the ImRMR filter method when it is combined with GSO generally outperforms the classification accuracy of GSO algorithm without ImRMR. Thus, the ImRMR is a promising method for identifying the relevant genes and omitting the redundant and noisy genes. We can conclude that the proposed ImRMR-GSO algorithm generates accurate classification performance with minimum number of selected genes when tested using all datasets as compared to the original GSO algorithm under the same cross validation approach. Therefore, the ImRMR-GSO algorithm is a promising approach for solving gene selection and cancer classification problems.

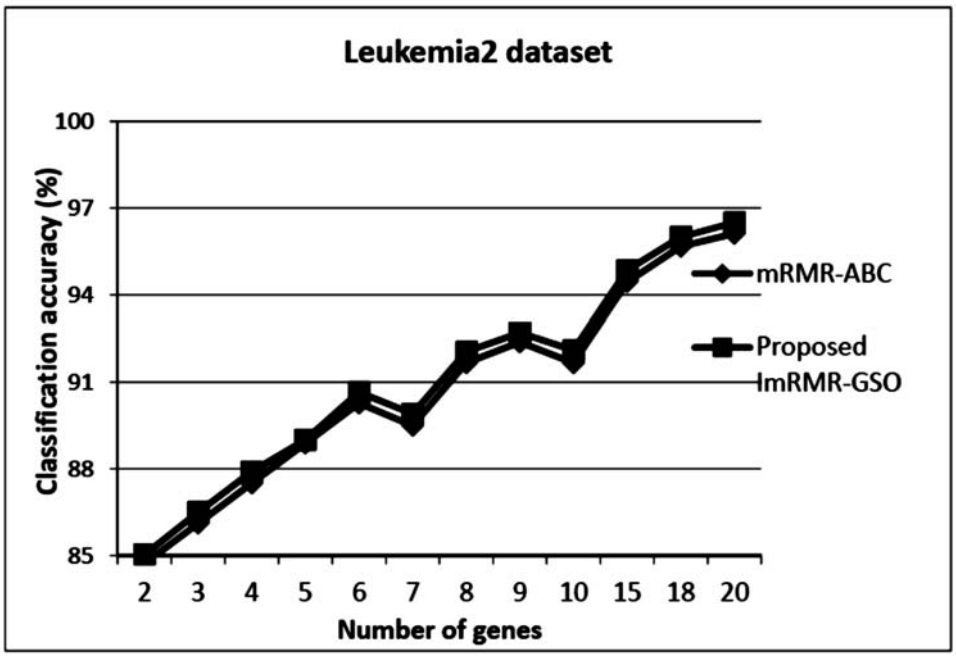

Figure 7: Feature selection results comparison for Leukemia2 dataset

Table 10
The best predictive genes that give highest classification accuracy for all microarray datasets using ImRMR-GSO algorithm

Datasets	*Predictive genes*	*Accuracy (%)*
Colon	Gene115, Gene161, Gene57, Gene70, Gene12, Gene132, Gene84, Gene62, Gene26, Gene155, Gene39, Gene14, Gene1924, Gene148, and Gene21	97.50
Leukemia1	M31994 at, U07563 cds1 at, Y07604 at, J03925 at, X03484 at, U43522 at, U12622 at, L77864 at, HG3707-HT3922 f at, D49950 at, HG4011-HT4804 s at, Y07755 at, M81830 at, and U03090 at	100
Lung	U77827 at, D49728 at, HG3976-HT4246 at, X77588 s at, M21535 at, L29433 at, U60115 at, and M14764 at	100
SRBCT	Gene795, Gene575, Gene423, Gene2025, Gene1090, Gene1611, Gene1389, Gene338, Gene1, and Gene715	100
Lymphoma	Gene1219X, Gene656X, Gene2075X, Gene3344X, and Gene345X	100
Leukemia2	Y09615 atD87683 at, U31973 s at, U68031 at, V00571 rna1 at, L39009 at, U37529 at, U35407 at, X93511 s at, L15533 rna1 at, X00695 s at, H46990 at, U47686 s at, L27624 s at, S76473 s at, X16281 at, M37981 at, M89957 at, L05597 at, and X07696 at	100

5. CONCLUSION

In this research paper, we proposed applying GSO algorithm for microarray gene expression profile. A new swarm based algorithm called hybrid gene selection approach to be combined with RFSVM as a classifier. It can be used to solve classification problems that deal with high-dimensional datasets, especially microarray gene expression profile. Up to our knowledge, the GSO algorithm has not yet been applied as a gene selection technique for a microarray dataset, so this is the first attempt. Our proposed ImRMR-GSO algorithm is a three-phase method; the ImRMR filter technique is adopted to identify the relative and informative gene subset from the candidate microarray dataset. Then the GSO algorithm is employed to select the predictive genes from the ImRMR genes subset. Finally,

the RFSVM classifier was trained and tested using the selected genes and returned the classification accuracy. Extensive experiments were conducted using six binary and multiclass microarray datasets. The results showed that the proposed algorithm achieves superior improvement when it is compared with the other previously proposed algorithms.

6. REFERENCES

1. L.Y. Chuang, C.H. Yang, K.C. Wu, and C.H. Yang, "A hybrid feature selection method for dna microarray data," Computers in Biology and Medicine, Vol. 41, No. 4, pp. 228–237, 2011.
2. H. Yu and S. Xu, "Simple rule-based ensemble classifiers for cancer dna microarray data classification," in Computer Science and Service System (CSSS), 2011 International Conference on, pp. 2555– 2558, 2011.
3. C. FENG and W. LIPO, "Applications of support vector machines to cancer classification with microarray data," International Journal of Neural Systems, Vol. 15, No. 06, pp. 475–484, 2005.
4. E. lba, J. García-Nieto, L. Jourdan, and E.G. Talbi, "Gene selection in cancer classification usingpso/svm and ga/svm hybrid algorithms," in Evolutionary Computation, IEEE Congress on, pp. 284–290, 2007.
5. S. Ghorai, A. Mukherjee, S. Sengupta, and P. Dutta, "Multicategory cancer classification from gene expression data by multiclass nppc ensemble," in Systems in Medicine and Biology (ICSMB), International Conference on, 2010, pp. 4–48.
6. G. Sheng-Bo, L.M.R., and T.M. Lok, "Gene selection based on mutual information for the classification of multi-class cancer," in Proceedings of the 2006 international conference on Computational Intelligence and Bioinformatics, Springer-Verlag, pp. 454–463, 2006.
7. L.M. Fu and C.S. Fu-Liu, "Multi-class cancer subtype classification based on gene expression signatures with reliability analysis," FEBS Letters, Vol. 561, No. 13, pp. 186 –190, 2004.
8. H.M. Alshamlan, G.H. Badr, and Y.A. Alohali, "The performance of bio-inspired evolutionary gene selection methods for cancer classification using microarray dataset," International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 4, No. 3, pp. 166–170, 2014.
9. Y. Li, and A. Ngom, "The non-negative matrix factorization toolbox for biological data mining," Source code for biology and medicine, Vol. 8, No. 1, pp. 1, 2013.
10. L. Sheng, R. Pique-Regi, S. Asgharzadeh,., and A. Ortega, "Microarray classification using block diagonal linear discriminant analysis with embedded feature selection," Acoustics, Speech and Signal Processing, pp. 1757-1760, 2009.
11. Automatic detection of lung cancer nodules by employing intelligent fuzzy cmeans and support vector machine ",Biomedical Research M.R. Daliri, "A hybrid automatic system for the diagnosis of lung cancer based on genetic algorithm and fuzzy extreme learning machines," Journal of medical systems, Vol. 36, No. 2, pp. 1001-1005, 2012.
12. S. Saraswathi, S. Sundaram, N. Sundararajan, M. Zimmermann, and M. Nilsen-Hamilton, "ICGA-PSO-ELM approach for accurate multiclass cancer classification resulting in reduced gene sets in which genes encoding secreted proteins are highly represented," Computational Biology and Bioinformatics, IEEE/ACM Transactions on, Vol. 8, No. 2, pp. 452-463, 2011.
13. C.V. Subbulakshmi, and S.N. Deepa, "Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier," The Scientific World Journal, 2015.
14. H.J. Rong, G.B. Huang, N. Sundararajan, and P. Saratchandran, "Online sequential fuzzy extreme learning machine for function approximation and classification problems," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, Vol. 39, No. 4, pp. 1067-1072, 2009.
15. C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," Journal of Bioinformatics and Computational Biology, Vol. 3, No. 2, pp. 185–205, 2005.
16. "Cluster based Key Management Authentication in Wireless Bio Sensor Network ", ,International Journal of pharma and bio sciences, Impact Factor = 5.121(Scopus Indexed).
17. Li, X. Tang, W. Zhao, J. Huang, "A new framework for identifying differentially expressed genes," Pattern Recognition, Vol. 40, pp. 3249– 3262, 2007.
18. T.M. Cover, J.A. Thomas, "Entropy, relative entropy and mutual information," Elements of Information Theory, John Wiley &Sons, Vol. 2, pp. 1-55, 2006.
19. R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Transactions on Neural Networks, Vol. 5, No. 4, pp. 537–550, 1994.

20. Automatic detection of lung cancer nodules by employing intelligent fuzzy cmeans and support vector machine “,Biomedical Research .
21. U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” Proceedings of the National Academy of Sciences of the United States of America, Vol. 96, No. 12, pp. 6745–6750, 1999.
22. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, and C.D. Bloomfield, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” Science, Vol. 286, No. 5439, pp. 531–527, 1999.
23. D.G. Beer, S.L. Kardia, C.C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, and M.L. Lizyness, “Gene-expression profiles predict survival of patients with lung adenocarcinoma,” Nature Medicine, Vol. 8, No. 8, pp. 816–824, 2002.
24. J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, and P.S. Meltzer, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” Nature Medicine, Vol. 7, No. 6, pp. 673–679, 2001.
25. A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, and J.I.. Powell, “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling,” Nature, Vol. 403, No. 6769, pp. 503–511, 2000.
26. S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, and S.J. Korsmeyer, “MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia,” Nature Genetics, Vol. 30, No. 1, pp. 41–47, 2001.
27. A.Y. Ng, “Preventing ‘overfitting’ of cross-validation data,” in Proceedings of the 14th International Conference on Machine Learning (ICML ’97), pp. 245–253, 1997.
28. New Zealand University of Waikato, “Waikato environment for knowledge analysis,” http://www.cs.waikato.ac.nz/ml/weka/ downloading.html.
29. R Mohanasundaram and PS Periasamy, “Hybrid Swarm Intelligence Optimization Approach for Optimal Data Storage Position Identification in Wireless Sensor Networks”, The Scientific World Journal (Hindawi Publishing Corporation) Vol. 2015, Pages 1-12, 2015.