Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 9, September -2017

Microarray Gene Expression Data Pre-Processing Using PPCA and Classification using RF-SVM Algorithm

Ms. N. Kanchana, MCA, M.Phil.(Ph.D),
Assistant Professor and Part Time Ph.D. Research Scholar,
Department of Computer Science,
Dr.G.R.Damodaran College of Science,Coimbatore-641014. Tmail Nadu, India.

Dr.N.Muthumani,M.Sc.(CC), M.Phil, Ph.D,
Associate Professor, Department of Computer Applications,
Sri Ramakrishna College of Arts and Science,
Coimbatore-641006. Tamil Nadu, India.

Abstract – *Various recent research have shown that microarray gene expression data is useful for cancer classification and 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 Support Vector Machine Recursive Feature Elimination with Minimum Redundancy–Maximum Relevancy (SVM-RFE with MRMR Filter) algorithm is proposed to minimize the redundancy among the selected genes. It also improves the accuracy of classification and yields smaller gene sets on several benchmark cancer gene expression datasets. This method outperforms compared to other popular gene selection methods. The RF-SVM (Random Forest-SVM Classifies) algorithm 2 is applied to classify the genes and our experimental results shows that the proposed algorithm classifies accurately compare to other existing algorithms. The SVM-RFE with MRMR Filter algorithm 1 which is applied before classification for feature selection also performed well with small amount of predictive genes when tested using both datasets and compared against previously suggested schemes. Finally the result proves that the proposed RF-SVM (Random forest - SVM classifier) is a promising approach for cancer classification problems.*

Keywords: *Gene Expression, Cancer gene Classification, Gene Selection, SVM-RFE with MRMR Filter, RF-SVM.*

I. 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.

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].

The classifier characteristics such as SVM weights in SVM-RFE provide a criterion to rank genes based on their relevancy, but they do not account for the redundancy among the genes. Our aim is to combine classifier characteristics with a filter criterion that could minimize the redundancy among selected genes, resulting in a selection of a small subset of genes and improved classification accuracy. In this paper, we propose an approach that incorporate mutual-information-based MRMR filter in SVM-RFE to minimize the redundancy among the selected genes. As seen later, our approach, referred to as *SVM-RFE with MRMR filter*, improved the accuracy of classification and yielded smaller gene sets on several benchmark cancer gene expression datasets. Experiments show that our method outperforms MRMR and SVM-RFE methods, as well as other popular methods on most datasets, and selects genes that are biologically relevant in discriminating cancer samples and have properties belonging to the same pathway.

II. 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.

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].

III. PROPOSED METHODOLOGY

In this section, the novel SVM-RFE with MRMR Filter 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 SVM-RFE and several improvements have been recently suggested[19]. SVM-RFE, starting with all the genes, removes the gene that is least significant for classification recursively in a backward elimination manner[18]. The MRMR filter method alone will not yield optimal accuracy because the classifier performs independently and is not involved in the selection of genes[17]. On the other hand, SVMRFE does not take into account the redundancy among genes. Our aim is to improve the gene selection in SVM-RFE by introducing an MRMR filter to minimize the redundancy among relevant genes[16]. As seen later, this improves the classification accuracy by compromising relevancy and redundancy of genes relating to cancer. In our approach of SVM-RFE with MRMR filter, the genes are ranked by a convex combination of the relevancy given by SVM weights and the MRMR criterion. Finally when Classification using Modified Support Vector Machine (RFSVM) is used to get improved classification. The system flow diagram is illustrated in the figure 1.

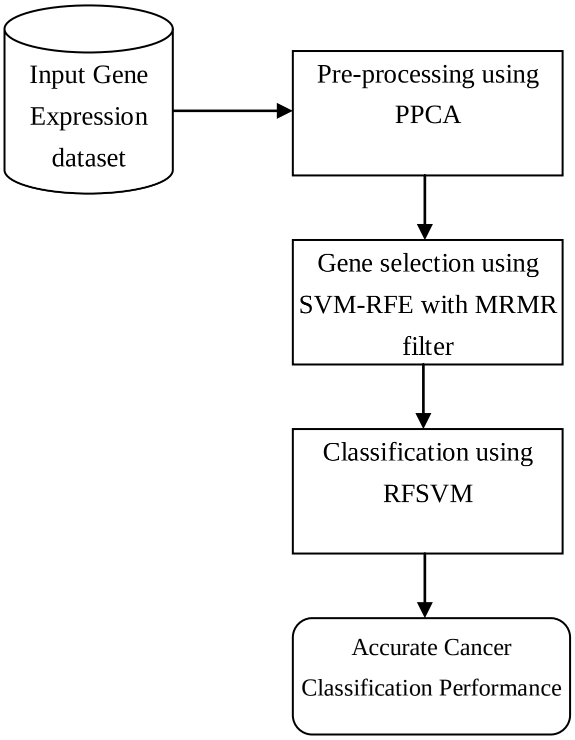

Input Gene Expression dataset
Pre-processing using PPCA
Gene selection using SVM-RFE with MRMR filter
Classification using RFSVM
Accurate Cancer Classification Performance

Figure 1: Architecture view representation of the contribution

A. 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.

B. 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

$$t = Wx + \mu + \square$$

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 ϵ refers to a noise model that is tacit isotropic Gaussian $\mathrm{i.e.}\ \square \sim N(0, \sigma I))$ 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:

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

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

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

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.

C. 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:

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

Where $P_x(x)$, $x = 1,2, \dots, N_x$ 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:

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

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).

D. SVM-RFE with MRMR Filter

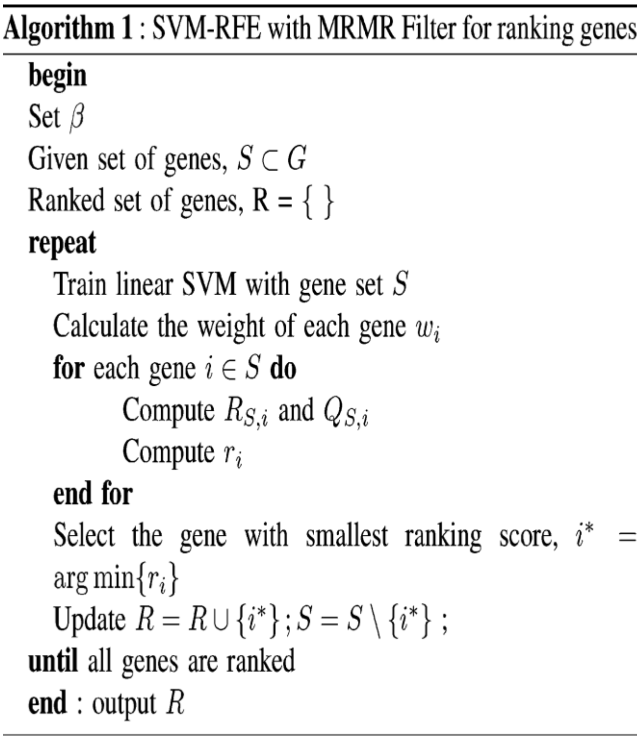
The MRMR filter, when used alone, may not yield optimal accuracy because the classifier performs independently and is not involved in the selection of genes[16][17]. On the other hand, SVMRFE does not take into account the redundancy among genes. Our aim is to improve the gene selection in SVM-RFE by introducing an MRMR filter to minimize the redundancy among relevant genes[18][19][20]. As seen later, this improves the classification accuracy by compromising relevancy and redundancy of genes relating to cancer. In our approach of SVM-RFE with MRMR filter, the genes are ranked by a convex combination of the relevancy given by SVM weights and the MRMR criterion. For *i* th gene, the ranking measure *ri* is given by

$$r_i = \beta\,|w_i| + (1-\beta)\frac{R_{S,i}}{Q_{S,i}}$$

where the parameter *β* ∈ [0, 1] determines the tradeoff between SVM ranking and MRMR ranking, and the relevancy *RS,i* of gene *i* in the set *S* on classification is given by

$$R_{S,i} = \frac{1}{|S|}\sum_{l} I(l,i) \qquad \forall i \in S.$$

To facilitate the backward selection, we use gene-wise MRMR criterion for ranking in the present method. Also, we use |*wi* | as the gene relevancy measure from SVM to better compromise with redundancy of genes. Algorithm 1 illustrates an iteration of SVM-RFE with MRMR filter method of ranking genes: the least important gene at a time is identified after ranking the genes in the subset $S \subset G$. In each iteration, one (or more) of least significant genes are removed and the remaining subset will go through the removal process iteratively, until the removal of any more genes does not improve the classifier performance.



Algorithm 1 : SVM-RFE with MRMR Filter for ranking genes

```
begin
Set $\beta$
Given set of genes, $S \subset G$
Ranked set of genes, R = { }
repeat
  Train linear SVM with gene set $S$
  Calculate the weight of each gene $w_i$
  for each gene $i \in S$ do
      Compute $R_{S,i}$ and $Q_{S,i}$
      Compute $r_i$
  end for
  Select the gene with smallest ranking score, $i^* = \arg\min\{r_i\}$
  Update $R = R \cup \{i^*\}; S = S \setminus \{i^*\}$ ;
until all genes are ranked
end : output $R$
```

E. **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 subsetscontains 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 raisedelse it getsdecreased. The data subsets are then rearranged and once more the SVM classifier is employed for conducting the classification at every subset. The weights are updatedagainbased 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 2: 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 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

IV. 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 SVM-RFE with MRMR Filter 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: SVM-RFE with MRMR Filter control parameters

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

In this study, the performance of the proposed SVM-RFE with MRMR Filter 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:

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

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 SVM-RFE with MRMR Filter 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 SVM-RFE with MRMR Filter 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.

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 SVM-RFE with MRMR Filter algorithm performs better than the original ABC algorithm in every single case (i.e., all datasets using a different number of selected genes).

Table 4: Comparison between SVM-RFE with MRMR Filter, mRMR-ABC classification performance when applied with the RFSVM classifier for colon dataset

Classification Accuracy in (%)		
Number of genes	**mRMR-ABC**	**Proposed SVM-RFE with MRMR Filter**
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

Table 5: Comparison between SVM-RFE with MRMR Filter, mRMR-ABC classification performance when applied with the RFSVM classifier for leukemia1 dataset

Classification Accuracy in (%)		
Number of genes	**mRMR-ABC**	**Proposed SVM-RFE with MRMR Filter**
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 SVM-RFE with MRMR Filter, mRMR-ABC classification performance when applied with the RFSVM classifier for Lungdataset

Classification Accuracy in (%)		
Number of genes	**mRMR-ABC**	**Proposed SVM-RFE with MRMR Filter**
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 SVM-RFE with MRMR Filter, mRMR-ABC classification performance when applied with the RFSVM classifier for SRBCT dataset

Classification Accuracy in (%)		
Number of genes	**mRMR-ABC**	**Proposed SVM-RFE with MRMR Filter**
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 SVM-RFE with MRMR Filter, mRMR-ABC classification performance when applied with the RFSVM classifier for lymphoma dataset

Classification Accuracy in (%)		
Number of genes	**mRMR-ABC**	**Proposed SVM-RFE with MRMR Filter**
2	86.36	86.90
3	90.90	91.20
4	92.42	92.80
5	96.96	97.10

Table 9: Comparison between SVM-RFE with MRMR Filter, mRMR-ABC classification performance when applied with the RFSVM classifier for Leukemia2 dataset

Classification Accuracy in (%)		
Number of genes	**mRMR-ABC**	**Proposed SVM-RFE with MRMR Filter**
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

Table 10: The best predictive genes that give highest classification accuracy for all microarray datasets using SVM-RFE with MRMR Filter 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

The comparison results for the binary-class microarray datasets: colon, leukemia1, and lung are shown in Figure 2,3, and 4, respectively while Figures5, 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 SVM-RFE with MRMR Filter algorithm performs better than the original ABC algorithm in every single case (i.e., all datasets using a different number of selected genes).

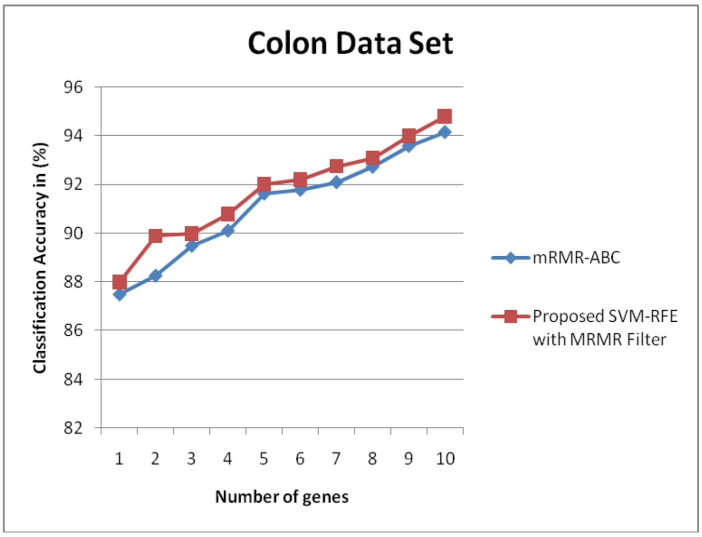


Figure 2: Feature selection results comparison for colon dataset

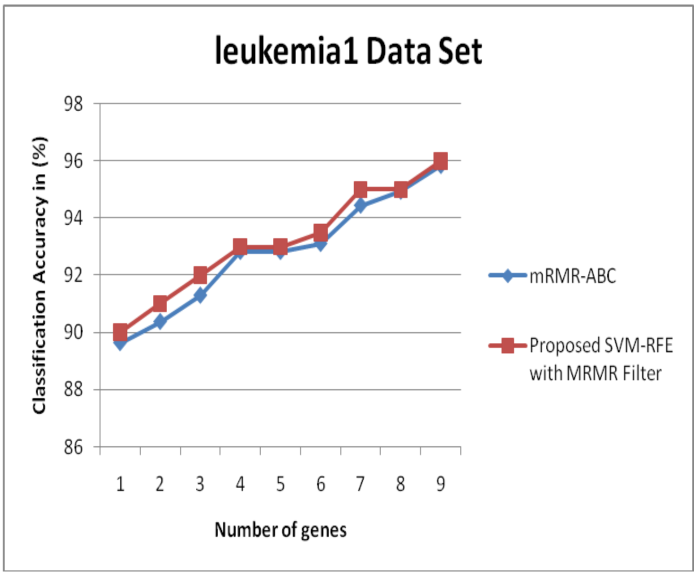


Figure 3: Feature selection results comparison for leukemia1 dataset

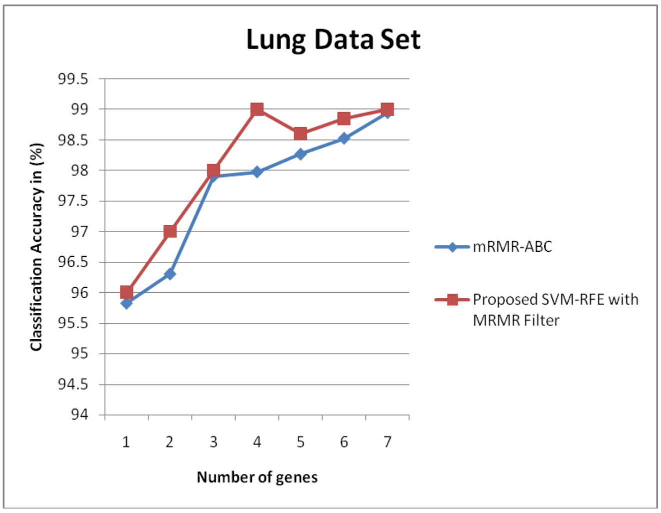


Figure 4: Feature selection results comparison for Lung dataset

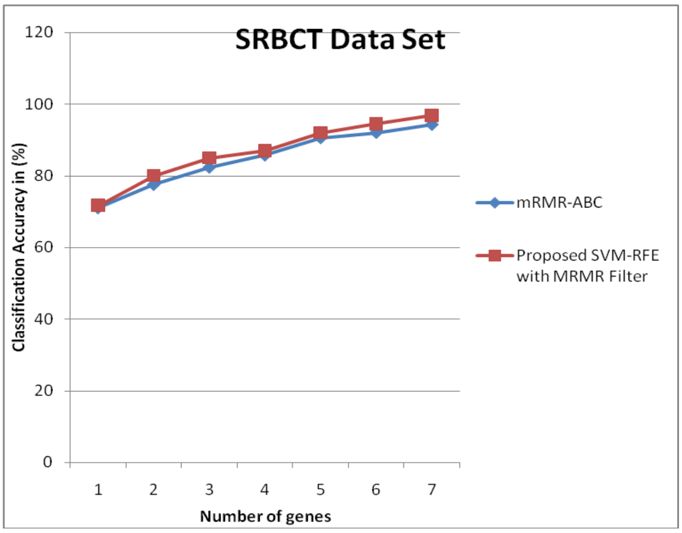


Figure **5: Feature selection results comparison for SRBCT dataset**

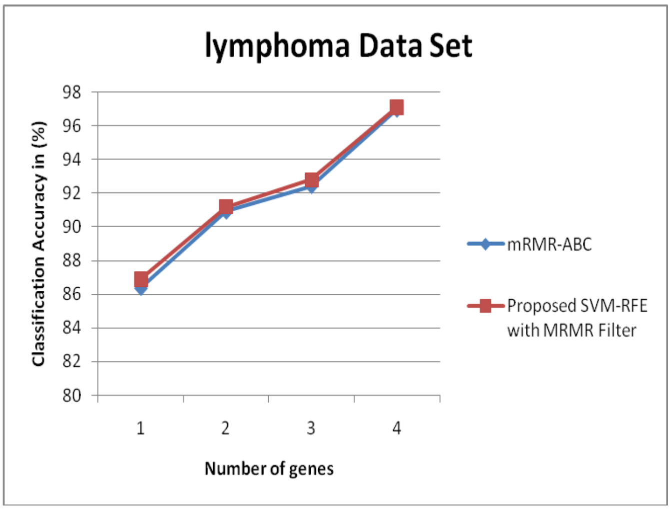


Figure 6: Feature selection results comparison for Lymphoma dataset

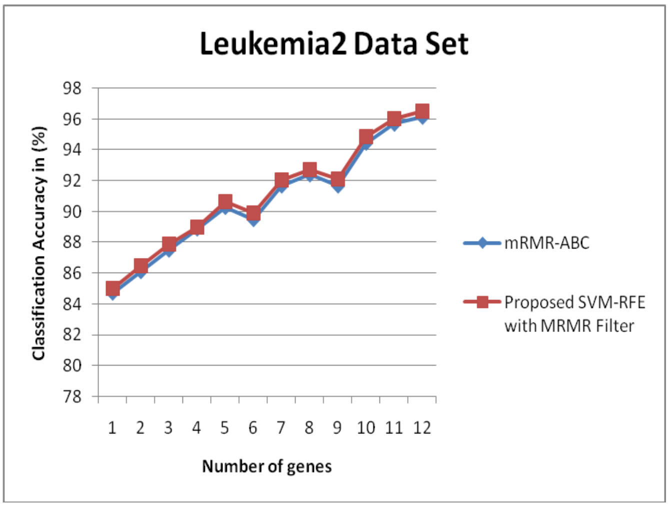


Figure 7: Feature selection results comparison for Leukemia2 dataset

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

V. CONCLUSION

In this research paper, we proposed SVM-RFE with MRMR Filter algorithm for microarray gene expression profile. However, SVM-RFE with MRMR filter selects a fewer number of genes and attempts to eliminate redundant genes some of which may have biological functions important to cancer. Therefore, once a good classification is achieved, the original gene set need to be scanned to find genes that are biologically similar to those selected. A new classification algorithm called hybrid gene selection approach applied by combing RF with SVM called 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 this algorithm has not yet been applied as a gene selection technique for a microarray dataset, so this is the first attempt. Our proposed SVM-RFE with MRMR Filter algorithm is a two-phase method; the MRMR filter technique is adopted to identify the relative and informative gene subset from the candidate microarray dataset. Then the SVM-REF algorithm is employed to select the predictive genes from the 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.

REFERNCES

[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, 2011, pp. 2555– 2558.

[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] A. E, G.-N. J, J. L., and T. E., “Gene selection in cancer classification usingpso/svm and ga/svm hybrid algorithms,” in Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, 2007, pp. 284–290.

[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), 2010 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 - Volume Part III, ser. ICIC’06. Springer-Verlag, 2006, pp. 454–463.

[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] Li, Y., &Ngom, A. (2013). The non-negative matrix factorization toolbox for biological data mining.Source code for biology and medicine, 8(1), 1.

[10] Lingyan Sheng, Pique-Regi, R., Asgharzadeh, S., Ortega, A,” Microarray classification using block diagonal linear discriminant analysis with embedded feature selection, “Acoustics, Speech and Signal Processing, 2009.

[11] Herold, D., Lutter, D., Schachtner, R., Tomé, A. M., Schmitz, G., & Lang, E. W. (2008, August). Comparison of unsupervised and supervised gene selection methods. In Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE (pp. 5212-5215). IEEE.

[12] Daliri, M. R. (2012). A hybrid automatic system for the diagnosis of lung cancer based on genetic algorithm and fuzzy extreme learning machines.Journal of medical systems, 36(2), 1001-1005.

[13] Saraswathi, S., Sundaram, S., Sundararajan, N., Zimmermann, M., &Nilsen-Hamilton, M. (2011). 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, 8(2), 452-463.

[14] Subbulakshmi, C. V., &Deepa, S. N. (2015). Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier. The Scientific World Journal, 2015.

[15] Rong, H. J., Huang, G. B., Sundararajan, N., &Saratchandran, P. (2009). Online sequential fuzzy extreme learning machine for function approximation and classification problems.Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 39(4), 1067-1072.

[16] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” J. Bioinformatics Comput. Biol., vol. 3,pp. 185–205, 2005.

[17] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and minredundancy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[18] I. Guyon, J. Weston, S. Barhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Mach. Learn., vol. 46, pp. 389–422, 2002.

[19] D. Kai-Bo, J. Rajapakse, H. Wang, and F. Azuaje, “Multiple SVM-RFE for gene selection in cancer classification with expression data,” IEEE Trans. Nanobiosci., vol. 4, no. 3, pp. 228–234, Sep. 2005.

[20] Y. Tang, Y.-Q. Zhang, and Z. Huang, “Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis,”IEEE Trans. Comput. Biol. Bioinformatics, vol. 4, no. 3, pp. 365–381, Jul.–Sep. 2007.

[21] Y. Tang, Y.-Q. Zhang, Z. Huang, X. Hu, and Y. Zhao, “Recursive fuzzy granulation for gene subset extraction and cancer classification,” IEEE Trans. Inf. Technol. Biomed., vol. 12, no. 6, pp. 723–730, Nov. 2008.

[22] Adnan Idris, Muhammad Rizwan, Asifullah Khan, Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies, Computers & Electrical Engineering, Volume 38, Issue 6, November 2012, Pages 1808-1819.

[23] U. Alon, N. Barka, D. A. Notterman et al., "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.

[24] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, vol. 286, no. 5439, pp. 531–527, 1999.

[25] D. G. Beer, S. L. R. Kardia, C.-C. Huang et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," Nature Medicine, vol. 8, no. 8, pp. 816–824, 2002.

[26] J. Khan, J. S. Wei, M. Ringner et al., "Classification and ´ diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature Medicine, vol. 7, no. 6, pp. 673–679, 2001.

[27] A. A. Alizadeh, M. B. Elsen, R. E. Davis et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, vol. 403, no. 6769, pp. 503–511, 2000.

[28] S. A. Armstrong, J. E. Staunton, L. B. Silverman et al., "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," Nature Genetics, vol. 30, no. 1, pp. 41–47, 2001.

[29] 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. New Zealand University of Waikato, "Waikato environment for knowledge analysis," http://www.cs.waikato.ac.nz/ml/weka/downloading.html