# Fusion Feature Selection: New Insights into Feature Subset Detection in Biological Data Mining

**Rajangam ATHILAKSHMI[1*], Ramadoss RAJAVEL[1], Shomona Gracia JACOB[2]**

[1] SSN College of Engineering, Department of ECE, Kalavakkam, Chennai 603110, India
athilakshmir@ssn.edu.in (*Corresponding author*), RajavelR@ssn.edu.in

[2] Independent Research Advisor, Computers and Biological Applications,Oman
graciarun@gmail.com

**Abstract:** In DNA microarray research, the increase in gene expression samples and feature dimensions become a challenge for feature selection. This makes it necessary that a more efficient and improved classification algorithm be developed so as to select optimal features in gene expression data. This study presents a new feature selection algorithm that combines the Correlation Feature Selection (CFS) and the Velocity Clamping Particle Swarm Optimization (VCPSO) algorithm. This hybrid model takes advantage of both the filters and the wrappers. It also selects the subsets with optimal features to classify genes by using different classifiers such as Support Vector Machine (SVM), Random Forest(RF),Naïve Bayes(NB) and Decision Tree(DT). Two bioinformatics problems become the basis of evaluation for hybrid mechanisms. These are neurodegenerative brain disorder protein data and microarray cancer data. Reducing the redundancy and finding optimal gene features is the need of the hour. Our experiments show that CFS-VCPSO-SVM selection method eliminates the redundant features and classifies the gene expression data with maximum accuracy.

**Keywords**: Microarray data analysis, Correlated Feature Selection, Velocity Clamping Particle Swarm Optimization, Fusion Feature Selection.

## 1. Introduction

The traditional method of classifying microarray gene expression data involves filter- or pure wrapper-based algorithms. Filters are fast as far as their performance is concerned but show poorer learning results. On the other hand, wrappers guarantee better learning results but they are very slow when applied to high- dimensional datasets. This leads to the fusion of algorithms which incorporates the benefits of the efficient wrapper into the efficient filter for extracting the informative genes. It also improves the classification of gene expression data by enhancing both accuracy and search efficiency of the model (Hsu, Hsieh &Lu, 2008).In this research, we attempted to utilize the benefits of both filters and wrappers by modifying the existing feature selection techniques as detailed in the proposed methodology. As a first step, the feature sets were filtered based on their high correlation to target class and on the lower correlation between features. Then, the reduced feature sets was tuned by using a wrapper procedure named Velocity Clamping Particle Swarm Optimization (VCPSO). The above-mentioned fusion methodology was applied to select the relative features from the neurodegenerative brain disorder protein data and the microarray cancer datasets. Both datasets characterize innumerable gene features which in turn trigger a new need for identifying prognostic (significant) genes that play an important part in detecting the disease earlier. Moreover the existence of irrelevant and redundant gene features in the datasets deteriorates the computing efficiency and also the classification accuracy of machine learning algorithms (Golub et al., 1999; Wu et al., 2012). Therefore it becomes inevitable to reduce the presence of irrelevant and unnecessary genes from the dataset using efficient feature selection methods.

The proposed fusion feature selection strategy namely the CFS-VCPSO algorithm combines the correlation calculation and swarm optimization techniques to reduce the search complexity and overcome the above-mentioned challenges. The performance of the proposed fusion feature selection method was also validated by comparing it with the existing methods on the classification accuracy parameter. Further on, the strength of the proposed algorithm was tested on the selected datasets by using four different classifiers. This paper is organized as follows: literature survey is discussed in section 2, the proposed fusion feature selection methodology is then described in section 3, while the gene expression datasets and parameter settings are defined in section 4. The experimental results are presented in section 5 and the conclusions are set forth in section 6.

## 2. Literature Survey

The main task in feature selection is to identify an optimal set of features for a given problem based on which to construct a classification model. Here, early research concerning the filters

https://doi.org/10.24846/v28i3y201909

and then the wrapper techniques are reviewed. In order to find the ideal subset of features, the filter-based feature selection framework includes both gene-gene mutual information and gene-class mutual information. It shows significant results in classifying high-dimensional data (Hoque, Bhattacharyya & Kalita, 2014).

In 2005, a feature selection framework called Minimum Redundancy Maximum Relevance (MRMR) was proposed. It captures data which considers maximum relevance towards class and minimum redundancy among genes in classifying microarray gene data. Thus, it selects features with improved classification accuracy and a stronger generalization property (Ding & Peng, 2005). Another work on filters applied the CFS technique to a diverse collection of bioinformatics datasets and built the effective classification model that reduces the problem of class inconsistency, high-dimensionality, and information redundancy (Wald, Khoshgoftaar &Napolitano, 2014). (Tan et al., 2008) proposed a hybrid framework for feature selection that binds a genetic algorithm (GA) with different existing feature selection algorithms. The method selects the optimal number of genes from DNA microarray gene expression data. In this context, it was concluded that hybrid approaches are more effective in finding optimal genes due to higher classification accuracy. An effective feature selection scheme based on employing mutual information maximization (MIM) method in improving the classification accuracy of multi-label problems was developed by (Zhang et al., 2009). (Shen, Shi & Kong, 2008) used a hybrid feature selection algorithm integrating tabu search and PSO in order to select an optimal number of genes on microarray datasets but the results were not satisfactory. (Lyu et al., 2017) introduced a fusion filter based on Maximal Information Coefficient (MIC) and Gram-Schmidt Orthogonalization (GSO). This method eliminated irrelevant features while taking two criteria into consideration. First, it improved the degree of relevance between feature variables and target variables. Next, it applied the orthogonalization strategy on the selected candidate feature subsets. The obtained results revealed that modified methods removed irrelevant genes better than the classical method, MRMR.

(Zhou et al., 2006) developed a Mutual Information Rough Set (MIRS) model along with the Binary Particle Swarm Optimization (BPSO) for classification of cancer genes in microarray gene expression data. It was concluded that this model was superior in classifying microarray data when compared to traditional feature selection methods. In 2013 a gene selection algorithm called Rank Weight Feature-selection (RWFS) based on weights was introduced. It assigned weights to attributes selected by means of different feature selection methods followed by ranking. It was concluded that the respective methods were effective in predicting gene features of microarray cancer datasets (Ramani & Jacob, 2013a). (Lu et al., 2017) proposed a fusion algorithm which combined Mutual Information Maximization (MIM) and Adaptive Genetic Algorithm (AGA). The method removed the data redundancies and reduced the size of the data for classification. In 2013 a computational strategy combining different feature selection techniques such as gain ratio, and correlation-based subset evaluator for predicting the class of lung cancer tumor was introduced. It identified a feasible number of features in differentiating small cell lung cancer from non-small cell lung cancer with an improved classification accuracy. The predicted class information with regard to lung cancer tumor could reveal the protein function, the genetic markers for the two diseases thereby enabling the development of possible targets for drug formulation (Ramani & Jacob, 2013b).

The above-mentioned algorithms made use of some filters and/or wrappers for feature selection. Several classical filter-wrapper approaches were employed for classification of microarray gene expression data. However, the pure filter could not guarantee the learning results of the classifier, since it ignores the peculiar heuristics and preferences of the classifier which might lower the classification accuracy. On the other hand, the wrappers detected high-feature dependencies and became computationally inefficient as the search space grew.

This work proposes a novel fusion feature selection strategy which integrates the heuristic measure of CFS (Correlation Feature Selection) algorithm with the searching strategy of VCPSO (Velocity Clamping Particle Swarm Optimization) algorithm. It selects the non-redundant and relevant genes from among the gene expression data and categorizes the respective data more accurately. The mechanism capitalizes on the effectiveness of the filters and the learning accuracy of the wrappers. Moreover, the combination of the efficient filter with a different search strategy is required since it accelerates the learning process of classifiers and stabilizes the classification accuracy. CFS measures the correlation between feature subsets and finds feature subsets that are highly relevant to the class and least relevant to each other while the VCPSO solves the early convergence issue of PSO by clamping the velocity within a search space. The main contribution of this proposed fusion framework is highlighted as follows:

- Novelty: The proposed strategy is a new fusion variant taking both learning results and search space into consideration. It is an efficient computational method which solves both the redundancy problem and the early convergence issues of PSO thereby significantly improving the performance of the algorithm.

- Effectiveness: The proposed hybrid algorithm capitalizes on the advantages of the CFS and VCPSO. The optimal gene set selected by our algorithm improves classification accuracy compared with existing feature selection approaches.

- Robustness: To analyze the performance of the proposed fusion approach, four different classifiers were executed on the selected gene expression data with the selected feature sets.

## 3. Fusion Feature Selection: CFS-VCPSO Selection

### 3.1 Correlated Feature Selection

The correlation-based Feature Selection (CFS) algorithm was used to decrease the search space of the high dimensional datasets. This technique filters feasible feature subsets from a given sample space by including only the mutually uncorrelated features but has greater predictive ability towards a class (Hall,1999). If the features are mutually uncorrelated then redundancy is eliminated whereas the greater relevance of the features with class ensures better prognostic ability.

The CFS criterion is defined as follows:

$$\mathrm{CFS} = \max_{\mathrm{HM_s}}\left[\frac{r_{cf_1} + r_{cf_2} + \ldots + r_{cf_n}}{\sqrt{k + 2(r_{f_1f_2} + \ldots + r_{f_if_j} + \ldots + r_{f_kf_l}}}\right] \quad (1)$$

To measure the worth of a feature subset S consisting of n features, the algorithm uses following merit criteria:

The heuristic merit of a subset,

$$\mathrm{HM_s} = \frac{n\rho_{g,c}}{\sqrt{n} + (n-1)\rho_{g,g}} \quad (2)$$

where $\rho_{g,c}$ is the average value of gene-class correlation. $\rho_{g,g}$ is the average value of gene-gene correlation. The high value of HMs indicates the correlations between the gene features and the class in a better way and lowers the redundancy among the genes in the subset.

### 3.2 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a computational technique based on swarm intelligence that generates a sequence of improved solutions from a given set of solutions (Eberhart & Kennedy, 1995). Velocity Clamping PSO (VCPSO) is a distinct version of PSO which prevents the particle from leaving the search space by clamping its velocity within the search limit. In standard particle swarm optimization, parameters such as swarm size, inertia weight, and neighborhood size play an important part in order to observe the convergence. A number of variations to the standard PSO are developed to enhance the convergence speed and quality solutions discovered by the PSO (Eberhart &Shi, 2000; Zhan et al., 2009; Zhang,2015) This work proposes an additional velocity clamping strategy based on boundary handling mechanisms when the velocity of a particle exceeds the defined boundaries. In PSO, an innumerable particle moves through the search space to obtain a best feasible solution for a particular objective with a lower number of iterations. Before moving the particle to a new position, the algorithm confirms whether the particle lies within the search space. If the maximum velocity of a particle lies within the search space then the particle is allowed to move to the new position, or else the velocity of the particle is set to $V_{max,j}$. This strategy is mathematically represented by the following equation.

$$V_{ij}^{new}(t+1) = \begin{cases} V_{ij}(t+1), V_{ij}(t+1) < V_{max,j} \\ V_{max,j}, \text{Otherwise} \end{cases} \quad (3)$$

If the particle's minimum velocity is smaller than $V_{min}$, then it is clamped to the velocity by the following equation.

$$V_{ij}^{new}(t+1) = \begin{cases} V_{ij}(t+1), V_{ij}(t+1) > V_{min,j} \\ V_{min,j}, \text{Otherwise} \end{cases} \quad (4)$$

The initializations of maximum and minimum velocities are given by equations (5) and (6).

$$V_{max,j} = \lambda(x_{max,j} - x_{min,j}) \tag{5}$$

$$V_{min,j} = \lambda(x_{min,j} - x_{max,j}) \tag{6}$$

where $x_{max,j}$ and $x_{min,j}$ are the maximum and minimum positions of particles in jth dimension obtained from initial test runs of a particle and $\lambda$ is a constant factor between [0,1]. The complete framework of the proposed model CFS-VCPSO is shown in Figure 1 and explained in detail in section 3.3.

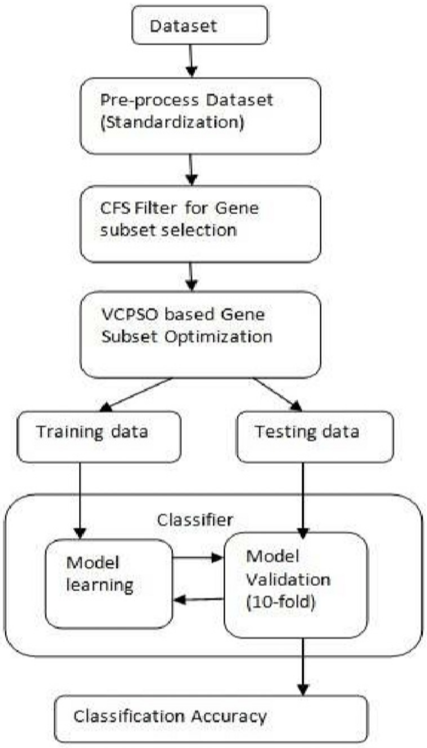
Dataset
Pre-process Dataset (Standardization)
CFS Filter for Gene subset selection
VCPSO based Gene Subset Optimization
Training data
Testing data
Classifier
Model learning
Model Validation (10-fold)
Classification Accuracy

**Figure 1.** Proposed model of CFS-VCPSO selection

## 3.3 CFS-VCPSO-Selection

The selection started with a data pre-processing step in the pre-selection phase, wherein the missing values of gene expressions were replaced by the average values in every dataset. The whole dataset was standardized by using the following equation so as to possess an average value equivalent to zero and standard deviation equal to one.

$$x_{new} = \frac{x - \mu}{\sigma} \tag{7}$$

Here $\mu$ is the average value and $\sigma$ is the standard deviate value for the given data. $x_{new}$ is the new value for a gene expression e. Further on, one applied the multivariate CFS filters that separated the features that are well related to the class, however irrelevant to each other. The reduced feature set $D_R$ was further optimized in the Gene optimization phase using a meta-heuristic wrapper approach Velocity Clamping Particle Swarm Optimization (VCPSO) for which the model achieved maximum classification accuracy. Four different classifiers are used to estimate the predictive ability of the proposed approach in which the stratified 10-fold cross-validation was performed. Specifically, in the Gene optimization phase, the objective is to find an optimal gene subset $\hat{x}$ such that

$$f(\hat{x}) \geq f(x), \forall x \in D_R \tag{8}$$

where 'x' is a dimensional vector of reduced feature set $D_R$ and f(x) is the fitness function. In PSO, a swarm of particles is represented as potential solutions, and each particle is associated with two vectors, i.e., the velocity vector ($V^t$) and position vector ($X^t$) and it is defined as

$$V_m^t = (V_{m1}^t, V_{m2}^t, .., V_{mn}^t) \tag{9}$$

$$X_m^t = (X_{m1}^t, X_{m2}^t, .., X_{mn}^t) \tag{10}$$

The position of the particle with the present fitness value is denoted by '$p_{best}$' and the position of the particle with the best fitness value of the swarm is denoted by '$g_{best}$'. The following formula is used to compute the velocity.

$$V^{t+1} = \omega^t V^t + r_1 c_1 (p_{best} - x) + r_2 c_2 (g_{best} - x) \tag{11}$$

The current velocity $V^{t+1}$ is computed by adding two components to the previous velocity $V^t$ of the particle. The difference between the present fitness '$P_{best}$' and the current position 'x' of the particle constitutes the first component whereas the difference of best fitness value '$g_{best}$' and the current position 'x' of the particle constitutes the second component along with learning constants $c_1$, $c_2$ and random numbers $r_1$,$r_2$. Based on this computed velocity, the forthcoming position of the particle is updated using the following equation.

$$X^{t+1} = X^t + V^{t+1} \tag{12}$$

The movements of each particle are guided by the particle's known influenced position ($p_{best}$) and swarm's best-known position($g_{best}$).In each iteration, every particle reforms its present position and velocity based on the $p_{best}$ and $g_{best}$ value of the particle. Finally, the swarm's new movements are decided based on the particle's reformed position.

The combination of techniques involving correlation calculation and swarm optimization is proposed as the fusion methodology. The key aspect of the model is that it integrates the benefits of fast and efficient dimensionality reduction of the multivariate filter (CFS) and simple yet intelligent Particle Swarm Optimization (PSO) approach. The feature subsets returned by VCPSO were used to train the SVM classifier model. The performance of the model was evaluated by 10-fold cross-validation by giving test data as input. Finally, feature subsets with high classification accuracy and a high ROC value were taken as the optimal genes returned by the model.

## 4. Gene Expression Datasets and Parameter Settings

This research aims at achieving maximum classification accuracy in high dimensional biological datasets. The first dataset is the Common Genes Alzheimer Parkinson (CGAP) based on neuro-degenerative brain-disorder data. The complete description of the data extraction and feature nomenclature is described by (Jacob & Athilakshmi, 2016). The next five datasets include brain tumor, glioblastoma, lung cancer, leukemia and gastric cancer. These datasets were downloaded from the Biolabs Data Set Repository which stores both experimental values and the gene names (Mramor et al., 2007). The proposed fusion feature selection algorithm combined the CFS algorithm with bio-inspired PSO along with the respective modifications (summarized in Algorithm 1). The user does not need to mention any thresholds or number of features to be selected while using the CFS algorithm which is the main benefit of this algorithm. The advantage of using PSO is that it is based on the swarm intelligence and the speed of the search process is very high. The proposed model attempts to select relevant and non-redundant genes by combining the fast computational CFS filter and efficient wrapper VCPSO in a single approach. In the above- mentioned approach, the inertia weight value ‘ω’ falls between 0.9 and 0.4 and it gets decremented at each iteration. Stratified 10-fold-validation was employed in order to analyze the performance of our proposed model. The parameters used in the Velocity Clamping PSO of Gene optimization phase are listed in Table 1. The main characteristics of the gene expression data of CGAP dataset and microarray cancer datasets are illustrated in Table 2.

**Algorithm 1: CFS-VCPSO feature selection:**

```
1: Read the gene attributes g1, g2, g3... gn into an array f[].CFS
2: Pre-process the dataset according to Eq. (7)
3: Apply CFS algorithm to remove the redundant feature.
    a) Find the correlation between the gene attributes in the subset $\rho_{g,g}$
    b) Find the correlation between the gene attributes and the class $\rho_{g,c}$
    c) Measure the worth of a feature subset according to equation (2)
4: Repeat step 3 with different feature subsets.
5: Print the reduced dataset $D_R$ that contains the feature attributes with high class and low feature correlation value based on equation (1).
6: Read $D_R$ and set VCPSO parameters according to table 1.
7:  Initialize swarm population
8:  repeat
9:  for all particles i in the swarm do
10: Evaluate the fitness function $f(X_i)$
11:   if $f(X_i) > f(P_{best (i)})$ then
12: Update swarm’s best position, $P_{best (i)} = X_i$
13:    end if
14: if $f(X_i) > f(g_{best (i)})$ then
15: Update swarm’s global position, $g_{best(i)} = X_i$
16:    end if
17: end for
18: for all particles i in the swarm do
19:    for all features j in the swarm do
20:       $V_{ij}^{t+1} = \omega * V_{ij}^{t} + c_1 r_1 (P_{best(ij)} - x_{ij}) + c_2 r_2 (g_{best(ij)} - x_{ij})$
21: Limit particle velocity $V_{ij}^{t+1}$ according to
    Eq. (3)  and Eq. (4)
22:             if $V_{ij}^{t+1} > V_{max}$ then
23:                 $V_{ij}^{t+1} = V_{max}$
24:             end if
25:             if $V_{ij}^{t+1} < V_{min}$ then
26:                 $V_{ij}^{t+1} = V_{min}$
27:             end if
28: Update next position of particle till fitness
    function converges according to Eq. (12)
29:          end for
30:     end for
31: Until all iterations are not done
32 Train the features returned by PSO using SVM
    classifier.
33 Test the trained model by giving test data to
    the model.
34: return optimal gene set $\hat{x}$ according to Eq. (8)
35: return maximum classification accuracy
    achieved by the model.
```

**Table 1.** Parameters for Velocity Clamping PSO algorithm

| Parameters | Values |
|---|---|
| Swarm strength | 20 |
| Total number of iterations | 20 |
| Inertia,ω | 0.4-0.9 |
| Random numbers(C1,C2) | 2,2 |
| Minimum velocity,Vmin | -6 |
| Maximum velocity,Vmax | 6 |

**Table 2.** Characteristics of gene expression data of six datasets

| Datasets | No. of total genes | Total no. of samples | No. of Classes | References |
|---|---|---|---|---|
| CGAP | 1437 | 111 | 3 | [20] |
| Brain Tumor | 7129 | 40 | 5 | [21] |
| Glioblastoma | 12625 | 50 | 4 | [21] |
| Lung Cancer | 12600 | 203 | 3 | [21] |
| Leukemia | 5147 | 72 | 4 | [21] |
| Gastric Cancer | 4522 | 30 | 3 | [21] |

**Table 3.** Comparison of average number of genes selected for six datasets

| Datasets | ReliefF | MIM | RWFS | CFS-PSO | CFS-VCPSO |
|---|---|---|---|---|---|
| CGAP | 954 | 923 | 54 | 122 | **33** |
| Brain Tumor | 6323 | 6129 | **3** | 199 | 9 |
| Glioblastoma | 6745 | 5625 | **5** | 255 | 31 |
| Lung Cancer | 5401 | 4941 | **3** | 317 | 22 |
| Leukemia | 4203 | 3980 | **6** | 276 | 27 |
| Gastric Cancer | 3412 | 3122 | **4** | 145 | 12 |

**Table 4.** Comparison of mean classification accuracy obtained for six datasets

| Datasets | ReliefF | MIM | RWFS | CFS-PSO | CFS-VCPSO |
|---|---|---|---|---|---|
| CGAP | 55.2 | 66.7 | 71.9 | 91.5 | **94.9** |
| Brain Tumor | 74.3 | 79.3 | 77.5 | **98.1** | 98 |
| Glioblastoma | 66.1 | 68.3 | 90 | 92.2 | **98** |
| Lung Cancer | 74.3 | 79.3 | 94.1 | 98.3 | **99.1** |
| Leukemia | 71.6 | 77.7 | 65 | 92.9 | **98.3** |
| Gastric Cancer | 80.3 | 74.2 | 93.3 | 97.7 | **99** |

**Table 5.** Comparison of ROC obtained for six datasets

| Datasets | ReliefF | MIM | RWFS | CFS-PSO | CFS-VCPSO |
|---|---|---|---|---|---|
| CGAP | 0.64 | 0.82 | 0.86 | 0.89 | **0.94** |
| Brain Tumor | 0.68 | 0.79 | 0.87 | 0.91 | **0.98** |
| Glioblastoma | 0.71 | 0.78 | 0.88 | 0.90 | **0.97** |
| Lung Cancer | 0.79 | 0.81 | 0.82 | 0.87 | **0.99** |
| Leukemia | 0.51 | 0.77 | 0.81 | 0.88 | **0.99** |
| Gastric Cancer | 0.76 | 0.79 | 0.83 | 0.92 | **0.98** |

**Table 6.** Performance of classifiers on selected gene subset of CFS-VCPSO selection

| Datasets | Support Vector Machine(%) | Decision Tree (%) | Naive-Bayes (%) | Random Forest (%) |
|---|---|---|---|---|
| CGAP | **94.9** | 78.4 | 75.7 | **94.9** |
| Brain Tumor | **98** | 95 | 95 | 95 |
| Glioblastoma | **98** | 94 | 94 | 96 |
| Lung Cancer | **99.1** | 97.1 | 97.1 | 91.2 |
| Leukemia | **98.3** | 87.5 | **98.3** | 94.4 |
| Gastric Cancer | **99** | 96.7 | 75.7 | 78.4 |

## 5. Experimental Results

The results obtained on neurodegenerative brain disorder data and microarray cancer datasets with four other algorithms namely ReliefF (Kononenko,1994), MIM-Mutual Information Maximization (Lu et al.,2017), RWFS-RankWeightFeatureSelection (Ramani & Jacob,2013a) and CFS-PSO (Yang et al.,2008) evaluate the effectiveness of our proposed fusion methodology for feature selection, on the basis of classification accuracy and the number of genes selected. The metrics used for measuring the performance of the classifier were classification accuracy and Receiver Operating Characteristic (ROC) curve. SVM was applied with the same parameter setting for the selected gene subsets of all the five algorithms. The parameters for the SVM execution set are as follows: the penalty coefficient and the gamma value were set at 0.12 and 0.13; while the kernel function used was sigmoid. For all datasets, on average, the proposed method CFS-VCPSO and RWFS selected a small number of features and it is shown in Table 3. Although the number of features selected by RWFS is small compared to our proposed method, it does not yield sensible classification accuracy. The classification accuracy rates are shown in Table 4. For two datasets (Gastric Cancer and Lung cancer) the proposed method managed to improve classification accuracy to nearly 100%. The next high classification accuracy ~98% was noticed for Brain tumor, Glioblastoma and Leukemia datasets. For CGAP dataset, the estimated classification accuracy of the proposed method was 94.9%. The next significant performance for all six datasets is reported by CFS-PSO. It is obvious from the Table 4, that the proposed method is superior to all feature-selection methods.

Next, the performance of the selected feature subset of all five feature selection algorithms is evaluated by the second metric ROC for all datasets. The curve plots a false positive rate on the x-axis, the true positive rate on the y-axis and the diagonal line corresponds to the random classifier. For the CGAP dataset, the estimated ROC by our proposed method was 0.94. For brain tumor and gastric cancer dataset, it was 0.98. For lung cancer and leukemia it was 0.99 and for glioblastoma it was 0.97 as it can be seen in Table 5. The results of ROC curve of five feature selection methods based on the 10-fold cross-validation for all datasets are shown in Figures 2,3,4,5,6, and 7 respectively. From the plot, it can be noticed that the CFS-VCPSO method results in a higher running curve and a larger ROC value for all datasets when compared to all feature selection methods. This means the feature subsets selected based on the proposed approach have the best classification performance among all five feature selection methods. The fitness of the selected gene subset of CFS-VCPSO selection is further evaluated by three classification algorithms Random Forest (RF), Naïve Bayes (NB) and Decision Tree (DT).The SVM classifier showed the best performance for all datasets and achieved the highest accuracy that is 99% for Lung Cancer and Gastric cancer datasets as shown in Table 6. Next to SVM, the RF classifier showed an equivalent optimal performance in CGAP data while the NB classifier showed an equivalent best performance for the Leukemia Dataset. From this experiment, we came to the conclusion that the SVM is the most suitable classifier for the CFS-VCPSO-Selection algorithm. In brief, all four classifiers reach classification accuracy rates higher than 90% for all datasets, which demonstrates the efficiency of CFS-VCPSO selection method.

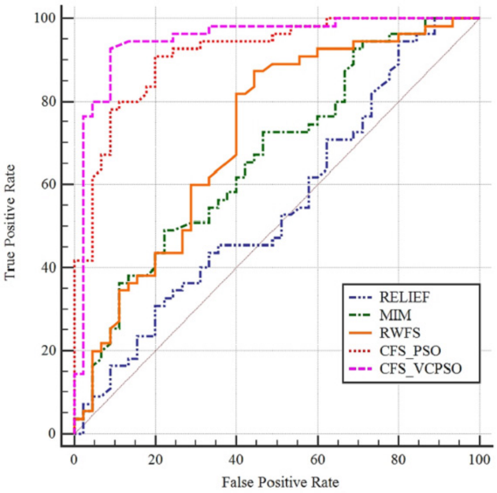

**Figure 2.** CGAP data

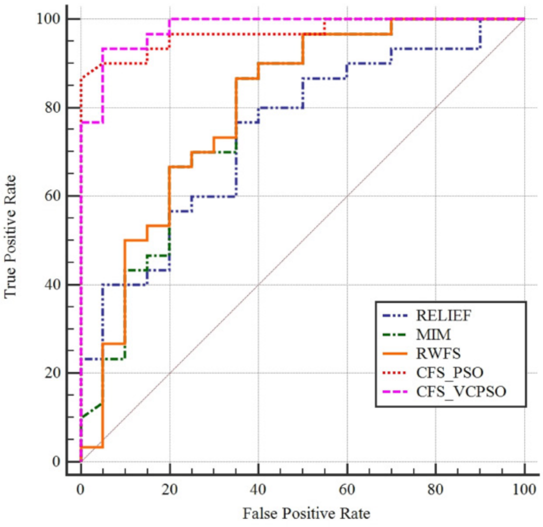

**Figure 3.** Brain Tumor

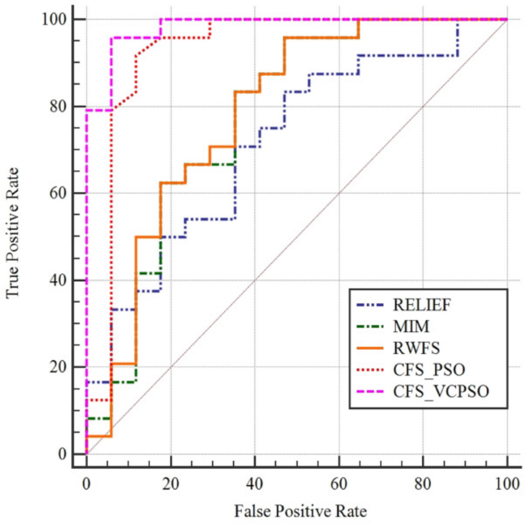

**Figure 4.** Glioblastoma

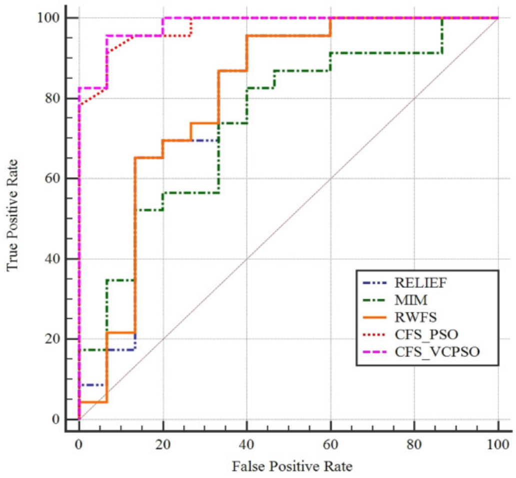

**Figure 5.** Leukemia

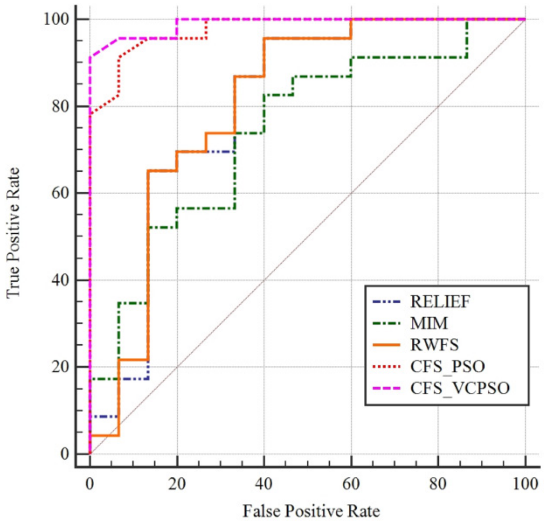

**Figure 6.** Lung Cancer

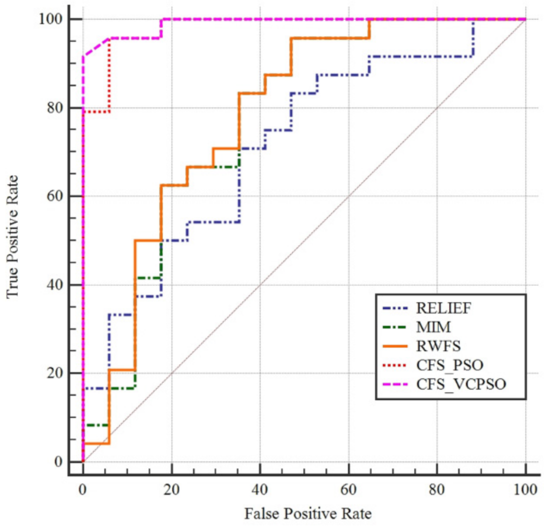

**Figure 7.** Gastric Cancer

## Acknowledgements

This research work was carried out as part of the project funded by Science and Engineering Research Board (SERB), Department of Science and Technology (DST) funded project under Young Scientist Scheme – Early Start-up Research Grant- titled “Investigation on the effect of Gene and Protein Mutants in the onset of Neurodegenerative Brain Disorders (Alzheimer’s and Parkinson’s disease): A Computational Study” with Reference No- SERB – YSS/2015/000737.

# REFERENCES

1. Ding, C. & Peng, H. (2005). Minimum Redundancy Feature Selection from Microarray Gene Expression Data, *Journal of Bioinformatics and Computational Biology*, *3*(2),185-205.

2. Eberhart, R. C. & Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Proceedings of International Symposium on Micro Machine and Human Science*, Nagoya, Japan (pp. 39-43).

3. Eberhart, R. C. & Shi, Y. (2000). Comparing inertia weights and constriction factors in particle swarm optimization. In *Proceedings of International Conference on Evolutionary Computation*, La Jolla, CA, USA (pp. 84-88).

4. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, *286*(5439), 531-537.

5. Hall, M. A. (1999). *Correlation-based feature selection for machine learning.* Doctoral dissertation. The University of Waikato.

6. Hoque, N., Bhattacharyya, D. K. & Kalita, J. K. (2014). MIFS-ND: a mutual information-based feature selection method, *Expert Systems Applications*, *41*(14), 6371-6385.

7. Hsu, H. H., Hsieh, C. W. & Lu, M. D. (2011). Hybrid feature selection by combining filters and wrappers, *Expert Systems Applications*, *38*(7), 8144-8150.

8. Jacob, S. G. & Athilakshmi, R. (2016). Extraction of Protein Sequence features for Prediction of Neuro-degenerative Brain Disorders: Pioneering the CGAP database. In *Proceedings of the International Conference on Informatics and Analytics*, Pondicherry, India (p. 30).

9. Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In *Proceedings of European Conference on Machine Learning*, Springer, Berlin, Heidelberg (pp. 171-182).

10. Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y. & Gao, Z. (2017). A hybrid feature selection algorithm for gene expression data classification, *Neurocomputing*, *256*(c), 56-62.

11. Lyu, H., Wan, M., Han, J., Liu, R. & Wang, C. (2017). A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining, *Computers in Biology and Medicine*, *89*, 264-274.

12. Mramor, M., Leban, G., Demsar, J. & Zupan, B. (2007). Visualization-based cancer microarray data classification analysis, *Bioinformatics, 23*(16), 2147-2154.

13. Ramani, R. G. & Jacob, S. G. (2013). Benchmarking Classification Models for Cancer Prediction from Gene Expression Data: A Novel Approach and New Findings, *Studies in Informatics and Control, 22*(2), 133-142. DOI: 10.24846/v22i2y201303

14. Ramani, R. G. & Jacob, S. G. (2013). Improved classification of lung cancer tumors based on structural and physicochemical properties of proteins using data mining models, *PLoS One*, *8*(3), e58772.

15. Shen, Q., Shi, W. M. & Kong, W. (2008). Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data, *Computational Biology and Chemistry, 32*(1), 53-60.

16. Tan, F., Fu, X., Zhang, Y. & Bourgeois, A. G. (2008). A genetic algorithm-based method for feature subset selection, *Soft Computing*, *12*(2)*,* 111-120.

17. Wald, R., Khoshgoftaar, T. M. & Napolitano, A. (2014). Using Correlation-Based Feature Selection for a diverse collection of Bioinformatics Datasets. In *Proceedings of International Conference on Bioinformatics and Bioengineering*, Boca Raton, FL, USA (pp. 156-162).

18. Wu, M. Y., Dai, D. Q., Shi, Y., Yan, H. & Zhang, X. F. (2012). Biomarker identification and cancer classification based on microarray data using Laplace Naive Bayes model with mean shrinkage, *IEEE/ACM Transactions on*

*Computational Biology and Bioinformatics*, *9*(6), 1649–1662.

19. Yang, C. S., Chuang, L. Y., Ke, C. H. & Yang, C. H. (2008). A Hybrid Feature Selection Method for Microarray Classification, *IAENG International Journal of Computing*, *35*(3), 1-3.

20. Zhan, Z. H., Zhang, J., Li, Y. & Chung, H. S. H. (2009). Adaptive particle swarm optimization, *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, *39*(6), 1362-1381.

21. Zhang, M. L., Peña, J. M. & Robles, V. (2009). Feature selection for multi-label naive Bayes classification, *Information Sciences*, *179*(9), 3218-3229.

22. Zhang, Y. (2015). A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications, *Mathematical Problems in Engineering*, *2015*(1), 1-38. Article ID 931256.

23. Zhou, W., Zhou, C., Zhu, H., Liu, G. & Chang, X. (2006). Feature Selection for Microarray Data Analysis Using Mutual Information and Rough Set Theory. In *Proceedings of International Conference Intelligent Computing Computational Intelligence and Bioinformatics*, Berlin, Heidelberg (pp 424-432).