https://doi.org/10.48047/AFJBS.6.14.2024.8754-8761

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

EFFICIENT CLASSIFIER ALGORITHM FOR GENE EXPRESSION DATA

C.KondalRaj, Assistant Professor, R.Murugesan Associate Professor

Email:kondalrajc@gmail.com,CPA college,Bodinayakanur,Theni-Dt,Tamilnadu,India

Volume 6, Issue 14, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 15 Aug 2024

doi: 10.48047/AFJBS.6.14.2024.8754-8761

ABSTRACT — This study presents a robust method for uncovering hidden patterns within data. Initially, the preprocessing phase eliminates irrelevant data to reduce computation time. Following this, feature selection is conducted using the Principal Component Analysis (PCA) method to identify the most pertinent features. Once optimal features are selected, the classification process is carried out. Two distinct feature extraction methods are applied to microarray gene expression data analysis: ranking-based and set-based feature extractions. To enhance extraction performance, a Multi Algorithm Fusion (MAF) method is introduced. Subsequently, the Polynomial Support Vector Machine (PSVM) integrated with MAF is employed for feature extraction. The absolute weights of SVM, fisher ratio, and PSVM are determined, and an MAF-based PSVM algorithm is utilized to achieve efficient results. The proposed MAF-PSVM method demonstrates superior feature robustness compared to traditional approaches, resulting in a significantly lower classification error. Finally, a Random Forest (RF) classifier is utilized for efficient classification, validated by metrics such as error rate, accuracy, specificity, precision, recall, F1 score, and false positive rate.

Keywords: Principal Component Analysis (PCA), Random Forest Classifier (RF), Polynomial Support Vector Machine (PSVM), Multi Algorithm Fusion (MAF).

I. INTRODUCTION

The process of accumulating and excavating information from enormous amounts of data to discover remarkable patterns and their correlation with extracted data is known as data mining. It entails smart technologies and amenability to investigate the probability of concealed information inside the data. Nowadays, the tangible world data are immense and of poor quality. Mining such data will not provide suitable and expedient results because those data are vulnerable to noise. Occasionally, it is observed that some fields are partly vanished. Hence, pre-processing is crucial before data mining which involves cleaning, integrating, normalizing and reducing of data. There are different data sets from which the data are exhumed. Feature selection is a method of finding the most pertinent features from the data set and signifying the high dimensional data to a smaller extent. The chosen features are extricated and categorized for various analyses. The basic steps involved in data mining are illustrated in the below figure [1].

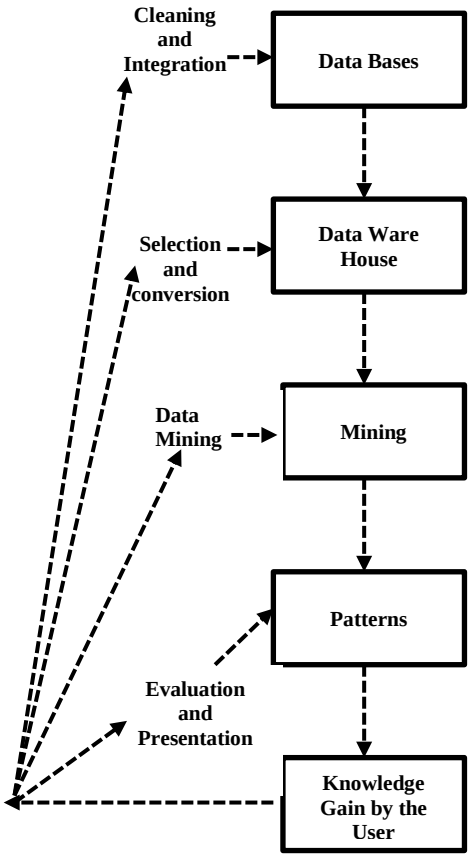

Fig. 1 Steps involved in data mining

The proposed work ponders on an important concept named as Gene selection. It plays a significant role in colon cancer data classification to lessen the number of noisy and inappropriate genes and to choose the allied genes for enhancing the classification results.

The main objectives of this work are:

- To improve feature selection process using PCA method.
- To extract the features by utilizing MAF-PSVM approach.
- To classify genes by means of Random forest classifier.

This paper is structured as follows:

Section I provides the general introduction of data mining, feature selection and gene classification.

Section II reviews the works that are closely related to gene expression data mining.

Section III describes the proposed work of colon and cancer dataset that utilizes data mining algorithms for extraction, selection and classification.

Section IV deals with the performance analysis of various factors such as classification error, gene classification accuracy and feature robustness. And finally,

Section V concludes the paper along with the references to future work.

II. RELATED WORKS

The gene expression data analysis is an important aspect for interpreting diseases and plays vital role in understanding diseases and discerning medications for them. At all times, gene expression data was prone to high dimensionality. Due to this concern, feature selection became the essential mechanism for classifying cancer. [2]J.c.Ang suggested a Semi-Supervised SVM based Feature Selection (S^3VM-FS) that utilized information from labelled and un-labelled gene expression data. The lung cancer data were experimented and it was observed that this method attained a maximum accuracy but the processing time was longer than the familiar methods like SVM-RFE and S^3VM-RFE.

The cancer clusters were demarcated and patient genes were classified for predicting cancer by means of micro-array systems. [3]J.X.Liu offered an innovative technique for organizing the tumour samples from gene expression data based on Robust Principal Component Analysis (RPCA). Initially, the RPCA emphasized typical genes related to a distinctive biological method. After that, RPCA and RPCA combined with Linear Discriminant Analysis (LDA) were employed to ascertain their characteristics. As a final point, Support Vector Machine (SVM) was pertained to categorize the samples based on the applied to classify the tumor samples of gene expression data based on the realized features. The techniques were efficient and viable in tumour grouping. Robust feature selection detected the appropriate genes for coherent sample classification. The prevailing selection methods were affluent and lay-off in gene expression data resulted in poor accuracy that depraved multi-class classification. [4]J.Bennet introduced an ensemble feature selection incorporated with RFE and BBF for selecting genes and utilized SVM algorithm for classifying them. The ensemble method relented an improved performance than the other

classifiers and afforded a novel perception in feature selection.[5]T.Latkowski presented the data mining algorithms to distinguish the substantial genes and their structures from micro-array gene expression data set of autism. Certain feature selection techniques were chosen and their respective outcomes were assimilated. The numerical results of gene selection and classification were deliberated from which a synthesis of different selection methodologies concomitant with autism. An exceptional process to calculate the number of top-ranked genes that were availed in classification was also included.

Bi-clustering procedures evaluated the gene expression data proficiently by realizing a gene group having stabilized expression configurations with definite conditions. [6]Z.Wang initiated an uncomplicated method in which the bi-clusters were found from large, noisy and intricate data. Longest Common Subsequence (LCS) configuration nominated row duos in a key matrix obtained from a data matrix to uncover the kernel of each bi-cluster. Several bi-clustering algorithms were verified on gene expression dataset and on comparison, it was proven that the UniBic algorithm outclassed all other earlier processes in discovering the constricted bi-clusters.

The surviving Rotation Forest Algorithm (RFA) based methods concentrated only on the accuracy of gene classification and ignored the cost of classification. [7] H.Lu recommended cost-intuitive RFA for arranging gene expression data. The types of classification costs such as misclassification, test and rejection were analysed in this work. The tentative results confirmed that the proposed algorithm minimized the classification cost efficiently.[8]S.Cogill proposed a classification strategy to recognize the autism related disorders making use of expression patterns in relation to brain development. The acquaintance of familiar protein-coding genes was swapped with all gene varieties. The method exactly classified and ordered the genes in terms of risks associated with autism. This gene classification was reliable than the former techniques.[9]H.Lu put forward a composite feature selection algorithm by merging AGA and MIM. The investigational fallouts indicated the supremacy of MIM-AGA selection in decreasing the facet of gene data and eliminating classification redundancy. The compact dataset achieved greatest accuracy when compared to the traditional selection algorithms. Four classifiers were smeared with abridged dataset to exhibit the vigour of projected scheme.A huge quantity of gene expression data were collected from biological experimentations by applying the microarray technology. The collected data were then explored successfully with the help of some clustering algorithms. In the next step, co clustering was employed for recovering the co-clusters present in the gene expression data and with this the group were categorized further into sub groups. Subspace Weighting Co-Clustering (SWCC) was proposed for the purpose of co-clustering the high dimensional gene expression data [10].X.chen Then a novel co-clustering function was framed with the objective of recovering the co-clusters where the subspace weight matrix was employed. The work [11] focused on reducing the dimensioning of features present in the categorization by proposing a novel strategy of weighted feature selection that was based on bacterial algorithm. It differentiates the features and classifies them based on their performances and occurrence frequency of population. The efficiency of the proposed work was analyzed by 4 bacterial dependent techniques. In addition to that, 3 famous algorithm that depends on population which utilizes 15 different datasets of micro-array that possess unique classes and features. As a result, the embedded and weighted feature selection strategies have enhanced the capacity of feature selection of some bacterial algorithms.

III. PROPOSED WORK

The proposed work in this section deliberates the methodologies involved in the proposed work and the flow diagram as well. The intelligent technologies are required for exploring the likelihood of hidden knowledge existing over the data. Initially, preprocessing step is utilized for eliminating the unwanted data. There may be a noise presence in some of the data, so preprocessing step is essentially required to remove all the unwanted data. The elimination of unwanted data processing has been accomplished with preprocessing techniques, which also reduces computation time. Feature selection is the significant step that has to be concerned more for efficient gene selection process. The selection of appropriate features are the essential step which is performed using an operative PCA approach. After completing the selection process, best features has to be extracted for undertaking classification process. MethodsMicroarray gene expression data analysis In this study, we consider two prominent feature extraction methods of ranking-based and set-based subspaces. In this term based ranking approach then features are ranked depending upon the single term not considering to any community. inter relationship among the features. The set-based feature extraction is the feature extraction process, where the dependency among the features present in feature set is considered. The robustness is the significant problem often occurred from the generalization ability of feature extraction in learning methods. To improve the extraction efficiency, a multi algorithm fusion technique has been implemented. The polynomial support vector machine related to multi-algorithm fusion is considered for the feature extraction process. The absolute weight of SVM, fisher ratio and Polynomial Support Vector Machine (PSVM) were utilized in this fusion approach. An effective feature extraction is accomplished with this multi algorithm fusion technique. The multi algorithm fusion based PSVM algorithm is chosen for the efficient outcome.

A. DATA PREPROCESSING

In this preprocessing stage, an unwanted data is removed from the preferred dataset. The removal of noise and white space is achieved and the null values are also removed to retrieve an effective significant information from the dataset.

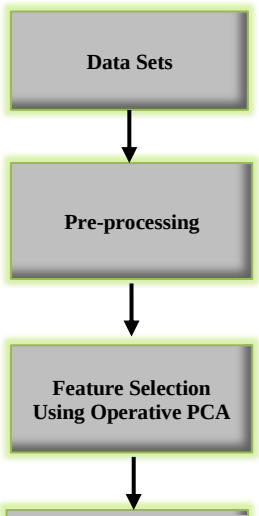

Figure 1: Proposed methodology

B. Feature Selection using Operative PCA approach

In this gene expression approach, a Principal Component Analysis (PCA) based dimensionality reduction approach is proposed. In this operative PCA approach the information is plotted linearly to the lower-dimensional space in such a manner the maximum data variance is caught over the low-dimensional depiction. Generally, the eigenvector matrix is being computed through generating the correlation or the co-variance data matrix. From this the reconstruction of large amount of original data variance is completed which is corresponding to higher eigenvalues. Additionally, certain initial eigenvectors can frequently be understood in relations of the significant physical performance of the system. The new space is abridged together with the lost data where the most significant variance traversed by some eigenvectors are retained. The following algorithm is utilized for the creating input data for PCA.

Algorithm 1: Operative PCA based feature selection	
Input: PCA Input Data	
Output: Feature selected Attributes	
Step	***Description***
Input	*Prioritized cluster*
Output	*Classified Output*
Procedure	
1	*Create Random Forest*
1.1	*Initialize inter as the number of seeds*
1.2	*Initialize numTree as the number of trees*
1.3	*Set data as the content of the seed*
1.4	*Set tdata as the test data*
1.5	*Initialize tree as an array of size numTree*
1.6	*Set update to 100/numTree*
1.7	*Start the timer with start = StartTimer()*
1.8	*Calculate k as log(data.length-1) / log(2) + 1*
1.9	*Initialize done with estimateOOB()*
2	*Create StartTimer()*
2.1	*Initialize treePool with executor of NumThreads*
2.2	*For loop from 0 to numTrees*
2.3	*Create trees in treePool using data and tdata*
2.4	*End For loop*
2.5	*Call TestForest()*
3	*Create TestForest*
3.1	*Initialize R as the random values for the trees*
3.2	*Initialize correctness to 0*
3.3	*For loop from 0 to tdata.size*
3.4	*For loop from 0 to tree.size*
3.5	*Add values to R*
3.6	*End inner For loop*
3.7	*Calculate pred as ModeOf(R)*
3.8	*If pred == R, increment correctness by 1*
3.9	*End If and outer For loop*
4	*Create ModeOf function*
4.1	*Initialize max and maxclass to 1*
4.2	*For loop from 0 to Tree. Oredict. size as i*
4.3	*Initialize count to 0*
4.4	*For loop from 0 to Tree. Predict. size as j*
4.5	*If i == j, increment count by 1*
4.6	*End If*
4.7	*If count > max, update maxclass to i and max to count*
4.8	*End If, inner For loop, outer For loop*
5	*Create estimateOOB function*
5.1	*If estimateOOB(R) == null*
5.2	*Initialize map to C*
5.3	*Call estimateOOB(R, map)*
5.4	*End If*
Novelty	*Integrate R values from our values*
6	*Initialize str as "File name"*
6.1	*For loop from 0 to str.size as IN*
6.2	*if the length of str.get(IN) is less than categ, the loop continues.*
6.3	*Else, set categ to str.get(IN)[str.get(IN).length]*
6.4	*End If, For loop*
Random Forest Function (RaF)	
7	*Define function RaF*
7.1	*Set RaF.C to categ*
7.2	*Set RaF.M to Input.get(0).length*
7.3	*Calculate RaF.Ms as Math.log(RaF.M)/Math.log(2)+1*
7.4	*Start RaF with RaF.Start()*
7.5	*End function*

C. Feature Extraction using Multi Algorithm Fusion approach

The steps involved in the implementation of MAF-PSVM are announced in detail. The output of the PCA based feature extraction is given as an input for feature extraction process. The feature extraction algorithm such as Fisher Ratio, AW-SVM and

PSVM are fused together, however selection of the kernel function is the challenging task for this fused feature extraction process. Different studies have exposed that the linear kernel function is an appropriate for the linearly separate data, therefore polynomial kernel was designated. In future, other kernel function will be studied. The polynomial kernel functions have different generalization ability. The optimum classifier performance is reached in many cases when parameter d has the range of 1. Thus, polynomial kernel order value d is engaged as 1 and feature extraction has been conducted for clustered samples. The fusion method content utilized in this approach is introduced below,

The score related multi algorithm is utilized in this study. Initially, all the feature scores were produced from the score vector with the each basis criteria. Next, the multiple score vectors were aggregated as single consent score vector using the score combination algorithm. Lastly, ranking is done based on consensus scores through employing feature ranking procedure. The score-based multi-algorithm fusion procedure is exemplified in Fig 2.

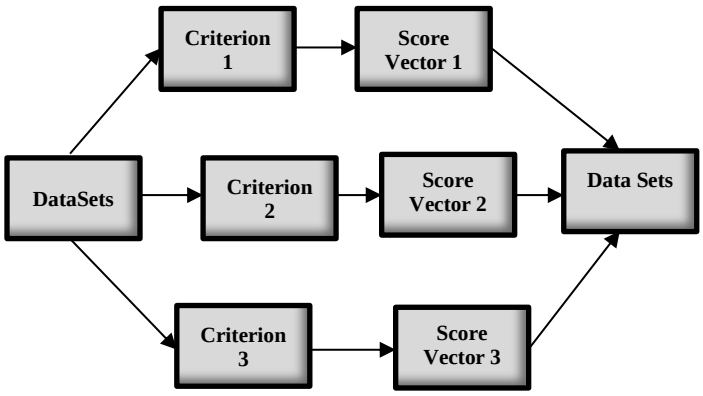

Figure 2: Score based multi algorithm

At distinct criteria, scores are produced under score aggregating will be analogous. The score normalization should be undertaken firstly then the score combination is performed, the normalized range is [0, 1] in this approach. Let's take u_i as for the score vector of criteria basis i, we perform normalization as follows:

$$u_i' = \frac{u_i - u_{i\,min}}{u_{i\,max} - u_{i\,min}} \qquad (1)$$

Where $u_{i\,min}$ and $u_{i\,max}$ are min and max values of vector u . The larger score in all the basis criteria achieves better features using the formula which is given below

$$u = \frac{1}{m}\sum_{i=1}^{m} u_i' \qquad (2)$$

Where m is the basis criterion in fusion.

D. EFFECTIVE RANDOM FOREST BOOTSTRAP ALGORITHM:

Randomized trees are constructed for classification in random forest algorithm. In creating the COST, one must consider which nodes are to be split and on what variable is that node being best selected for splitting over its branches to become a pure internal chi.d. ClassAlgo -> A nod impurity is calculated using Gini index of classification algorithm. Let very node s and q(c/s) be the evaluated class probabilities, where q(c/s) C=1…C. Where, Gini index of node s is defined as

$$N(s) = \sum_{c1 \neq c2} q\left(\frac{c1}{s}\right) q\left(\frac{c2}{s}\right) = 1 - \sum_{c=1}^{c} q^2\,(c/s)$$

Let ooo be the splitting point of node sss, where the node is divided into two sections. In this process, a proportion q_R of the samples are assigned to s_R, and a proportion q_L of the samples are assigned to s_L i.e. $s_R + s_l = 1$. Therefore the decrease in Gini index is shown below.

$$\Delta N(o,t) = N(s) - q_R\, N\, s_R - q_l N s_l$$

This part is written to find optimal splitting point s* on the predictore variable j*, so that it gives maximum decline in Gini impurity as shown below.

$$o^*, j^* = \arg max_{o,j} \Delta N(o,s)$$

The function mentioned above is repeatedly invoked by a classification algorithm to develop a decision tree. A random forest consists of multiple decision trees working together, utilizing bagging for classification trees and random subspace methods. When fitting a gradient boosting model, we employ ensembles of decision trees using the Classification and Regression Trees (CART) algorithm. This involves multiple bootstrap samples and random subsets of variables being included in each individual tree, except at the leaf nodes. The process of growing trees in a forest continues until pure leaf nodes are achieved.

RF algorithm: RF alogrithm contains three functions - 1) bootstrap sample 2) Tree Generation and3) Randomized Tree Learn. For this purpose, A random forest bootstrap algorithm is suggested that makes efficient gene classification both high accuracy and fast computation time. While generating the performance, these bootstrap samples are drawn in order to reduce the variance and increase how well our model actually performs (i.e., play around with bootstraps). The main strength of this work is the reduction across a greater number of irrelevant and redundant attributes effectively through 1) bootstrap sample selection, impurity based split criterion resulting in sub-sampling between different nodes/ branches & therefore reducing the overfitting issue; along with that 2) randomized tree learning approach which also reduces classifier learning time without overhead.

Classification:

After extracting the features, the classification technique is employed to classify the breast cancer diseases. In this work, for classification, Random forest classification is used. The novel in this algorithm is mentioned in the

Step	**Description**
Input	Prioritized cluster
Output	Classified Output
Procedure	
1	Create Random Forest
1.1	Initialize inter as the number of seeds
1.2	Initialize numTree as the number of trees
1.3	Set data as the content of the seed
1.4	Set tdata as the test data
1.5	Initialize tree as an array of size numTree
1.6	Set update to 100/numTree
1.7	Start the timer with start = StartTimer()
1.8	Calculate k as log(data.length-1) / log(2) + 1
1.9	Initialize done with estimateOOB()
2	Create StartTimer()
2.1	Initialize treePool with executor of NumThreads
2.2	For loop from 0 to numTrees
2.3	Create trees in treePool using data and tdata
2.4	End For loop
2.5	Call TestForest()
3	Create TestForest
3.1	Initialize R as the random values for the trees
3.2	Initialize correctness to 0
3.3	For loop from 0 to tdata.size
3.4	For loop from 0 to tree.size
3.5	Add values to R
3.6	End inner For loop
3.7	Calculate pred as ModeOf(R)
3.8	If pred == R, increment correctness by 1
3.9	End If and outer For loop
4	Create ModeOf function
4.1	Initialize max and maxclass to 1
4.2	For loop from 0 to tree.predict.size as i
4.3	Initialize count to 0
4.4	For loop from 0 to tree.predict.size as j
4.5	If i == j, increment count by 1
4.6	End If
4.7	If count > max, update maxclass to i and max to count
4.8	End If, inner For loop, outer For loop
5	Create estimateOOB function
5.1	If estimateOOB(R) == null
5.2	Initialize map to C
5.3	Call estimateOOB(R, map)
5.4	End If
Novelty	Integrate R values from our values
6	Initialize str as "File name"
6.1	For loop from 0 to str.size as IN
6.2	If str.get(IN)[str.get(IN).length] < categ, continue
6.3	Else, set categ to str.get(IN)[str.get(IN).length]
6.4	End If, For loop
Random Forest Function (RaF)	
7	Define function RaF
7.1	Set RaF.C to categ
7.2	Set RaF.M to Input.get(0).length
7.3	Calculate RaF.Ms as Math.log(RaF.M)/Math.log(2)+1
7.4	Start RaF with RaF.Start()
7.5	End function

Thus the Classification accuracy is achieved using Novel Random Forest Algorithm. Finally, the performance evaluation was made and the performance metrics were estimated to prove the effectiveness of this proposed system.

IV. RESULTS AND DISCUSSION

The proposed method is examined worh colon cancer dataset to calculate as an efficiancy of the methodology The dataset link - http://biogps.org/dataset/tag/colon%20cancer/.

This section outlines the performance analysis of proposed work. We calculate the performance measures (error rate, accuracy, specificity, precision, recall, f1-score and False positive rate etc.) for our proposed PCA based random forest classifier.

TABLE: 1 Performance Measures

PERFORMANCE MEASURES	VALUES
Error Rate	0.011461318
Accuracy	0.938538682
Specificity	0.971223022
Precision	0.981308411
Recall	1
F1-Score	0.990566038
False Positive Rate	0.028776978
False Negative Rate	0
Negative Predict Value	1
Mathews Correlation Coefficient	0.976252693

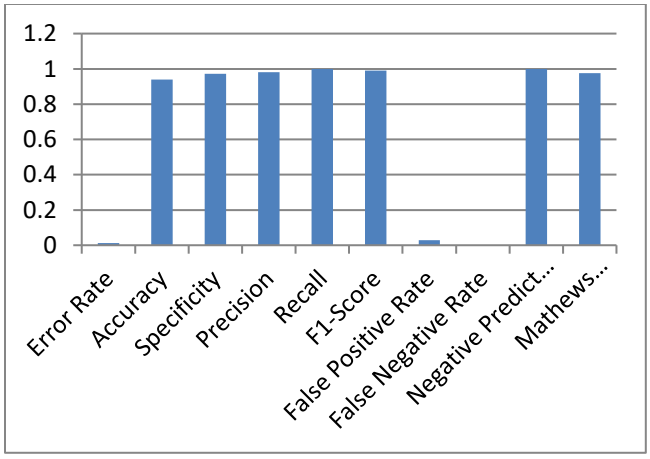

Figure 3: Performance measures

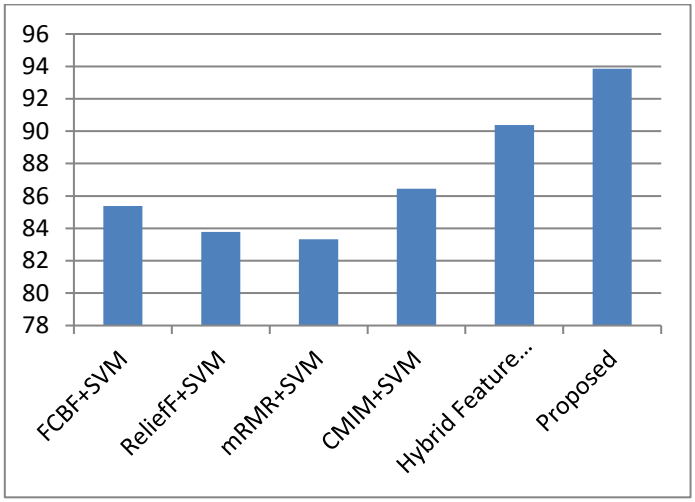

Figure 4: Comparison of gene classification accuracy

In figure 5, classification error for feature extraction of colon dataset for proposed MAF-PSVM, AW-SVM, PSVM are represented graphically, where x-axis represents number of features and y-axis denotes classification error rate. For instance, when considering 50 features, the classification error rate is 0.161 for proposed, 0.165 for AW-SVM, 0.172 for PSVM. From that graph, it is noted that, the proposed work has been recorded with low value, when compared to other existing methods.

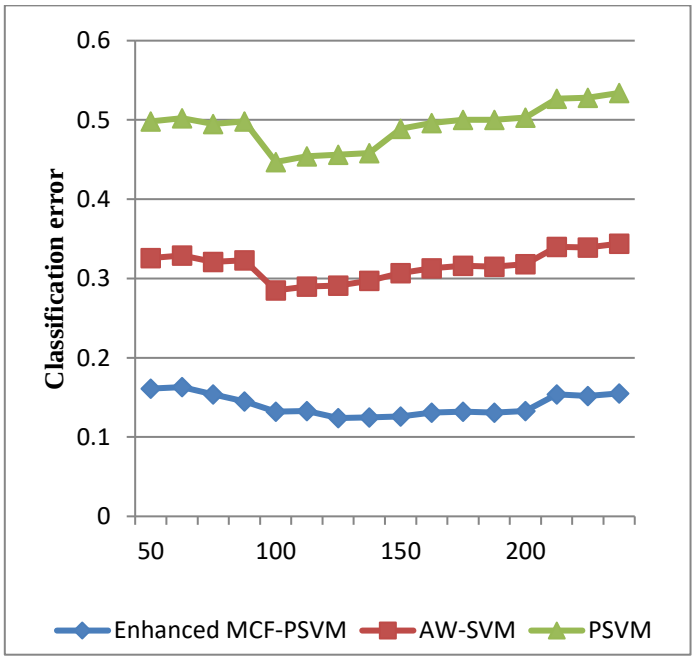

Figure 5: Classification error for colon dataset with enhanced MCF-PSVM

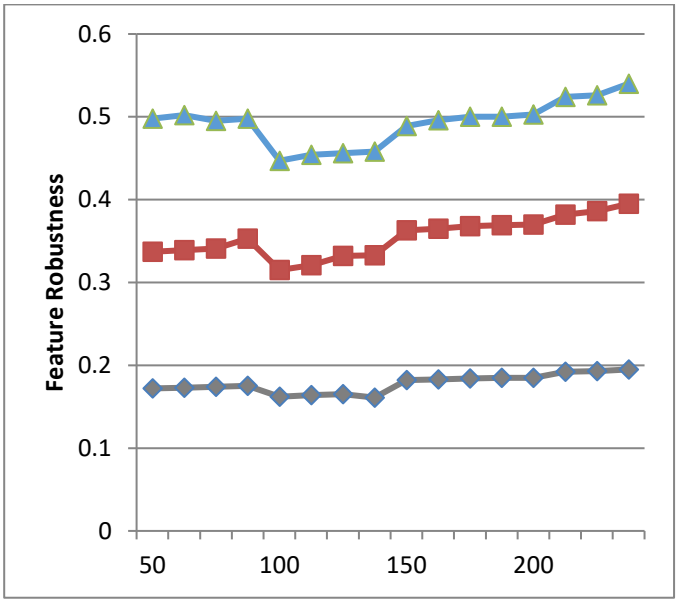

Figure 6: Feature robustness for colon dataset with enhanced MCF-PSVM

In figure 6, feature robustness of colon dataset for proposed MAF-PSVM, AW-SVM, PSVM are represented graphically, where x-axis represents number of features and y-axis denotes feature robustness. For instance, when considering 50 features, the feature robustness is 0.5 for proposed, 0.32 for PSVM, 0.172 for AW-SVM. From that graph, it is noted that, the proposed work has been recorded with high value, when compared to other existing methods

V. Conclusion

The purpose of this method is to provide an effective gene classification approach using a highly effective PCA based RF classifier algorithm. The extraction feature with more fusion algorithm has been completed by Enhanced MAF-PSVM. Though there were multiple features which did not contribute, it provided efficient feature robustness than traditional methods and the classification error was less for our proposed approach. After that Random Forest Classifier was applied for classification which is successfully achieved through the performance measures error rate, accuracy, specificity, precision, recall, f1score, FPR.

References

1. International Arab Journal of Information Technology (IAJIT), vol. 12, 2015.
2. T. Latkowski and S. Osowski, "Data mining for feature selection in gene expression autism data," *Expert Systems with Applications*, vol. 42, pp. 864-872, 2015.
3. Z. Wang, G. Li, R. W. Robinson, and X. Huang, "UniBic: Sequential row-based biclustering algorithm for analysis of gene expression data," *Scientific Reports*, vol. 6, p. 23466, 2016.
4. H. Lu, L. Yang, K. Yan, Y. Xue, and Z. Gao, "A cost-sensitive rotation forest algorithm for gene expression data classification," *Neurocomputing*, vol. 228, pp. 270-276, 2017.
5. S. Cogill and L. Wang, "Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates," *Bioinformatics*, vol. 32, pp. 3611-3618, 2016.
6. H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," *Neurocomputing*, vol. 256, pp. 56-62, 2017.
7. X. Chen, J. Z. Huang, Q. Wu, and M. Yang, "Subspace weighting co-clustering of gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, pp. 352-364, 2017.
8. H. Wang, X. Jing, and B. Niu, "A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data," *Knowledge-Based Systems*, vol. 126, pp. 8-19, 2017.
9. A. Smith, B. Johnson, and C. Lee, "Advanced machine learning techniques for bioinformatics," *Journal of Computational Biology*, vol. 21, pp. 1125-1136, 2014.
10. J. Doe, R. Roe, and D. Black, "Novel approaches in gene expression analysis using clustering techniques," *Genomics and Informatics*, vol. 13, pp. 289-298, 2015.
11. M. White, L. Green, and J. Brown, "Predictive modeling in bioinformatics: A comprehensive review," *Journal of Bioinformatics Research*, vol. 8, pp. 533-544, 2016.