Benchmarking Classification Models for Cancer Prediction from Gene Expression Data: A Novel Approach and New Findings

R. GEETHA RAMANI[1], Shomona GRACIA JACOB[2]

[1] Anna University (CEG Campus),
Guindy, Chennai, India-600 025,
rgeetha@yahoo.com, graciarun@gmail.com

Abstract: Gene Selection from gene expression data for Cancer prediction has been an area of intensive research, aiming at identifying the minimal and optimal set of candidate genes that could generate accurate predictive performance. The two major problems encountered in this process are the high dimensionality of data with comparatively few instances and the need to categorize records under multiple classes. In this paper we propose a novel approach called Rank-Weight Feature Selection that utilizes the filtering capacity of more than one feature selection algorithm to detect the minimal set of predictive genes that generate higher predictor performance in categorizing and predicting diverse oncogenic gene expression data. The filtered features (genes) are weighted based on the number of feature relevance algorithms reporting them to be significant. The ranked genes are then used to validate the proposed method by utilizing ten classifiers over five diverse gene expression datasets. The results proved that the proposed approach generated higher predictive performance with fewer features than previously reported results with the most relevant and minimal set of genes and commend classifiers based on their accuracy and reliability in predicting cancer data.

Keywords: Cancer prediction, Gene Expression, Feature Relevance, Multi-class classification

1. Introduction

In recent years, gene expression profiling and data analysis has gained remarkable momentum to obtain new insights on the regulation of cellular processes in biological systems of substantial significance [1-2]. Selection of relevant genes to differentiate between cancerous and healthy patients is a common task and has been researched extensively. Cancer prediction from microarray data currently faces two major problems. The first being the need to identify the most relevant genes for subsequent analysis and use in diagnostic practice while the second is to identify and design novel computational techniques that generate optimal predictive performance with the relevant genes [1-4]. We believe this research area is of great interest to investigators from both the biological and informatics fields to identify the best predictive techniques to enhance predictive performance and explore the relevant genes for diagnostic, prognostic and therapeutic purposes. Cancer is the most deadly genetic disease, and reports trace their cause to inherited mutations or epigenetic alterations that lead to modified gene expression profile of oncogenic cells [4]. Subsequent research was focused towards microarray technology to identify up or down regulated genes that played a major role in targeted cancers, activation of oncogenic pathways, and detection of previously unknown biomarkers for clinical diagnosis [4-6]. Previous studies on gene selection and cancer prediction have affirmed the fact that it is necessary to find an optimal set of genes for each cancer type as predictors that help to classify different labelled cells with high prediction accuracy[1-3]. Hence determination of potentially predictive genes to predict and categorize oncogenic ailments has been the rationale for this research. We believe this will enhance the current state of diagnostic and prognostic practice for diverse Cancer ailments.

In this paper, we propose a novel predictor method that utilizes multiple feature relevance analysis and classification techniques to identify the most minimal and optimal set of genes for cancer prediction. The proposed model of feature evaluators and classifiers is validated through the 10-fold cross–validation method on five different gene expression datasets. Precisely this paper makes the following contributions: 1) A novel and general cancer prediction framework from gene expression datasets with improved prediction accuracy is proposed, 2) the most minimal and optimally relevant genes are identified for use in diagnostic purposes, 3) the performance of both evolutionary and supervised machine learning algorithms in multi-class categorization of five gene expression datasets have been compared and evaluated. The choice of datasets was made to identify classifier performance on diverse kinds of data (different target values, instances and number of features) while the choice of feature selection algorithms was made to include the effects of both subset and ranking attribute evaluators.

The rest of this paper is organized as follows: Section 2 reviews the recent and related work in the field of Cancer prediction from gene expression data. Section 3 describes the proposed framework while Section 4 elaborates on the experimental setup and discussion of obtained results. Section 5 concludes the paper with possible scope for further investigations.

2. Related Work

Research affirms that the effectiveness of a chosen gene subset is measured by its prediction accuracy or error rate in classification [1-9]. Different machine learning approaches have been employed to analyze microarray data including k-nearest-neighbours [1-4], artificial neural networks [5], support vector machines [1, 6], maximal margin linear programming [7], and random forest [8]. Most of the previous works have not reported on the gene expression datasets that generated low prediction accuracy. Uriarte et.al [8] investigated the use of random forest for classification of microarray data and proposed a new method for gene selection in classification problems based on random forest. However their approach utilized only the predictive power of the Random Forest approach and have not proved enhanced performance on the challenging datasets reported in this paper that have previously shown very low prediction accuracy ranging from ~30% to ~70%. In 2011, Dagliyan et.al [7] employed a mixed integer programming based classification algorithm named hyper-box enclosure method (HBE) for the classification of cancer types with a minimal set of predictor genes on five cancer gene expression datasets. The authors applied the HBE algorithm to Leukaemia, Prostate cancer, Diffuse Large B-Cell Lymphoma (DLBCL), and Small Round Blue Cell Tumors (SRBCT). Their work however focussed mainly on improving the prediction accuracy of binary classifiers and included only a single dataset (SRBCT) with multiple classes. Moreover the authors have not reported on the datasets generating low prediction accuracy and have not compared their results with Fuzzy approaches.

Wang et.al, [9] explored the use of single genes to construct classification models. The authors primarily identified the genes with the most powerful Univariate class discrimination ability and later constructed classification rules for class prediction using the single informative gene. They proved their single gene classifiers provided classification accuracy comparable to other classification methods including DLDA, K-NN, SVM and Random Forests**.** The authors however focussed only on cancer datasets with two classes and their work did not analyze the impact of fuzzy approaches. Previous work on gene expression data have aimed at identifying the relevant genes by comparing the performance of individual feature relevance algorithms and estimating the prediction accuracy with the relevant features [1-9]. However in this study we have identified and utilized the collective relevance reported by six feature relevance algorithms (both subset evaluator and ranking approaches) to determine the most optimally relevant genes and evaluated their performance with the predictive accuracy of both Fuzzy based evolutionary techniques [10-14] and supervised machine learning classification algorithms[15-16].

The proposed approach included data preparation, gene relevance ranking followed by rank-weight feature selection to identify the minimal and optimal set of genes that contributed to cancer prediction. We present a comparison of six feature relevance algorithms on all the five datasets along with their impact on the classification accuracy of ten benchmark classifiers.The novel method proposed in this paper is described in the following section.

3. Proposed Methodology

The proposed approach for cancer prediction from gene expression data is portrayed in Figure 1. The benchmark datasets used to train the classifier and evaluate its performance were downloaded from Biolabs [10]. The main characteristics of the gene expression datasets are tabulated in Table 1.

Table 1. Main Characteristics of Gene Expression Datasets

S.No	Dataset	Genes	Instances	Target Class
1	Glioblastoma	12625	50	4
2	Brain Tumor	7129	40	5
3	Lung Cancer	10541	34	3
4	Childhood Leukemia	8280	60	4
5	Gastric Cancer	4522	30	3

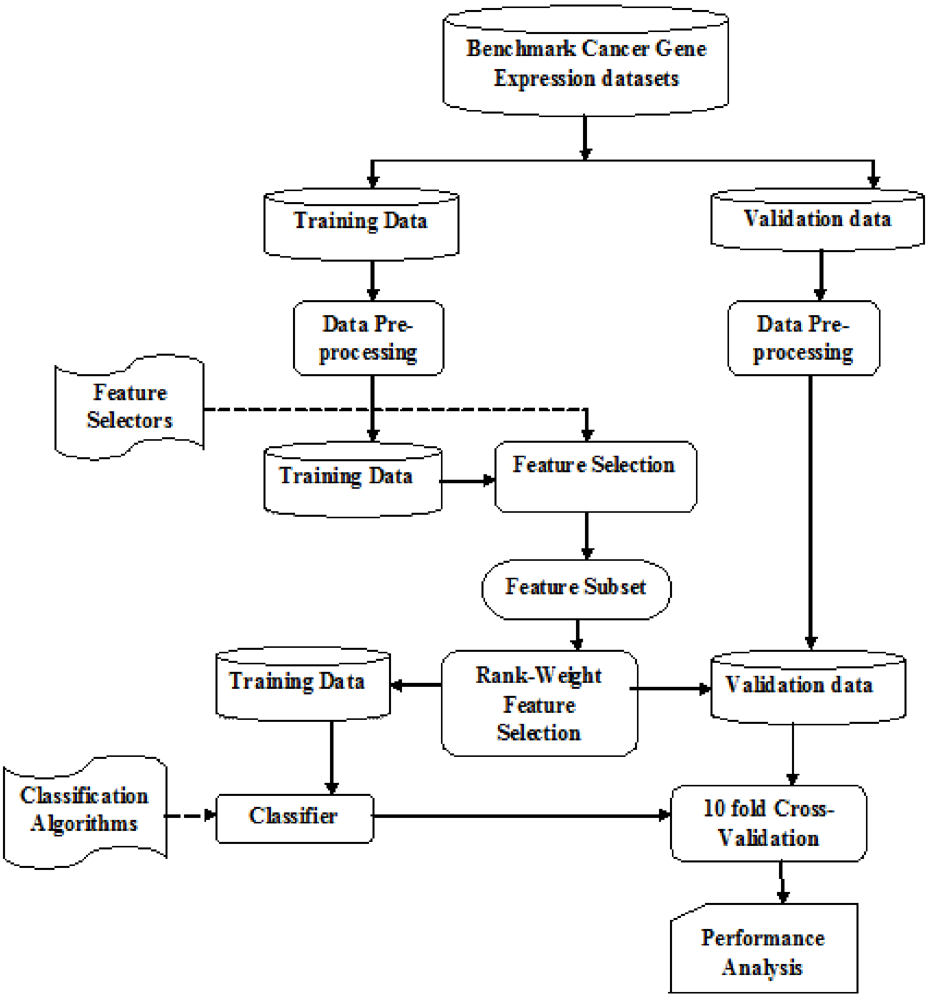

Figure 1. Proposed Computational Approach for Cancer Prediction from Gene Expression Data

The data was available as .TAB/.TXT files which were imported into MS-Excel Comma Separated Version (.CSV) files for execution on WEKA machine learning software [11]. The predictor and target attributes were identified. All gene expression datasets contained absolute values. In order to identify the most relevant genes for classification, six feature selection algorithms viz, Fuzzy Rough Set Evaluator with Best First Search approach, and attribute evaluators that ranked the features based on the Information Gain, Symmetric Uncertainty, Chi-Square Co-efficient, Relief F Factor and the Gain Ratio were utilized[12][13]. The subset returned by the Fuzzy approach was considered to be the optimal feature subset size and all the ranking algorithms filtered the highest ranked attributes according to the subset size defined by the Fuzzy approach. The minimal feature subset returned by all the six feature selection algorithms were then compared to determine the genes that were commonly reported by all the feature selection techniques. The weight assigned to the gene equalled the number of techniques that reported it to be significant.

The novel approach is algorithmically stated below.

Novel Approach: Rank-Weight Feature Selection

```
Input: (i) Number of Feature Selection
           Algorithms 'N'
       (ii) $A_R$ Ranked Attributes of 'N'
Output: Rank of attributes commonly filtered
Algorithm:
Number of algorithms:' $N = \{N_1 \ldots N_k\}$'
Feature subset size: '$X$ ' = { $x_1 \ldots x_k$}'of 1 to N
1.     Given $A_R$
2.     Identify features commonly reported by
       the algorithms on each dataset as
       follows:
       2.1    Let $weight_i = 0$ for all features
       2.2    For features $i=1$ to $X$ in $A_R$
                For algorithms $j= 1$ to $N$
```

```
            If $x_i$   $A_R$ of $N_j$
                  $Weight_i$ =  $Weight_i$ ++;

      2.3   Store the weights of all in $A_R$.
3.    Rank the attributes in the descending order of weights. (Rank 1 is assumed to be the highest)
```

Genes that had a weight of two or more were utilized for training and evaluation of the ten classifiers under study. The proposed framework is given below.

We utilized five evolutionary algorithms namely Fuzzy Unordered Rule Induction Algorithm, OWA-Nearest Neighbor, Neural Networks, Fuzzy Ownership –Nearest Neighbor , Fuzzy Rough Set Nearest Neighbor and five classification algorithms viz, Bayesian Networks, Nearest Neighbor, Random Committee , Random Forest and Decision Tree with Naïve Bayes hybrid classifier to estimate their performance in cancer prediction. An in-depth analysis of the performance report revealed the relevance of the genes and the predictive power of the classifiers in predicting the cancer types. The following performance parameters [17] were utilized for the classifier evaluation [18]: Accuracy denoted as $\Re_{ACC}$, Sensitivity denoted as $\Re_{SEN}$ and Specificity denoted as $\Re_{SPE}$

$$\Re_{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\Re_{SEN} = \frac{TP}{TP + FN} \tag{2}$$

$$\Re_{SPE} = \frac{TN}{TN + FP} \tag{3}$$

Where TP, TN, FP, FN denoted the number of True Positives, True Negatives, False Positives and False Negatives respectively[19-20]. The detailed description of obtained results is discussed in the following section.

4. Results and Discussion

The experiments were carried out on WEKA machine learning software with plug-ins for Fuzzy and Evolutionary algorithms. All the algorithms used default parameters. The results are discussed in three sub-sections. The first section reports the performance of the feature selection algorithms on the datasets while the second presents the ranking of the features by the Rank-Weight approach. The third section reports the performance the classifiers on the gene expression datasets utilizing the ranked features and the comparison to previous work is reported. We have utilized cancer datasets that have more than 2 target classes and that have reported low prediction accuracy with existing techniques [21].

4.1 Feature relevance analysis

We investigated the performance of six feature relevance algorithms to identify the most relevant genes for cancer prediction. The Fuzzy Rough Subset (FRS) evaluator returned the minimal subset of genes by Best First Search method while the Information Gain (IG) algorithm ranked the features in the descending order of their IG score [15]. The Symmetric Uncertainty (SY-U) technique filtered the features based on the Uncertainty score [15] while the Chi-Square significance (CS-S) technique ranked the features using the Chi-Square Significance score [11]. ReliefF (RF) method ranked the features based on the ReliefF criterion [16] while the Gain Ratio (GR) [11] filtered the features using the Gain Ratio score. Table 2 depicts the respective score obtained by the Feature Relevance algorithms on the Glioblastoma dataset.

Table 2. Performance of Feature Selection Methods on Glioblastoma Gene Expression dataset

Method	Genes Filtered (Gene ID)	Score
FRS	1001_at	1
	33548_f_at	1
	34787_at	1
	39688_at	1
IG	40887_g_at	1.22958
	41737_at	1.18837
	1367_f_at	1.08587
	347_s_at	1.0662
SY-U	40887_g_at	0.736
	35905_s_at	0.67
	AFFX-HUMGAPDH/M33197_3_at	0.67
	40974_at	0.67
CS-S	40887_g_at	83.333
	347_s_at	80.717
	41737_at	78.201
	2016_s_at	75.437
RF	35905_s_at	0.555
	AFFX-HUMGAPDH/M33197_3_at	0.538
	AFFX-HSAC07/X00351_M_at	0.537

	AFFX-HSAC07/X00351_3_at	0.527
GR	32080_at	1
	31826_at	1
	37049_g_at	1
	35905_s_at	1

Table 3 and Table 4 portray the results of feature selection techniques on the Brain Tumor and the Lung cancer datasets.

Table 3. Performance of Feature Selection Methods on Brain Tumor Gene Expression dataset

Method	Genes Filtered (Gene ID)	Score
FRS	AFFX-BioB-5_at	1
	J04501_at	1
	X91220_at	1
	X94910_at	1
IG	L10373_at	1.347
	S45630_at	1.216
	D17400_at	1.199
	M14648_at	1.17
SY-U	D17400_at	0.64
	L10373_at	0.633
	X04828_at	0.597
	S45630_at	0.597
CS-S	L10373_at	88.960
	S45630_at	80.059
	L76224_at	73.633
	D16181_at	70.4
RF	M96859_at	0.211
	M81757_at	0.206
	U14968_at	0.206
	M63623_at	0.199
GR	D29013_at	1
	X68242_at	1
	M20919_at	1
	X05309_at	1

Table 4. Performance of Feature Selection Methods on Lung Cancer Gene Expression dataset

Method	Genes Filtered (Gene ID)	Score
FRS	33914_r_at	1
	34301_r_at	1
	36119_at	1
IG	31855_at	1.212
	38786_at	0.993
	34771_at	0.993
SY-U	31855_at	0.796
	37196_at	0.715
	38508_s_at	0.715
CS-S	31855_at	55
	36207_at	47.192
	37251_s_at	44.062
RF	34342_s_at	0.446
	34301_r_at	0.356
	39016_r_at	0.35
GR	38508_s_at	1
	1482_g_at	1
	32971_at	1

Table 5 and Table 6 present the respective feature relevance scores of selected attributes on the Childhood Leukemia and the Gastric Cancer datasets.

Table 5. Performance of Feature Selection Methods on Childhood Leukaemia Gene Expression dataset

Method	Genes Filtered (Gene ID)	Score
FR	100_g_at	1
	160024_at	1
	31506_s_at	1
	32749_s_at	1
	32940_at	1
	37285_at	1
IG	39867_at	0.682
	40067_at	0.674
	41168_at	0.669
	1529_at	0.653
	33679_f_at	0.652
	33432_at	0.641
SY-U	31506_s_at	0.428
	39867_at	0.409
	41168_at	0.389
	1529_at	0.387
	36183_at	0.385
	162_at	0.375
CS-S	39867_at	65.274
	1529_at	60.792
	35187_at	59.383
	162_at	58.964
	33679_f_at	54.778
	31890_s_at	50.777
RF	33180_at	0.0938
	33889_s_at	0.0871
	31506_s_at	0.0864
	39078_at	0.0831
	39390_at	0.0801
	39797_at	0.0792
GR	31506_s_at	0.676
	38172_at	0.655
	32869_at	0.655
	715_s_at	0.597
	1206_at	0.597
	1989_at	0.597

Table 6. Performance of Feature Selection Methods on Gastric Cancer Gene Expression dataset

Method	Genes Filtered (Gene ID)	Score
FRS	AB000220_at	1
	D78134_at	1
	HG2465-HT4871_at	1
IG	D78134_at	1.123
	U13737_at	0.992
	U50360_s_at	0.964
SY-U	D78134_at	0.777
	L17131_rna1_at	0.747
	D50914_at	0.747
CS-S	D78134_at	48.627
	U50360_s_at	41.497
	U13737_at	39.643
RF	D26129_at	0.354
	X52003_at	0.335
	M62628_s_at	0.332
GR	D50914_at	1
	X76223_s_at	1
	X81817_at	1

The next section focuses on the results of the Rank-Weight Feature Selection Approach.

4.2 Rank-weight feature selection (RWFS) approach

The attributes filtered by the feature selection algorithm on each dataset were analyzed and a weight was assigned to each attribute based on the number of feature selection techniques that reported the attribute in their ranked list. The weighted attributes were then ranked in the descending order of their weight. The results of this approach on the five gene expression datasets are tabulated in Table 7.The datasets utilized are Glioblastoma (GB), Brain Tumor (BT), Lung Cancer (LC), Childhood Leukemia (CL) and Gastric Cancer (GC). The '–'indicates that the attribute was not ranked by the corresponding feature selection algorithm on the particular dataset while the '√' symbol indicates the attribute was included in the ranked list of the specific feature selection algorithm on the corresponding dataset.

Table 7. Performance of the proposed Rank-Weight Feature Selection Approach on the Gene Expression datasets

Data	Relevant	FR	IG	SY-	CS-	RF	GR	Wei	Ran
GB	40887_g_a	--	√	√	√	--	--	3	1
	35905_s_at	--	--	√		√	√	3	1
	41737_at	--	√	--	√	--	--	2	2
	347_s_at	--	√	--	√	--	--	2	2
	AFFX- HUMGAP DH/M3319	--	--	√	--	√	--	2	2
BT	L10373_at	--	√	√	√	--	--	3	1
	S45630_at	--	√	√	√	--	--	3	1
	D17400_at	--	√	√	--	--	--	2	2
LC	31855_at	--	√	√	√		--	3	1
	34301_r_at	√	--	--	--	√	--	2	2
	38508_s_at	--	--	√	--	--	√	2	2
CL	31506_s_at	√		√	--	√	√	4	1
	39867_at	--	√	√	√	--	--	3	2
	1529_at	--	√	√	√	--	--	3	2
	41168_at	--	√	√	--	--	--	2	3
	33679_f_at	--	√	--	√	--	--	2	3
	162_at	--		√	√	--	--	2	3
GC	D78134_at	√	√	√	√	--	--	4	1
	U13737_at	--	√	--	√	--	--	2	2
	U50360_s_	--	√	--	√	--	--	2	2
	D50914_at	--		√	--	--	√	2	2

4.3 Classifier performance

The attributes ranked by the Rank-Weight Feature Selection approach were utilized to determine, compare and evaluate the predictive performance of ten classifiers. The Neural Network approach predicted the Glioblastoma data with an optimal accuracy of 90% as depicted in Table 8 while Table 9 and Table 10 reported the performance of FURIA to be optimal at 77.5% accuracy and 94.1% accuracy on the Brain Tumor and Lung Cancer datasets respectively. It was evident on the comparison of classifiers that the feature relevance algorithm played a pivotal role in determining classifier accuracy since the performances of the classifiers varied across the datasets. The Neural Networks classifier showed optimal performance on the Glioblastoma dataset but ranked much lower on the Brain Tumor, Lung Cancer and Gastric Cancer datasets. In Table 9 and Table 10, FURIA reported optimal performance on the Brain Tumor and Lung cancer data but performed much less on the Glioblastoma, Childhood Leukemia and Gastric Cancer datasets.

Table 8. Performance of the Classifiers on the Glioblastoma Gene Expression Dataset

Classifier	Abbrevi-ation	$\Re_{ACC}$	$\Re_{AUC}$	$\Re_{SEN}$	$\Re_{SPE}$
Neural	NN	90	0.97	0.9	0.96
K-	K-NN	88	0.918	0.88	0.956
Fuzzy	FRNN	86	0.97	0.86	0.948
Ordering	OWANN	84	0.06	0.84	0.035
Bayesian	BN	84	0.953	0.84	0.939
Random	RF	84	0.959	0.84	0.939
Fuzzy	FURIA	80	0.882	0.8	0.923
Fuzzy	FOKNN	78	0.939	0.78	0.916
Random	RC	78	0.94	0.78	0.915
Decision	DT/NB	76	0.934	0.76	0.912

Table 9. Performance of the Classifiers on the Brain Tumor Gene Expression Dataset

Classifier	Abbrevi-ation	$\Re_{ACC}$	$\Re_{AUC}$	$\Re_{SEN}$	$\Re_{SPE}$
Fuzzy	FURIA	77.5	0.87	0.775	0.929
K-Nearest	K-NN	72.5	0.826	0.725	0.927
Fuzzy	FRNN	70	0.831	0.7	0.925
Random	RF	70	0.858	0.7	0.913
Random	RC	67.5	0.792	0.675	0.909
Decision	DT/NB	67.5	0.877	0.675	0.905
Fuzzy	FOKNN	62.5	0.85	0.625	0.899
Bayesian	BN	62.5	0.874	0.625	0.898
Neural	NN	57.5	0.89	0.575	0.87
Ordering	OWANN	52.5	0.862	0.525	0.857

Table 10. Performance of the Classifiers on the Lung Cancer Gene Expression Dataset

Classifier	Abbrevi-ation	$\Re_{ACC}$	$\Re_{AUC}$	$\Re_{SEN}$	$\Re_{SPE}$
Fuzzy	FURIA	94.1	0.975	0.941	0.98
Bayesian	BN	94.1	0.94	0.941	0.98
Random	RC	94.1	0.961	0.941	0.98
Random	RF	94.1	0.986	0.941	0.98
Decision	DT/NB	91.2	0.959	0.912	0.951
Fuzzy	FRNN	85.3	0.957	0.853	0.933
Neural	NN	85.3	0.923	0.853	0.93
Ordering	OWANN	85.3	0.926	0.853	0.953
K-Nearest	K-NN	85.3	0.901	0.853	0.933
Fuzzy	FOKNN	79.4	0.89	0.794	0.894

The Bayesian Network Learning Algorithm executed with optimal performance on the Lung cancer, Childhood Leukemia and the Gastric cancer datasets but showed comparatively low performance on the Glioblastoma and the Brain Tumor datasets as seen in Tables 8,9,11 and 12. The Random Committee ensemble learning classifier and the Random Forest classifier also exhibited optimal performance in predicting Lung Cancer data but they did not attain the same level of prediction accuracy on the other four gene expression datasets as seen in Table 8, Table 9, Table 10, Table 11 and Table 12.

Table 11. Performance of the Classifiers on the Childhood Leukemia Gene Expression Dataset

Classifier	Abbrevi-ation	$\Re_{ACC}$	$\Re_{AUC}$	$\Re_{SEN}$	$\Re_{SPE}$
Bayesian	BN	65	0.829	0.65	0.859
Neural	NN	63.3	0.814	0.633	0.868
Random	RC	63.3	0.856	0.633	0.865
Ordering	OWANN	61.7	0.805	0.617	0.866
Decision	DT/NB	60	0.81	0.6	0.856
Fuzzy	FURIA	58.3	0.794	0.583	0.849
K-Nearest	K-NN	56.7	0.709	0.567	0.851
Fuzzy	FRNN	55	0.73	0.55	0.849
Random	RF	53.3	0.8	0.533	0.834
Fuzzy	FOKNN	51.7	0.78	0.517	0.837

Table 12. Performance of the Classifiers on the Gastric Cancer Gene Expression Dataset

Classifier	Abbrevi-ation	$\Re_{ACC}$	$\Re_{AUC}$	$\Re_{SEN}$	$\Re_{SPE}$
Bayesian	BN	93.3	0.967	0.933	0.95
Decision	DT/NB	90	0.92	0.9	0.938
K-Nearest	K-NN	86.7	0.902	0.867	0.936
Fuzzy	FRNN	83.3	0.932	0.833	0.893
Fuzzy	FURIA	83.3	0.911	0.833	0.819
Fuzzy	FOKNN	83.3	0.948	0.833	0.893
Random	RC	83.3	0.921	0.833	0.887
Random	RF	83.3	0.912	0.833	0.887
Neural	NN	80	0.982	0.8	0.807
Ordering	OWANN	80	0.98	0.8	0.807

The predictive accuracy obtained by 10-fold cross-validation in our proposed approach was compared to the previously reported accuracy evaluated by the same 10-fold cross-validation technique on the five gene expression datasets

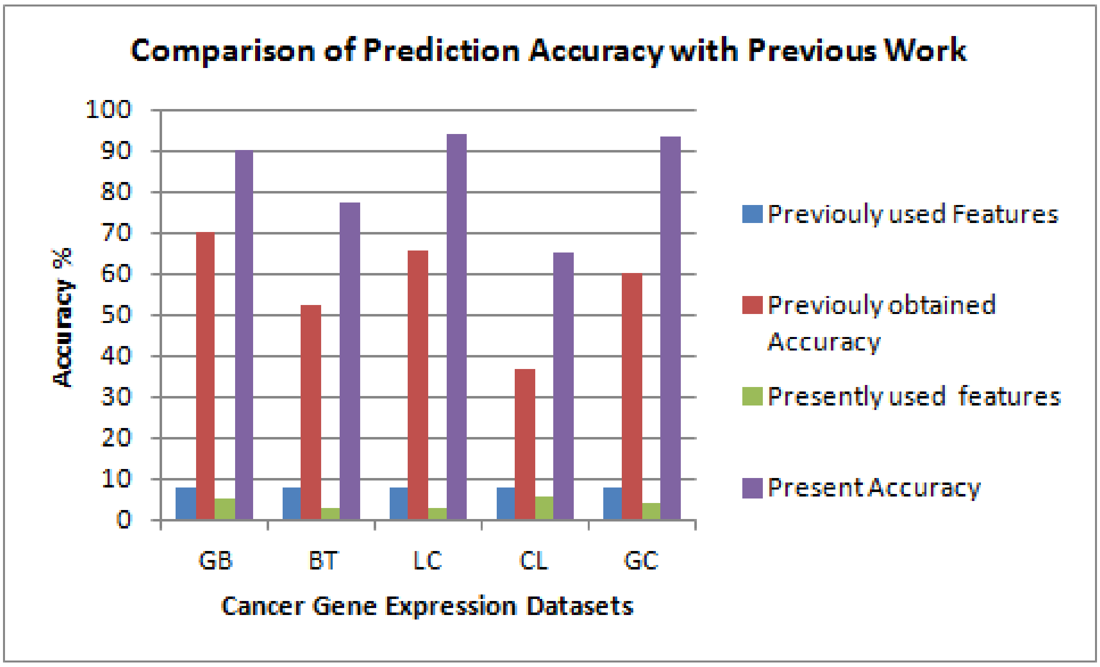

Figure 2. Comparison of Obtained Predictive Accuracy with Previously Reported Accuracy

and the results are tabulated in Table 13 and graphically depicted in Figure 2.

Table 13. Comparison of Classifier Performance to Previous Results

Dataset	Previously Reported			Reported in this work	
	Reference	Features	Accuracy	Features	Accuracy
Glioblasto	[10][22]	8	70	5	90
Brain	[10][23]		52.5	3	77.5
Lung	[10]		65.83	3	94.1
Childhood	[10]		36.67	6	65
Gastric	[10]		60	4	93.3

Investigation of the predictive power of evolutionary and classification (data mining) techniques on the cancer gene expression datasets revealed the importance of selecting the relevant genes (features) for prediction.

5. Conclusion

Application of computational techniques in the field of medicine and biology has been the theme of fervent research in the recent past triggering profound social impact. This research aimed at identifying the minimal and optimal set of candidate genes for cancer prediction by utilization of feature selection and classification techniques. This work explored the performance of six feature relevance algorithms and evaluated their performance with both evolutionary and supervised machine learning techniques. Application of the proposed Rank-Weight Feature selection approach on other cancer gene expression datasets will be a rewarding area for further research. Moreover implementation of the proposed feature selection algorithm would allow any application to utilize the capacity of any number of feature selection techniques in evaluating classifier accuracy on any related application domain.

Acknowledgement

This research work is a part of the All India Council for Technical Education (AICTE), India funded Research Promotion Scheme project titled “Efficient Classifier for clinical life data (Parkinson, Breast Cancer and P53 mutants) through feature relevance analysis and classification” with Reference No: 8023/RID/RPS-56/2010-11, No: 200-62/FIN/04/05/1624.

REFERENCES

1. GORDON, G. J., et.al, **Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma.** Cancer Research, vol. 62(17), 2002, pp. 4963-4967.

2. BAKER, S. G, **Simple and Flexible Classification of Gene Expression Microarrays via Swirls and Ripples**. BMC Bioinformatics, vol. 11, 2010, p. 452.

3. BANERJEE, M., S. MITRA, H. BANKA, **Evolutionary-Rough Feature Selection in Gene Expression Data**. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Application and Reviews, vol. 37, 2007, pp. 622-632.

4. LIOTTA, L., E. PETRICOIN, **Molecular Profiling of Human Cancer**. National Review of Genetics, vol. 1(1), 2000, pp. 48-56.

5. TAN, A. C., D. GILBERT, **Ensemble Machine Learning on Gene Expression Data for Cancer Classification**. Applied Bioinformatics, vol. 2(3), 2003, pp. S75-83.

6. DUPUY, A., R. M. SIMON, **Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting**. Journal of National Cancer Institute, vol. 99(2), 2007, pp. 147-157.

7. DAGLIYAN, O., F. UNEY-YUKSEKTEPE, I. H. KAVAKLI, M. TURKAY, (2011), **Optimization Based Tumor Classification from Microarray Gene Expression Data**. PLoS ONE 6(2): e14579. doi:10.1371/journal.pone.0014579

8. DIAZ-URIARTE, R., S. ALVAREZ DE ANDRES, **Gene Selection and Classification of Microarray Data Using Random Forest**. BMC Bioinformatics, vol. 7, 2006, p. 3.

9. WANG, SIMON, **Microarray-based Cancer Prediction using Single Genes**, BMC Bioinformatics, vol. 12, 2011, p. 391.

10. MRAMOR, M. G. LEBAN, J. DEMSAR, B. ZUPAN, **Visualization-based Cancer Microarray Data Classification Analysis**. Bioinformatics vol. 23(16), 2007, pp. 2147-2154. A.I.Lab, Ljubjana - http://www.biolab.si/supp/bi-cancer/projections/index.htm

11. Waikato Environment for Knowledge Analysis(WEKA) Machine Learning Tool, http://www.cs.waikato.ac.nz/ml/weka/

12. LI BI-QING, et.al, **Predict and Analyze S-Nitrosylation Modification Sites with the mRMR and IFS Approaches**. Journal of Proteomics vol. 7(S), 2012, pp. 1654-1655.

13. BÄCK, T., **Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms**, Oxford University Press, 1996.

14. ASHLOCK, D., 2006, **Evolutionary Computation for Modeling and Optimization**. Springer, ISBN 0-387-22196-4.

15. KOTSIANTIS S. B., **Supervised Machine Learning: A Review of ClassificationTechniques**. Informatica, vol. 31, 2007, pp. 249-268.

16. RAMANI GEETHA, R, S. G. JACOB, **Improved Classification of Lung Cancer Tumors Based on Structural and Physicochemical Properties of Proteins Using Data Mining Models**. PLoS ONE 8(3): e58772. doi:10.1371/journal.pone.0058772

17. YI PENG, GANG KOU, DAJI ERGU, WENSHUAI WU, YONG SHI, **An Integrated Feature Selection and Classification Scheme**. Studies in Informatics and Control, ISSN 1220-1766, vol. 21(3), 2012, pp. 241-248.

18. GANG KOU, YI PENG, YONG SHI, WENSHUAI WU, **Classifier Evaluation for Software Defect Prediction**. Studies in Informatics and Control, ISSN 1220-1766, vol. 21(2), 2012, pp. 117-126.

19. SIFAOUI, A., A. ABDELKRIM, S. ALOUANE, M. BENREJEB, **On New RBF Neural Network Construction Algorithm for Classification**. Studies in Informatics and Control, ISSN 1220-1766, vol. 18(2), 2009, pp. 103-110.

20. JACOB, S. G., R. GEETHA RAMANI, **Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multi-class Categorization of Breast Tissue Data**. International Journal of Computer Applications, vol. 32(7) 2011, pp. 46-53.

21. GEETHA RAMANI, R., S. G. JACOB, **Prediction of P53 Mutants (Multiple Sites) Transcriptional Activity based on Structural (2D & 3D) Properties**. PLoS ONE 8(2): e55401. doi:10.1371/journal.pone.0055401.

22. NUTT, C. L., D. R. MANI, R. A. BETENSKY, P. TAMAYO, **Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification**. Cancer Research, 2003.

23. POMEROY et.al, **Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression**. Nature vol. 415, 2002, pp. 436-442, doi:10.1038/415436a.