

A Robust and Efficient Feature Selection Algorithm for Microarray Data

Mehrab Ghanat Bari¹, Sirajul Salekin¹ and Jianqiu (Michelle) Zhang^{1*}

September 12, 2016

¹Department of Electrical and Computer Engineering, Univ. of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

1 Introduction

In this paper, we developed a new feature selection method, PSI, which based on the synergistic effects caused when the features combined to each other and are used as input of a classifier. The idea is that, irrelevant features, while combining other features are more likely to contribute negatively to classifier performance in train set. On the other hand, when informative genes are used in panels, the synergistic effect of them cause the classifier better learning, and this leads to better classification results. Our experiments show that, ranking features based on their performance when they are combined, could provide a good feature set to train a model to classify the unseen test set.

2 SVM vs LR in developing PSI

In the PSI algorithm, to compute the synergy scores of features, the individual and paired accuracy of preselected N features are needed, and it uses SVM classifier to do so. Here, we want to show the benefit of using SVM for the computation of accuracy of a feature or feature pairs, $Acc(F_i)$ and $Acc(F_i, F_j)$, respectively.

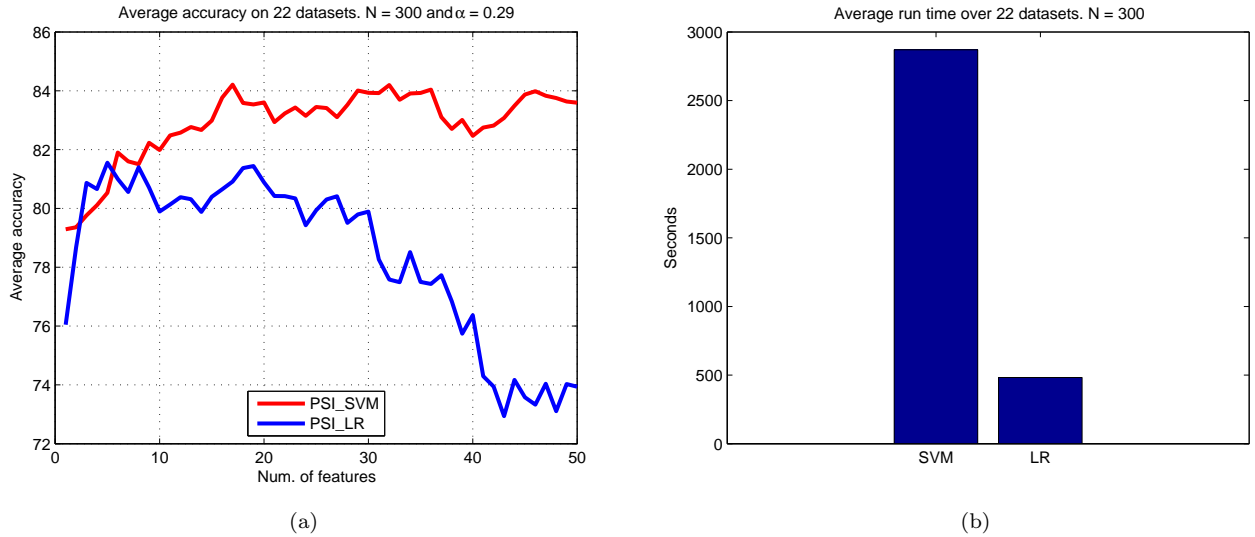


Figure S1: Comparing PSI accuracy performance and time complicity using SVM and LR, a) Average accuracy over 22 datasets b) Boxplot of the average of averages running time

*to whom correspondence should be addressed

As SVM scales with number of samples and Logistic Regression (LR) with number of features, this should be preferable for large datasets. It is a good practice to use LR instead of SVM in PSI body to see whether or not LR would be more efficient. The Figure S1 shows the average accuracy and run time of PSI when using LR and SVM over all the datasets. In both cases the $N = 300$, $\alpha = 0.29$ and 10-fold cross validation scheme was used. When $N = 300$, PSI is about 5 times faster than, when it uses LR instead of SVM. However, it causes very poor accuracy results as shown in the Figure S1.a. Note that, we proposed N to be equal to 100 in the final PSI while using SVM, which is at least 7 times faster than when $N = 300$ and still has the same good average accuracies. The effect of N on PSI performance is represented in the next section. Above results, convincing us to use SVM in PSI structure, and since we are using the original SVM function in MATLAB without any parameter setting, subsequently PSI uses SVM with the benefit of not having to select a cost parameters.

3 The effect of N on average performance

To find the best N for PSI, we examined the effect of N on the PSI's average performance and computational time. Since, bigger N makes PSI computationally expensive, the question is that how much a big N would increase the accuracy? As shown in the Figure S2, we compared PSI's average running time and accuracy for N equal to 100, 200 and 300. The mixing parameter, α , was set to 0.29 for all cases and 10-fold outer and 5-fold inner cross validation were used.

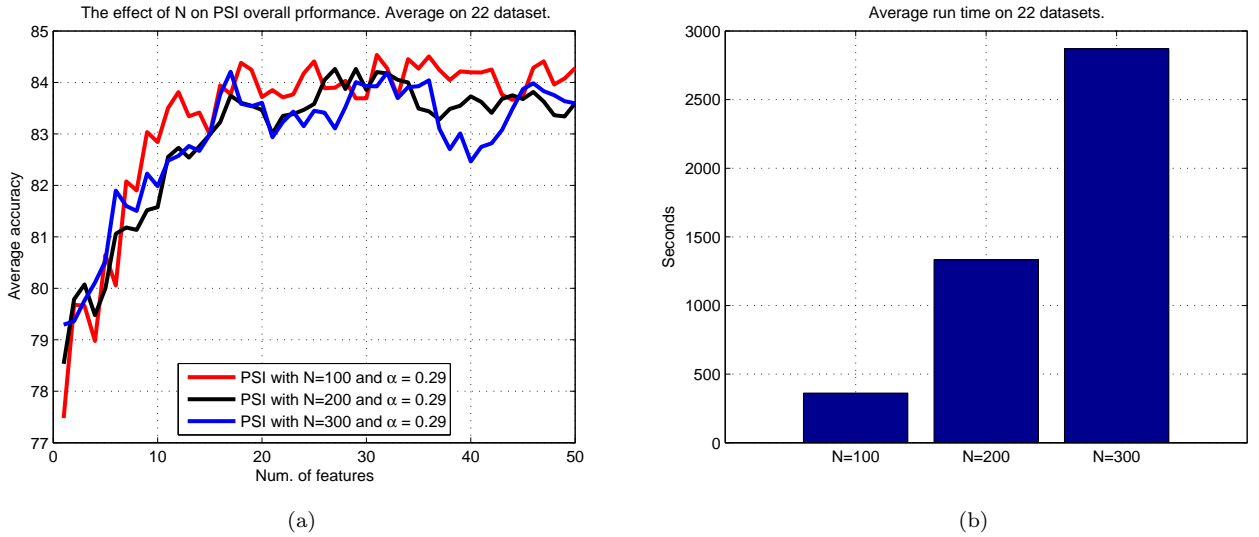


Figure S2: Comparing the The effect of N on average performance, a) Average accuracies, b) Average running time

The number of SVM classifiers need by PSI is equal to $10 \times 5 \times (N(N-1)/2)$ function N . The number 10 and 5 show the number of outer and inner cross validation. The number of SVM classifiers for $N = 100, 200$ and 300 are 247500, 995000 and 2242500 for each dataset respectively. Although, PSI was developed using parallel processing technique, by increasing N , running time increases rapidly without any improvement over PSI's results when N is 100.

Also for the datasets "Colon" and "CNS", which have 2000 and 7129 genes, we used all features to calculating the synergy scores. The number of classifiers are 99950000 and 1270400000 and the running time were 33 and 285 hours for "Colon" and "CNS" respectively, and in the both of the cases the accuracies were below than those of when N was equal to 100 (in average 2% less).

4 The effect of α on average performance

The mixing parameter α , is a parameter of PSI along with N the number of features considered (Set to 100). It is selected to be 0.29 based on the average of the α s cause best accuracy for each dataset in training step.

Figure S3 shows the effect of changing α on PSI's average accuracy on datasets "CNS", "GCM", "GSE27854" and "Prostate4", when PSI applies on unseen test sets. PSI average accuracy in most of 22 cases show a peak when α is around 0.16 like the cases "GCM" and "Prostate4". Because $\alpha = 0.16$ was seen in the outmost loop of the double CV scheme, then PSI may have an unfair advantage over the other methods, then we report the results by using $\alpha = 0.29$.

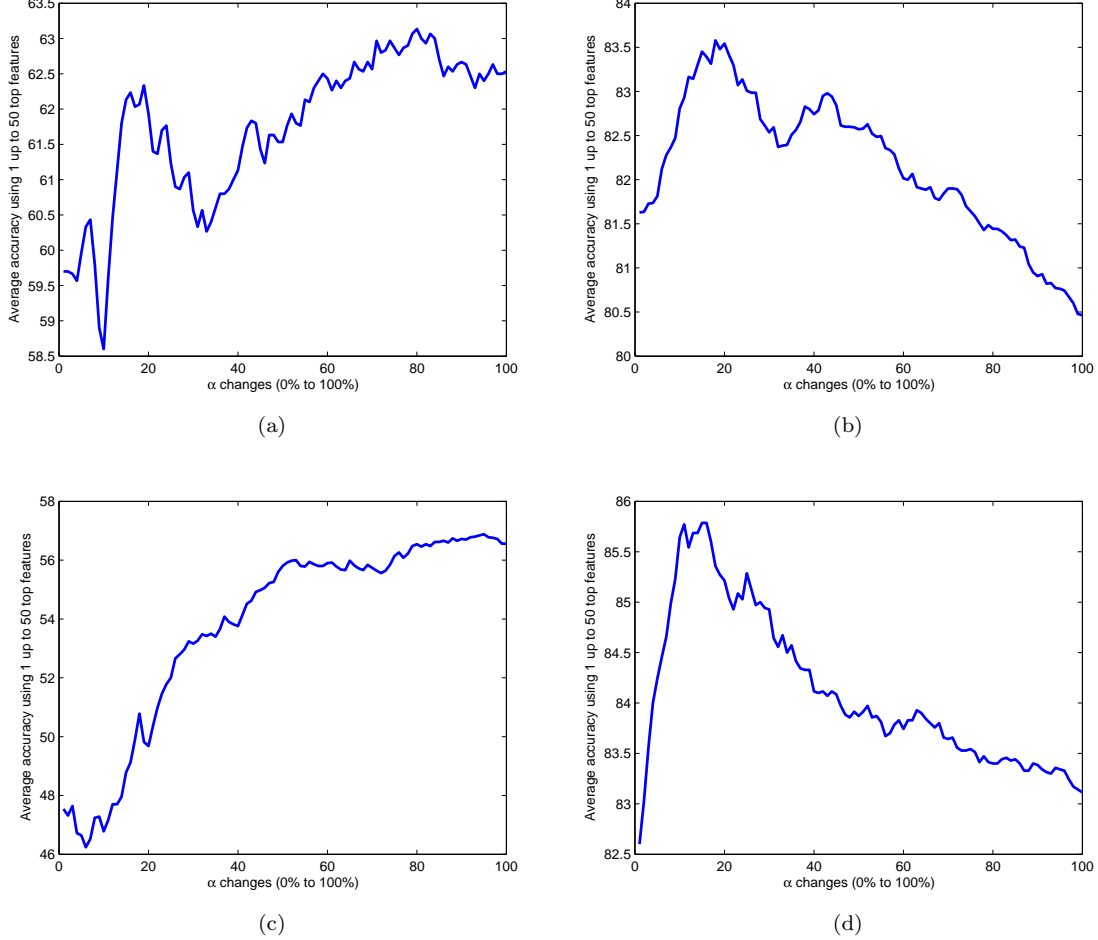


Figure S3: The effect of α on average accuracy of PSI using 1 up to 50 top features in dataset, a) CNS, b) GCM, c) GSE27854, d) Prostate3

5 KNN classifier

Each of the 14 feature selection methods in this study, were applied to all 22 datasets and the 50 features reported by each method, used to train the SVM classifier. PSI shows best average accuracy over all other method while having 4_{th} less computational times. PSI, also uses SVM to calculating the synergy scores. To make sure that PSI good performance is not because of classifier effect, we used top 50 features reported by all methods to train K-Nearest Neighbor (KNN) classifier, and then the trained KNN classifier were applied on test sets and the accuracy of each method were recorded. Figure S4 shows the average results over 22 datasets when top 1, 2, 10, 15, 25 and 50 features were used. The parameter K for KNN classifier is a user-defined positive constant and a common used choice to assign a value for K is equal to the square root of the number of samples [1]. It is a good starting point and given a dataset, we rounded its square root of the number of samples to the nearest odd number ($2 \times \text{round}(\sqrt{S} - 1/2) + 1$), and that was selected as the final K.

As shown in the Figure S4, PSI has the best accuracies in general like what we can see when the SVM used as final classifier. SAM method as explained in the main paper, when the number of features are small, has the

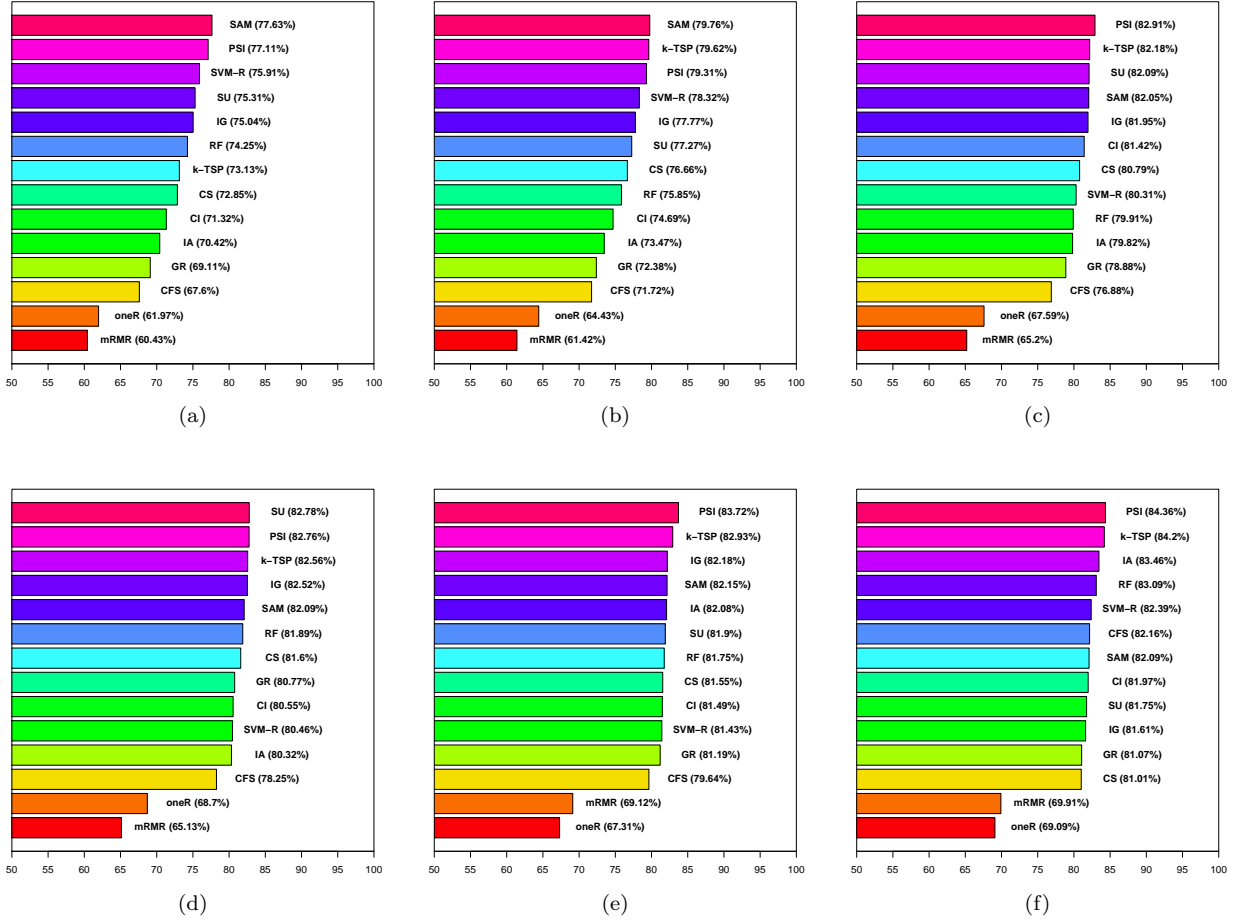


Figure S4: KNN classifier average 10-fold cross validation accuracy over 22 datasets using 1 in (a), 2 in (b), 10 in (c), 15 in (d), 25 in (e) and 50 in (f), features

highest average accuracy but as the number of features used are increased, the synergistic based methods, PSI and k-TSP could provide more informative features to the classifier and cause better performances. However, this is not k-TSP in reality, because it is proposed as classifier, which is based on majority of voting of negatively correlated pairs. It has been shown that when the features extracted by k-TSP used to train SVM classifier, the achieved performance is pretty better than of k-TSP as classifier [2]. Also, it shall be noted that the k-TSP is computationally is much more expensive the PSI. Our results show that, PSI in average is 35 times faster than k-TSP.

Also, Figure S5 shows box-plot of average accuracies achieved by using 1, 2, ..., 50 top features when SVM and KNN used as final classifiers. The 14 methods ranked by the median of the average accuracies they produced and it is clear in both case PSI has the best results. There are slight changes in other method's rank when different classifiers are used and these spares changes are acceptable due to internal characteristics of different classifiers. PSI, k-TSP and SAM are the three best method using SVM or KNN, and CI, the other synergistic method is in the 4th rank when SVM was used. The k-TSP has received those good performance because in this expremtent it uses SVM and KNN as classifiers and previous studies have shown than when k-TSP uses it own internal classifier, its performance will be lower than what were represented in this research, furthermore its time complexity is much more than PSI, since it uses combinations of two all features, while PSI needs to evaluate those of only $N = 100$ preselected features for the same dataset.

6 Wilcoxon test to examining the significance in accuracy gain

Figure S5 shows the box plot of accuracy results provided by 14 methods using top 10 features on 22 datasets, which are represented in the Table 3 in the main paper. We want to evaluate the significance of the gain in accuracy listed in Table 3. Wilcoxon test[3], is appropriate for evaluating the median difference in outcomes of two populations which are paired or dependent. The null hypothesis, H_0 , which indicates that the median of changes in accuracy comparing PSI and all others is equal 0. Table S11 shows the alternative hypothesis H_1 as well as test statistic, confidence interval and p-value when PSI's results compared to those of other methods. In all the comparisons, H_0 is rejected by Wilcoxon test at significance level equal to 0.05, and the conclusion is that data available in the Table 3, provide sufficient evidence to conclude that PSI is grater than those of each other methods at significance level of 0.05.

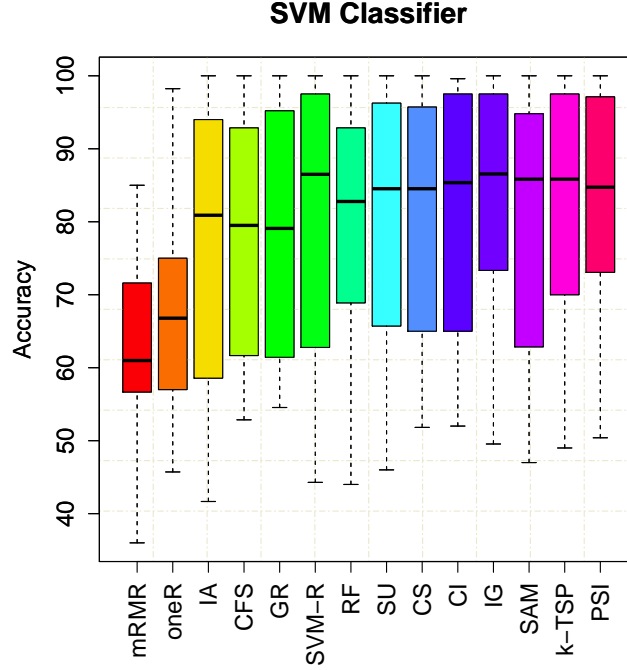


Figure S5: Boxplot of 10-fold cross validation accuracy results using top 10 features (SVM)

7 SVM, KNN and GLM classifiers comparison

Beside SVM and KNN classifiers we also used GLM classifier for comparison of 14 feature selection approaches. Figure S6 shows the boxplots of SVM, KNN, and GLM classifiers average performance over using 1 to 50 top features reported by all 14 methods on all 22 data sets in 10-fold cross-validation approach.

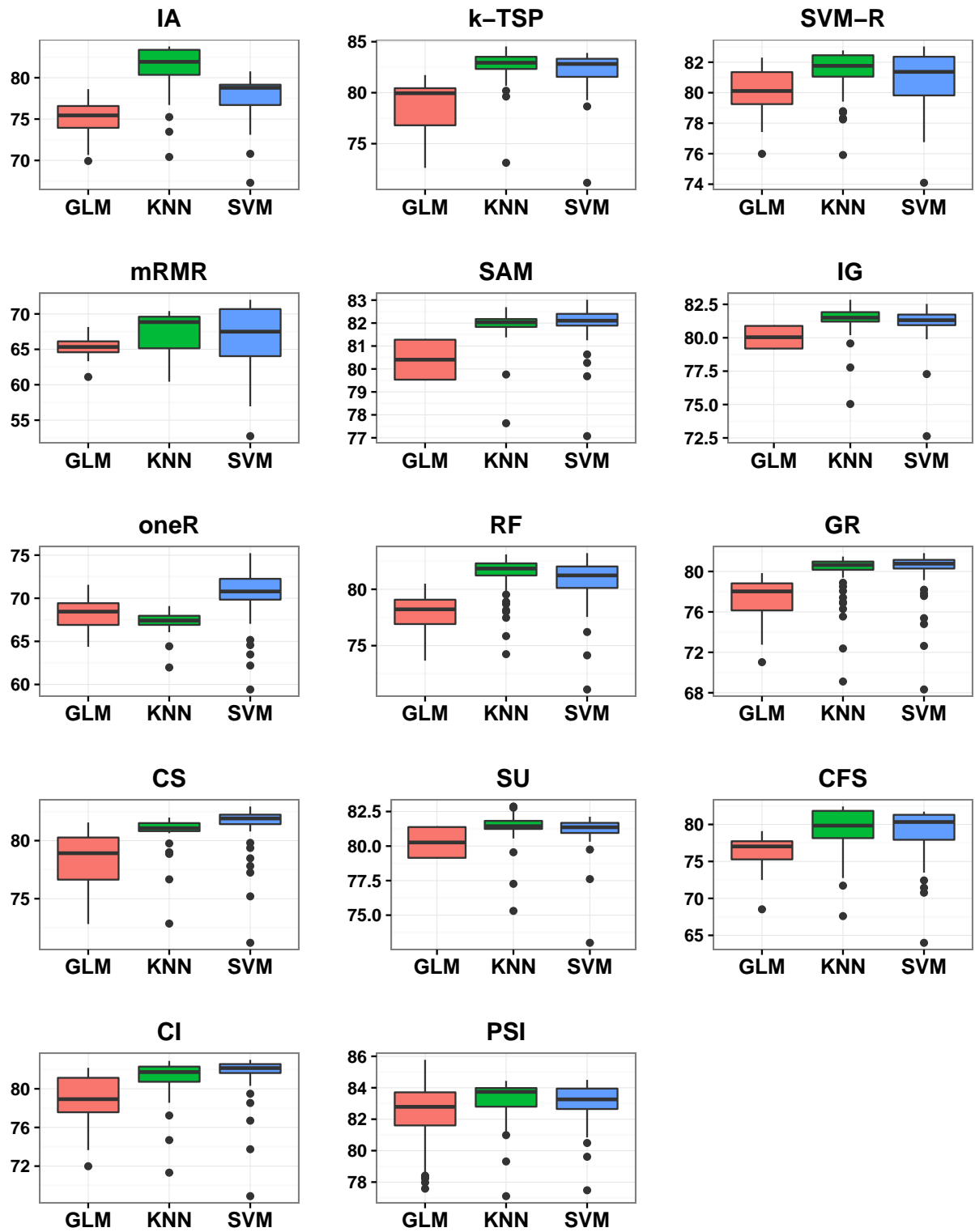


Figure S6: Boxplot of SVM,KNN, and GLM classifiers performance using 1 to 50 top features reported by all 14 methods

Table S1: Wilcoxon signed rank test with continuity correction.

	H_1	95% confidence interval	(pseudo)median	Test statistic	p-value	$-\log_2(\text{p-value})$
PSI v.s. IA	True*	(1.4 , 9.1)	4.73	198	0.0043	7.8
PSI v.s. k-TSP	True*	(-0.5 , 2.6)	1.39	157	0.15	2.6
PSI v.s. SVM-R	True*	(0.1 , 5.6)	2.64	176	0.037	4.7
PSI v.s. mRMR	True*	(14.5 , 25.3)	19.57	252	4.93e-05	14.3
PSI v.s. SAM	True*	(-0.9 , 2.5)	0.84	140	0.40	1.3
PSI v.s. IG	True*	(-1 , 2.1)	0.90	151	0.22	2.1
PSI v.s. oneR	True*	(9.1 , 21.4)	14.9	240	4.19e-05	14.5
PSI v.s. RF	True*	(0.7 , 5.8)	3.27	188	0.012	6.3
PSI v.s. GR	True*	(0.1 , 5.5)	2.70	179	0.028	5.1
PSI v.s. CS	True*	(-1.1 , 2.9)	1.12	147	0.28	1.8
PSI v.s. SU	True*	(0.01 , 2.4)	1.40	173	0.046	4.4
PSI v.s. CFS	True*	(3.1 , 9.1)	6.39	206	0.0017	9.1
PSI v.s. CI	True*	(-1.3 , 2.5)	0.44	150	0.45	1.1

* location shift is not equal to 0

8 Detailed results using SVM

To test PSI performance, we used 22 microarray binary class datasets, and we compared it on with 13 other existing methods. The 10-fold cross validation accuracy of different methods using top 1, 2, 15, 25 and 50 features are presented in Tables tables S2 to S6.

Table S2: 10-fold cross validation accuracy of different methods using top 1 feature.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
All	70.8	94.2	85.8	95.0	95.0	97.5	95.0	56.7	73.3	100	100	82.5	98.3	97.5
Brain	50.0	40.0	65.0	30.0	30.0	60.0	30.0	65.0	50.0	61.8	45.0	80.0	55.0	60.0
Breast1	50.0	58.6	65.7	51.4	52.9	52.9	61.4	44.3	47.1	51.4	58.6	70.0	62.9	52.9
Breast2	58.5	51.0	60.0	52.5	46.0	55.0	61.5	51.0	50.0	68.9	50.0	66.0	63.5	50.5
Carcinoma	80.0	82.5	82.5	82.5	87.5	70.0	90.0	42.5	62.5	88.9	77.5	82.5	95.0	82.5
CNS	63.3	46.7	45.0	40.0	43.3	51.7	48.3	45.0	51.7	81.4	51.7	76.7	46.7	50.0
Colon	61.7	71.7	63.3	65.0	71.7	75.0	81.7	61.7	58.3	84.7	83.3	81.7	58.3	73.3
DLBCL1	87.5	78.8	57.5	71.2	38.8	87.5	90.0	61.2	62.5	87.6	90.0	88.8	82.5	85.0
DLBCL2	80.0	87.5	65.0	85.0	62.5	80.0	70.0	32.5	45.0	91.4	95.0	95.0	85.0	82.5
GCM	66.1	73.9	53.2	56.8	76.4	70.7	70.4	61.1	55.7	74.6	68.6	75.0	68.2	72.5
GLI-85	62.2	72.2	66.7	71.1	77.8	71.1	73.3	53.3	72.2	73.6	81.1	67.8	77.8	70.0
GSE14333	52.3	54.5	48.2	50.9	54.1	49.1	48.6	50.9	50.5	45.7	51.4	49.1	50.5	48.6
GSE24514	77.5	87.5	77.5	87.5	77.5	85.0	92.5	65.0	77.5	88.9	70.0	87.5	85.0	82.5
GSE27854	45.0	41.0	57.0	53.0	41.0	51.0	44.0	48.0	53.0	43.4	46.0	56.0	46.0	52.0
Leukemia	82.9	87.1	64.3	91.4	84.3	95.7	85.7	45.7	62.9	95.7	82.9	94.3	90.0	94.3
Lung	92.4	91.8	82.4	94.7	94.1	91.2	81.8	48.8	94.7	93.2	94.7	85.9	95.9	91.8
Ovarian	85.2	89.6	58.8	89.6	86.4	88.8	86.0	62.8	60.4	87.4	88.8	76.8	88.8	89.6
Prostate1	72.0	82.0	59.0	80.0	88.0	83.0	85.0	57.0	52.0	80.4	73.0	86.0	85.0	86.0
Prostate2	52.2	58.9	57.8	58.9	65.6	62.2	66.7	47.8	58.9	78.1	47.8	75.6	75.6	62.2
Prostate3	70.7	67.9	43.6	53.6	57.9	67.9	50.0	53.6	55.7	76.4	55.7	67.9	70.7	67.9
SMK	59.4	61.1	55.6	55.0	63.9	62.8	57.8	50.6	53.3	60.8	65.6	62.8	59.4	64.4
SCBCT	96.0	88.0	94.0	88.0	86.0	90.0	96.0	56.0	60.0	89.4	88.0	88.0	90.0	90.0
Average	68.9	71.2	64.0	68.3	67.3	72.6	71.2	52.7	59.4	77.5	71.1	77.1	74.1	73.0

Table S3: 10-fold cross validation accuracy of different methods using top 2 features.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
All	81.7	96.7	95.0	96.7	95.8	98.3	100	65.8	72.5	100	100	88.3	99.2	97.5
Brain	60.0	45.0	80.0	50.0	40.0	80.0	55.0	60.0	50.0	61.8	50.0	80.0	60.0	80.0
Breast1	55.7	52.9	62.9	51.4	58.6	52.9	60.0	47.1	38.6	58.5	62.9	71.4	54.3	52.9
Breast2	62.0	52.0	61.5	44.5	41.0	55.5	63.0	54.0	47.5	66.9	54.0	62.0	62.0	55.0
Carcinoma	95.0	92.5	97.5	92.5	87.5	85.0	97.5	42.5	80.0	96.4	87.5	90.0	95.0	97.5
CNS	46.7	51.7	55.0	46.7	46.7	53.3	51.7	46.7	56.7	64.7	51.7	73.3	48.3	51.7
Colon	68.3	81.7	66.7	81.7	76.7	83.3	91.7	58.3	55.0	84.7	81.7	88.3	71.7	80.0
DLBCL1	93.8	86.2	71.2	77.5	50.0	91.2	97.5	62.5	66.2	97.5	90.0	93.8	87.5	88.8
DLBCL2	80.0	87.5	75.0	85.0	72.5	85.0	82.5	57.5	57.5	94	92.5	92.5	87.5	82.5
GCM	72.5	78.6	59.3	66.1	75.4	75.0	68.9	66.1	64.3	77.5	69.6	80.7	71.4	75.7
GLI-85	74.4	75.6	71.1	71.1	82.2	77.8	75.6	55.6	73.3	74.3	82.2	74.4	72.2	82.2
GSE14333	50.0	49.5	52.3	53.2	52.7	47.3	40.0	57.3	51.8	50.5	54.5	47.3	52.3	51.4
GSE24514	87.5	87.5	90.0	92.5	85.0	87.5	97.5	75.0	80.0	93.9	90.0	92.5	85.0	85.0
GSE27854	49.0	53.0	53.0	51.0	42.0	51.0	51.0	48.0	51.0	51.4	34.0	49.0	48.0	47.0
Leukemia	80.0	91.4	70.0	94.3	90.0	94.3	90.0	48.6	77.1	94.3	85.7	91.4	94.3	94.3
Lung	97.1	95.9	84.1	96.5	97.1	94.7	95.9	67.1	96.5	96.1	98.2	89.4	96.5	96.5
Ovarian	85.6	89.6	77.6	89.6	93.6	91.2	98.4	64.0	64.8	87.4	89.6	81.6	97.2	90.0
Prostate1	82.0	90.0	71.0	88.0	81.0	90.0	93.0	60.0	63.0	85.4	86.0	93.0	90.0	91.0
Prostate2	64.4	64.4	62.2	58.9	73.3	72.2	71.1	55.6	58.9	79.2	56.7	77.8	72.2	72.2
Prostate3	79.3	72.9	54.3	57.1	64.3	73.6	87.1	57.1	55.0	71.4	57.9	75.7	87.9	74.3
SMK	59.4	63.9	57.2	57.8	62.2	65.0	65.0	48.3	50.6	67.5	64.4	64.4	62.2	66.1
SCBCT	98.0	96.0	90.0	96.0	90.0	96.0	98.0	56.0	58.0	97.4	92.0	96.0	94.0	96.0
Average	73.7	75.2	70.8	72.6	70.8	77.3	78.7	57.0	62.2	79.6	74.1	79.7	76.8	77.6

Table S4: 10-fold cross validation accuracy of different methods using top 15 features.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
All	99.2	100	100	100	100	100	100	86.7	88.3	100	100	100	100	100
Brain	55.0	75.0	55.0	75.0	40.0	55.0	70.0	70.0	85.0	81.8	80.0	85.0	65.0	60.0
Breast1	70.0	60.0	64.3	57.1	61.4	57.1	58.6	52.9	50.0	60	65.7	57.1	55.7	61.4
Breast2	64.5	62.5	63.5	61.0	53.5	58.5	70.0	52.0	60.0	64.4	67.0	64.0	63.0	62.0
Carcinoma	97.5	100	97.5	100	97.5	100	97.5	57.5	75.0	98.9	97.5	97.5	97.5	97.5
CNS	61.7	51.7	58.3	51.7	38.3	63.3	51.7	58.3	60.0	58.1	61.7	60.0	56.7	68.3
Colon	78.3	85.0	85.0	81.7	76.7	78.3	78.3	70.0	73.3	83.1	81.7	83.3	85.0	80.0
DLBCL1	93.8	88.8	93.8	93.8	75.0	95.0	97.5	70.0	80.0	96.4	95.0	93.8	92.5	95.0
DLBCL2	92.5	92.5	85.0	90.0	97.5	92.5	87.5	57.5	62.5	93.9	82.5	92.5	90.0	92.5
GCM	86.1	82.1	75.0	77.9	76.8	81.4	79.3	71.8	76.4	81	75.0	82.9	83.9	84.3
GLI-85	83.3	84.4	78.9	82.2	81.1	82.2	87.8	66.7	84.4	88.1	78.9	83.3	86.7	80.0
GSE14333	55.9	48.6	56.4	50.5	50.9	47.3	46.4	53.2	57.3	54.1	54.5	56.4	43.2	50.5
GSE24514	97.5	92.5	95.0	95.0	92.5	95.0	92.5	80.0	96.4	90.0	95.0	95.0	95.0	97.5
GSE27854	48.0	53.0	42.0	55.0	50.0	57.0	48.0	39.0	52.0	44.4	47.0	52.0	43.0	50.0
Leukemia	92.9	95.7	91.4	95.7	94.3	95.7	95.7	61.4	70.0	98.5	94.3	94.3	95.7	92.9
Lung	98.8	99.4	96.5	98.8	98.8	99.4	98.2	84.1	95.9	100	98.8	98.2	97.6	99.4
Ovarian	100	97.6	98.0	98.8	97.6	98.4	99.6	73.6	76.4	99.4	98.8	98.4	100	98.0
Prostate1	94.0	87.0	86.0	88.0	93.0	92.0	93.0	58.0	65.0	90.4	95.0	91.0	87.0	92.0
Prostate2	75.6	75.6	70.0	67.8	70.0	76.7	78.9	58.9	54.4	77	67.8	75.6	73.3	77.8
Prostate3	85.7	87.9	72.1	80.0	83.6	88.6	92.9	55.0	63.6	83.5	80.7	84.3	80.7	87.1
SMK	68.9	70.6	65.0	63.3	63.3	67.8	71.1	60.0	59.4	73.6	65.6	67.2	62.8	72.8
SCBCT	98.0	100	98.0	98.0	100	100	100	74.0	72.0	100	100	100	98.0	100
Average	81.7	81.4	78.5	80.1	76.9	81.0	81.6	64.1	70.0	83	80.8	82.4	79.7	81.8

Table S5: 10-fold cross validation accuracy of different methods using top 25 features.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
All	100	100	100	100	100	100	100	90.8	93.3	100	100	100	100	100
Brain	75.0	60.0	75.0	65.0	55.0	65.0	80.0	60.0	70.0	86.8	70.0	90.0	65.0	55.0
Breast1	68.6	55.7	68.6	55.7	61.4	55.7	65.7	52.9	54.3	62.8	64.3	60.0	57.1	55.7
Breast2	65.0	63.5	59.5	63.5	56.0	63.5	69.0	54.0	61.5	67.4	64.0	64.0	70.5	63.5
Carcinoma	97.5	97.5	97.5	97.5	97.5	97.5	97.5	75.0	95.0	98.9	97.5	97.5	97.5	97.5
CNS	63.3	60.0	46.7	60.0	51.7	60.0	50.0	46.7	40.0	56.4	58.3	58.3	48.3	60.0
Colon	83.3	78.3	81.7	81.7	76.7	81.7	80.0	66.7	71.7	83.1	83.3	75.0	86.7	76.7
DLBCL1	95.0	93.8	95.0	93.8	85.0	90.0	97.5	73.8	83.8	96.4	96.2	96.2	96.2	91.2
DLBCL2	92.5	92.5	90.0	95.0	92.5	92.5	97.5	85.0	70.0	93.9	90.0	100	90.0	92.5
GCM	86.1	83.2	78.6	80.0	74.6	81.8	81.8	75.7	73.9	82.8	80.7	85.7	88.2	83.2
GLI-85	83.3	83.3	84.4	81.1	77.8	83.3	87.8	74.4	87.8	89.2	78.9	86.7	87.8	81.1
GSE14333	58.6	53.2	55.9	53.2	58.2	53.2	52.7	55.9	51.4	55.5	52.3	55.9	45.5	53.2
GSE24514	97.5	92.5	97.5	95.0	90.0	97.5	95.0	80.0	90.0	96.4	95.0	95.0	95.0	95.0
GSE27854	45.0	54.0	47.0	57.0	52.0	56.0	51.0	44.0	55.0	56.4	48.0	47.0	50.0	62.0
Leukemia	95.7	95.7	95.7	94.3	94.3	94.3	97.1	64.3	62.9	98.5	91.4	94.3	95.7	95.7
Lung	99.4	99.4	98.8	98.8	98.8	99.4	98.8	87.1	97.1	100	98.8	99.4	98.2	99.4
Ovarian	100	98.4	99.6	98.0	99.2	98.8	99.6	74.8	78.8	99.4	99.6	98.8	100	98.8
Prostate1	93.0	90.0	84.0	89.0	91.0	92.0	95.0	53.0	59.0	90.4	90.0	92.0	90.0	93.0
Prostate2	67.8	76.7	72.2	68.9	70.0	73.3	78.9	53.3	58.9	80.3	73.3	74.4	77.8	75.6
Prostate3	82.9	90.7	76.4	89.3	89.3	92.1	90.7	57.9	68.6	91.8	85.0	84.3	84.3	92.1
SMK	65.6	71.7	66.1	70.6	63.9	67.2	65.0	65.0	59.4	70.1	66.7	63.3	67.2	70.6
SCBCT	98.0	100	98.0	100	100	100	100	82.0	74.0	99.0	100	100	98.0	100
Average	82.4	81.4	80.4	81.2	78.9	81.6	83.2	66.9	70.7	84.4	81.1	82.6	81.3	81.4

Table S6: 10-fold cross validation accuracy of different methods using top 50 features.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
All	100	100	100	100	100	100	100	98.3	95.0	100	100	100	100	100
Brain	80.0	90.0	80.0	80.0	85.0	70.0	75.0	75.0	90.0	92	80.0	75.0	75.0	70.0
Breast1	65.7	51.4	65.7	51.4	64.3	51.4	64.3	54.3	60.0	64.5	68.6	65.7	61.4	51.4
Breast2	67.5	60.5	58.0	60.5	60.5	60.5	69.0	59.5	63.0	65.9	60.5	58.0	68.0	60.5
Carcinoma	97.5	97.5	97.5	97.5	97.5	97.5	97.5	80.0	97.5	98.9	97.5	97.5	97.5	97.5
CNS	61.7	60.0	48.3	60.0	55.0	60.0	53.3	53.3	60.0	56.4	60.0	56.7	50.0	60.0
Colon	83.3	83.3	81.7	83.3	78.3	83.3	81.7	71.7	73.3	76.4	83.3	76.7	83.3	83.3
DLBCL1	95.0	97.5	96.2	93.8	90.0	93.8	96.2	88.8	88.8	97.6	95.0	96.2	97.5	97.5
DLBCL2	90.0	90.0	95.0	90.0	92.5	90.0	97.5	87.5	67.5	98.9	97.5	97.5	92.5	90.0
GCM	84.3	83.9	78.6	80.0	77.1	83.6	83.6	76.8	75.7	85	82.5	82.1	91.4	84.3
GLI-85	86.7	83.3	84.4	82.2	86.7	86.7	84.4	74.4	83.3	88.1	84.4	85.6	87.8	83.3
GSE14333	58.6	53.2	50.0	53.2	52.7	53.2	50.9	50.0	56.8	53.7	53.6	51.8	50.0	53.2
GSE24514	95.0	92.5	97.5	97.5	90.0	95.0	95.0	77.5	95.0	96.4	92.5	95.0	95.0	92.5
GSE27854	47.0	49.0	59.0	49.0	53.0	51.0	50.0	43.0	54.0	53.4	40.0	40.0	43.0	51.0
Leukemia	95.7	97.1	98.6	94.3	92.9	94.3	95.7	71.4	80.0	97.1	94.3	95.7	95.7	95.7
Lung	99.4	98.8	99.4	98.8	98.8	98.8	98.8	92.9	98.2	100	98.8	99.4	98.8	98.8
Ovarian	100	99.6	100	99.6	99.2	99.6	100	75.6	76.0	100	99.6	99.2	100	99.6
Prostate1	93.0	92.0	87.0	93.0	88.0	92.0	94.0	61.0	68.0	98.2	94.0	94.0	92.0	89.0
Prostate2	75.6	74.4	70.0	73.3	67.8	68.9	74.4	66.7	55.6	74.3	78.2	75.6	74.4	70.0
Prostate3	83.6	91.4	80.7	90.7	87.9	92.1	92.9	70.7	67.9	85.7	90.0	83.6	89.3	91.4
SMK	64.4	65.0	65.6	64.4	61.7	66.7	63.9	63.3	59.4	68.1	66.1	62.2	67.8	67.2
SCBCT	100	100	98.0	100	98.0	100	100	92.0	90.0	100	100	100	100	100
Average	82.9	82.3	81.4	81.5	80.8	81.3	82.6	72.0	75.2	84.4	82.5	81.2	82.3	81.2

9 Detailed results using KNN

Similar to the previous section, the 10-fold cross validation accuracy of different methods using top 1, 2, 10, 15, 25 and 50 features are presented in Tables tables S7 to S12 for KNN classifier.

Table S7: 10-fold cross validation accuracy of different methods using top 1 feature.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
ALL	87.5	75.0	88.3	73.3	98.3	92.5	100	71.7	97.5	100	95.0	95.8	96.7	98.3
Brain	80.0	45.0	70.0	45.0	60.0	50.0	65.0	45.0	55.0	60.4	35.0	30.0	25.0	60.0
Breast1	62.9	51.4	58.6	57.1	55.7	65.7	52.9	57.1	60.0	50.4	61.4	45.7	50.0	55.7
Brest2	68.0	63.0	65.0	62.0	63.0	63.0	58.0	63.5	66.0	71.4	65.5	65.5	59.5	64.0
Carcinoma	85.0	77.5	85.0	60.0	82.5	92.5	75.0	65.0	92.5	87.9	80.0	85.0	80.0	70.0
CNS	71.7	61.7	63.3	56.7	61.7	56.7	63.3	61.7	46.7	82.1	56.7	58.3	58.3	68.3
Colon	78.3	68.3	66.7	61.7	70.0	78.3	80.0	48.3	66.7	82.1	73.3	75.0	68.3	73.3
DLBCL1	88.8	90.0	68.8	71.2	82.5	91.2	91.2	70.0	87.5	84.2	73.8	68.8	66.2	83.8
DLBCL2	95.0	82.5	62.5	42.5	80.0	57.5	95.0	35.0	85.0	90.4	87.5	65.0	82.5	75.0
GCM	77.5	71.8	70.7	69.6	75.4	69.6	67.5	66.4	72.5	76.5	73.6	75.4	67.5	72.9
GLI-85	66.7	66.7	70.0	65.6	68.9	71.1	84.4	65.6	74.4	73.7	72.2	78.9	65.6	75.6
GSE14333	55.5	53.2	57.3	54.5	56.4	59.5	52.7	53.2	57.7	52.2	59.1	56.8	52.3	57.7
GSE24514	92.5	82.5	77.5	70.0	87.5	92.5	85.0	92.5	85.0	92.9	82.5	75.0	85.0	85.0
GSE27854	51.0	51.0	47.0	46.0	49.0	40.0	48.0	48.0	52.0	44.4	51.0	52.0	48.0	47.0
Leukemia	92.9	81.4	70.0	70.0	94.3	87.1	82.9	67.1	90.0	94.7	88.6	88.6	91.4	95.7
Lung	90.0	94.1	84.1	83.5	91.8	90.0	94.7	94.7	94.7	92.8	92.9	92.9	94.7	92.9
Ovarian	76.4	88.4	67.2	62.8	89.6	88.0	88.8	64.8	89.6	86.0	89.6	86.8	89.6	89.6
Prostate1	86.0	72.0	54.0	52.0	86.0	79.0	74.0	63.0	88.0	79.4	80.0	87.0	78.0	85.0
Prostate2	76.7	57.8	61.1	56.7	64.4	68.9	61.1	48.9	75.6	76.0	60.0	63.3	52.2	64.4
Prostate4	78.6	84.3	52.9	51.4	82.9	64.3	60.7	63.6	84.3	74.0	82.9	62.1	60.7	82.9
SMK	57.2	55.6	53.3	57.8	65.0	57.2	61.1	58.3	59.4	54.3	52.2	57.2	58.9	61.7
SRBCT	90.0	96.0	94.0	60.0	92.0	94.0	92.0	60.0	90.0	90.4	90.0	84.0	90.0	92.0
Average	77.6	71.3	67.6	60.4	75.3	73.1	74.2	62.0	75.9	77.1	72.9	70.4	69.1	75.0
Sd	12.9	15.1	12.2	10.1	14.5	16.2	15.9	13.7	15.6	15.8	16.0	16.7	18.4	14.3

Table S8: 10-fold cross validation accuracy of different methods using top 2 feature.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
ALL	89.2	80.8	97.5	73.3	98.3	100	100	77.5	100	99.6	97.5	95.8	97.5	99.2
Brain	75.0	50.0	65.0	40.0	65.0	55.0	65.0	50.0	60.0	65.4	40.0	45.0	30.0	65.0
Breast1	64.3	52.9	72.9	51.4	57.1	65.7	54.3	47.1	60.0	57.5	60.0	51.4	48.6	55.7
Brest2	66.0	67.0	64.0	63.5	64.0	65.5	62.0	65.0	64.0	69.4	64.5	62.0	63.5	66.0
Carcinoma	87.5	87.5	95.0	65.0	92.5	92.5	85.0	70.0	95.0	97.9	85.0	85.0	85.0	80.0
CNS	73.3	61.7	61.7	53.3	61.7	58.3	60.0	60.0	55.0	65.4	61.7	55.0	50.0	68.3
Colon	90.0	71.7	65.0	55.0	76.7	91.7	80.0	61.7	68.3	85.4	81.7	76.7	80.0	80.0
DLBCL1	90.0	91.2	73.8	71.2	83.8	97.5	90.0	71.2	88.8	95.4	85.0	72.5	76.2	90.0
DLBCL2	95.0	82.5	70.0	60.0	82.5	80.0	95.0	55.0	87.5	92.9	95.0	70.0	95.0	82.5
GCM	80.4	75.7	69.6	69.3	77.5	76.1	68.6	66.8	70.4	79.7	77.9	74.6	66.4	76.8
GLI-85	76.7	80.0	68.9	72.2	83.3	74.4	84.4	65.6	76.7	79.3	74.4	80.0	71.1	80.0
GSE14333	54.1	49.5	54.1	60.5	54.5	53.6	55.5	54.5	55.9	52.2	56.8	50.5	55.0	55.5
GSE24514	90.0	90.0	95.0	75.0	82.5	97.5	92.5	85.0	95.0	90.4	90.0	87.5	92.5	87.5
GSE27854	54.0	59.0	48.0	44.0	39.0	46.0	46.0	51.0	52.0	48.4	51.0	55.0	51.0	44.0
Leukemia	90.0	80.0	68.6	64.3	91.4	92.9	78.6	77.1	90.0	90.4	85.7	90.0	91.4	92.9
Lung	89.4	97.1	87.1	81.8	96.5	97.1	95.9	97.1	94.7	93.9	96.5	95.3	97.1	97.1
Ovarian	81.6	87.6	79.2	69.2	88.8	98.0	88.4	64.0	96.8	86.0	88.4	93.6	88.8	90.4
Prostate1	91.0	79.0	67.0	62.0	90.0	91.0	82.0	60.0	92.0	86.4	90.0	84.0	83.0	89.0
Prostate2	73.3	61.1	64.4	50.0	67.8	66.7	58.9	57.8	75.6	74.8	68.9	70.0	54.4	67.8
Prostate4	85.7	85.7	61.4	58.6	86.4	89.3	70.0	67.1	89.3	76.8	85.0	72.1	68.6	85.7
SMK	62.2	57.2	57.8	55.6	70.6	65.0	66.7	53.9	62.2	63.2	61.7	62.2	57.2	67.8
SRBCT	96.0	96.0	92.0	56.0	90.0	98.0	90.0	60.0	94.0	94.4	90.0	88.0	90.0	90.0
Average	79.8	74.7	71.7	61.4	77.3	79.6	75.8	64.4	78.3	79.3	76.7	73.5	72.4	77.8
Sd	12.9	15.1	13.8	10.4	15.4	17.5	15.7	12.0	16.3	15.4	16.2	15.7	19.2	14.7

Table S9: 10-fold cross validation accuracy of different methods using top 10 feature.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
ALL	100	98.3	100	85.8	100	100	100	81.7	100	100	100	100	100	100
Brain	75.0	70.0	75.0	55.0	80.0	70.0	65.0	75.0	55.0	80.4	65.0	80.0	55.0	75.0
Breast1	68.6	57.1	55.7	60.0	61.4	55.7	62.9	48.6	57.1	59.0	55.7	58.6	55.7	60.0
Brest2	69.0	68.0	67.5	65.0	65.0	71.0	64.0	60.0	63.5	69.9	62.0	64.5	64.5	64.5
Carcinoma	100	97.5	87.5	55.0	97.5	97.5	95.0	75.0	97.5	97.9	97.5	97.5	97.5	97.5
CNS	58.3	63.3	63.3	53.3	65.0	58.3	68.3	65.0	60.0	67.1	61.7	55.0	65.0	66.7
Colon	86.7	81.7	85.0	73.3	86.7	88.3	90.0	65.0	83.3	88.7	85.0	86.7	85.0	85.0
DLBCL1	88.8	96.2	88.8	77.5	92.5	93.8	87.5	75.0	92.5	91.7	92.5	71.2	91.2	93.8
DLBCL2	95.0	85.0	82.5	57.5	90.0	95.0	92.5	52.5	92.5	92.9	92.5	92.5	90.0	90.0
GCM	79.6	81.1	75.7	72.5	80.4	83.2	71.8	75.0	79.6	82.5	81.4	75.7	76.1	82.1
GLI-85	85.6	78.9	83.3	73.3	86.7	77.8	84.4	72.2	82.2	86.0	84.4	84.4	75.6	88.9
GSE14333	54.1	59.1	53.6	60.0	58.6	54.1	52.7	52.3	54.1	55.9	53.2	53.6	56.8	57.7
GSE24514	97.5	97.5	97.5	75.0	95.0	95.0	92.5	92.5	95.0	97.9	95.0	95.0	95.0	92.5
GSE27854	53.0	50.0	51.0	48.0	40.0	49.0	47.0	49.0	52.0	49.4	49.0	49.0	51.0	41.0
Leukemia	92.9	92.9	90.0	74.3	92.9	97.1	92.9	75.7	91.4	93.3	90.0	91.4	92.9	92.9
Lung	91.8	97.6	91.8	82.9	97.6	98.2	98.2	95.3	97.6	98.0	98.2	97.6	97.6	97.6
Ovarian	92.8	99.2	96.0	70.8	96.0	100	92.4	72.8	99.6	98.0	95.6	96.4	93.6	96.4
Prostate1	90.0	91.0	73.0	57.0	89.0	90.0	87.0	66.0	90.0	92.4	90.0	90.0	90.0	90.0
Prostate2	78.9	76.7	53.3	55.6	75.6	74.4	67.8	58.9	67.8	77.1	75.6	77.8	66.7	77.8
Prostate4	83.6	87.1	67.1	57.1	88.6	93.6	80.7	66.4	94.3	81.1	86.4	80.7	78.6	88.6
SMK	66.1	65.0	61.7	63.3	71.7	67.8	69.4	61.1	65.6	68.2	70.6	62.2	61.7	68.9
SRBCT	98.0	98.0	92.0	62.0	96.0	98.0	96.0	52.0	96.0	96.4	96.0	96.0	96.0	96.0
Average	82.1	81.4	76.9	65.2	82.1	82.2	79.9	67.6	80.3	82.9	80.8	79.8	78.9	81.9
Sd	14.9	15.7	15.7	10.4	15.8	16.9	15.6	13.0	17.3	15.3	16.4	16.4	16.6	15.9

Table S10: 10-fold cross validation accuracy of different methods using top 15 feature.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
ALL	100	99.2	99.2	90.0	100	100	100	81.7	100	100	100	100	100	100
Brain	70.0	60.0	65.0	65.0	75.0	65.0	75.0	65.0	50.0	85.4	65.0	80.0	65.0	75.0
Breast1	68.6	60.0	60.0	58.6	57.1	62.9	60.0	58.6	62.9	61.8	62.9	55.7	62.9	60.0
Brest2	67.0	67.0	67.0	64.5	65.5	73.5	67.0	65.5	64.5	67.4	66.0	59.5	67.0	67.5
Carcinoma	97.5	97.5	97.5	57.5	97.5	97.5	97.5	80.0	97.5	100	97.5	97.5	97.5	97.5
CNS	58.3	56.7	56.7	51.7	71.7	58.3	66.7	60.0	61.7	63.7	68.3	58.3	65.0	68.3
Colon	88.3	81.7	85.0	73.3	88.3	88.3	90.0	75.0	81.7	90.4	88.3	88.3	90.0	88.3
DLBCL1	92.5	96.2	91.2	72.5	92.5	92.5	88.8	76.2	95.0	94.2	91.2	70.0	91.2	93.8
DLBCL2	97.5	90.0	92.5	62.5	92.5	97.5	92.5	67.5	90.0	95.4	97.5	95.0	90.0	92.5
GCM	82.1	82.5	77.1	74.3	80.7	84.3	78.9	74.6	76.8	80.4	81.4	75.7	78.2	80.0
GLI-85	82.2	80.0	78.9	70.0	86.7	82.2	82.2	74.4	81.1	83.7	82.2	81.1	80.0	87.8
GSE14333	58.6	58.2	55.5	55.0	58.6	50.9	57.3	50.0	52.7	53.6	50.0	50.5	53.6	58.6
GSE24514	95.0	97.5	97.5	77.5	95.0	95.0	97.5	90.0	95.0	95.4	97.5	95.0	95.0	92.5
GSE27854	55.0	44.0	49.0	44.0	52.0	47.0	50.0	55.0	48.0	44.4	43.0	59.0	50.0	44.0
Leukemia	94.3	94.3	92.9	70.0	94.3	95.7	94.3	77.1	95.7	94.7	92.9	94.3	94.3	94.3
Lung	91.8	97.6	92.9	83.5	97.6	97.6	97.6	95.3	97.6	99.2	98.2	97.6	97.6	97.6
Ovarian	95.6	99.2	96.4	73.6	96.8	99.2	95.6	74.8	99.6	97.2	96.4	96.8	96.4	97.6
Prostate1	90.0	86.0	74.0	57.0	91.0	89.0	87.0	66.0	91.0	91.4	90.0	92.0	90.0	92.0
Prostate2	80.0	77.8	55.6	53.3	73.3	81.1	73.3	51.1	75.6	79.3	76.7	74.4	67.8	75.6
Prostate4	80.0	82.1	70.0	57.1	87.9	93.6	82.9	65.0	90.7	83.3	86.4	83.6	84.3	89.3
SMK	65.6	66.7	67.8	63.9	71.1	67.2	69.4	60.6	65.0	62.6	67.8	66.7	65.0	67.2
SRBCT	96.0	98.0	100	58.0	96.0	98.0	98.0	48.0	98.0	96.4	96.0	96.0	96.0	96.0
Average	82.1	80.6	78.3	65.1	82.8	82.6	81.9	68.7	80.5	82.8	81.6	80.3	80.8	82.5
Sd	14.6	17.0	16.9	11.1	14.8	16.9	15.1	12.5	17.7	16.6	16.8	16.3	15.9	15.7

Table S11: 10-fold cross validation accuracy of different methods using top 25 feature.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
ALL	100	99.2	100	95.0	100	100	100	85.8	100	100	100	100	100	100
Brain	75.0	65.0	80.0	75.0	70.0	60.0	75.0	50.0	60.0	85.4	55.0	80.0	65.0	70.0
Breast1	65.7	60.0	54.3	60.0	57.1	62.9	62.9	52.9	67.1	63.3	57.1	61.4	57.1	57.1
Brest2	68.5	63.0	64.0	62.0	65.0	73.0	66.5	68.0	68.0	66.4	65.0	67.5	65.0	65.0
Carcinoma	97.5	95.0	97.5	75.0	97.5	97.5	97.5	87.5	97.5	100	100	97.5	100	97.5
CNS	61.7	65.0	55.0	60.0	63.3	58.3	65.0	63.3	63.3	63.7	65.0	58.3	63.3	63.3
Colon	90.0	86.7	86.7	71.7	88.3	88.3	90.0	76.7	86.7	88.7	88.3	88.3	88.3	88.3
DLBCL1	91.2	95.0	92.5	82.5	91.2	93.8	88.8	72.5	88.8	95.4	91.2	70.0	92.5	91.2
DLBCL2	97.5	95.0	95.0	75.0	95.0	97.5	95.0	65.0	90.0	97.9	95.0	97.5	92.5	95.0
GCM	81.8	82.5	79.3	76.1	82.5	84.6	78.9	68.6	78.9	82.2	81.1	75.7	80.4	83.9
GLI-85	82.2	80.0	80.0	76.7	85.6	86.7	83.3	76.7	83.3	81.5	84.4	80.0	83.3	86.7
GSE14333	53.6	56.4	54.5	54.5	55.0	55.5	50.5	50.0	55.0	53.1	55.0	54.1	55.0	55.0
GSE24514	95.0	95.0	92.5	77.5	95.0	97.5	95.0	87.5	95.0	97.9	95.0	97.5	95.0	95.0
GSE27854	52.0	48.0	49.0	46.0	51.0	46.0	43.0	52.0	51.0	55.4	51.0	60.0	47.0	50.0
Leukemia	92.9	97.1	95.7	74.3	94.3	94.3	94.3	60.0	94.3	93.3	92.9	92.9	94.3	94.3
Lung	91.8	98.2	94.7	82.4	97.6	98.8	97.1	95.3	96.5	99.2	97.6	98.2	95.3	97.6
Ovarian	93.2	100	99.2	70.0	96.4	99.6	97.6	76.0	100	98.4	96.4	98.4	96.4	97.2
Prostate1	91.0	90.0	80.0	65.0	87.0	89.0	91.0	55.0	87.0	91.4	92.0	92.0	89.0	91.0
Prostate2	77.8	76.7	62.2	62.2	76.7	81.1	73.3	51.1	70.0	78.2	81.1	80.0	71.1	76.7
Prostate4	84.3	77.1	73.6	58.6	85.7	92.1	84.3	65.0	92.9	86.8	85.0	88.6	89.3	85.0
SMK	66.7	70.0	68.3	63.3	69.4	70.0	71.7	63.9	66.1	66.5	67.8	67.8	68.3	70.0
SRBCT	98.0	98.0	98.0	58.0	98.0	98.0	98.0	58.0	100	96.4	98.0	100	98.0	98.0
Average	82.2	81.5	79.6	69.1	81.9	82.9	81.8	67.3	81.4	83.7	81.5	82.1	81.2	82.2
Sd	14.9	16.3	17.0	11.2	15.8	16.9	16.4	13.6	16.0	15.6	16.7	15.5	16.6	16.0

Table S12: 10-fold cross validation accuracy of different methods using top 50 feature.

Dataset	Methodes													
	CI	CS	CFS	GR	IA	IG	k-TSP	mRMR	oneR	PSI	RF	SAM	SVM-R	SU
ALL	100	99.2	100	95.8	100	100	100	87.5	100	100	100	100	100	100
Brain	75.0	70.0	75.0	70.0	70.0	75.0	80.0	50.0	75.0	90.4	45.0	90.0	60.0	65.0
Breast1	62.9	64.3	64.3	57.1	55.7	60.0	65.7	57.1	68.6	61.8	55.7	60.0	55.7	55.7
Brest2	66.5	65.5	63.5	63.0	67.0	68.5	65.0	68.0	65.5	67.4	67.0	66.0	67.0	67.0
Carcinoma	97.5	97.5	97.5	75.0	97.5	97.5	97.5	92.5	97.5	97.9	97.5	97.5	97.5	97.5
CNS	63.3	65.0	56.7	55.0	63.3	63.3	70.0	60.0	61.7	68.7	63.3	61.7	63.3	63.3
Colon	88.3	86.7	88.3	73.3	88.3	90.0	88.3	73.3	81.7	88.7	88.3	90.0	88.3	88.3
DLBCL1	93.8	96.2	95.0	81.2	92.5	93.8	92.5	81.2	90.0	95.4	92.5	85.0	92.5	93.8
DLBCL2	97.5	92.5	97.5	67.5	92.5	100	97.5	75.0	92.5	100	92.5	97.5	92.5	92.5
GCM	81.8	81.8	79.3	75.7	81.8	85.0	80.0	73.2	79.6	82.5	83.6	75.7	78.9	83.6
GLI-85	85.6	84.4	85.6	75.6	88.9	83.3	88.9	83.3	87.8	81.5	88.9	82.2	87.8	87.8
GSE14333	55.5	53.2	55.5	55.0	57.7	55.9	57.7	51.8	55.5	57.2	57.7	57.7	57.7	57.7
GSE24514	97.5	97.5	97.5	87.5	95.0	95.0	95.0	90.0	95.0	97.9	95.0	97.5	95.0	95.0
GSE27854	45.0	47.0	49.0	45.0	43.0	62.0	40.0	57.0	51.0	53.4	47.0	61.0	42.0	43.0
Leukemia	92.9	98.6	92.9	77.1	92.9	94.3	92.9	62.9	94.3	93.3	92.9	91.4	92.9	91.4
Lung	92.4	97.1	97.6	82.9	94.7	99.4	93.5	91.8	98.2	97.5	94.7	97.6	94.7	95.9
Ovarian	92.8	99.6	99.2	71.2	98.0	100	97.6	77.6	100	98.0	97.6	97.6	97.6	98.4
Prostate1	90.0	90.0	86.0	65.0	90.0	88.0	90.0	52.0	88.0	90.4	91.0	89.0	89.0	89.0
Prostate2	77.8	71.1	71.1	70.0	76.7	78.9	80.0	48.9	74.4	79.3	78.9	80.0	75.6	76.7
Prostate4	85.7	77.9	89.3	62.9	85.0	91.4	87.9	60.0	88.6	86.8	84.3	88.6	87.1	86.4
SMK	64.4	68.3	68.9	66.1	70.0	71.1	70.0	62.8	67.8	66.5	68.9	70.0	68.3	69.4
SRBCT	100	100	98.0	66.0	98.0	100	98.0	64.0	100	100	100	100	100	98.0
Average	82.1	82.0	82.2	69.9	81.8	84.2	83.1	69.1	82.4	84.4	81.0	83.5	81.1	81.6
Sd	16.0	16.5	16.5	11.6	16.3	14.9	15.7	14.2	15.3	15.3	17.7	14.6	17.1	16.6

10 references

References

- [1] Duda, R. O. *et al.* (2012) Pattern classification. *John Wiley & Sons*.
- [2] Shi, P. *et al.* (2011) Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. *Bmc Bioinformatics*, **12**(1), 375.
- [3] Bauer, D. F. *et al.* (1972) Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, **67**(339), 687-690.