Iterative feature perturbation method as a gene selector for microarray data

Juana Canul-Reich, Lawrence Hall, Dmitry Goldgof and Steven Eschrich

Abstract—Gene expression microarray datasets often consist of a limited number of samples with a large number of expression measurements, usually on the order of thousands of genes. Therefore, dimensionality reduction is a core process prior to any classification task. In this work, the iterative feature perturbation method (IFP), an embedded gene selector, is introduced and applied to 4 cancer microarray datasets: colon cancer (cancer vs. normal), leukemia (subtype classification), Moffitt colon cancer (prognosis predictor) and lung cancer (prognosis predictor). We compared results obtained by IFP to those of SVM-RFE and the t test using a linear support vector machine as the classifier in all cases. The IFP approach resulted in comparable or superior accuracy as compared to SVM-RFE in 3 of the 4 datasets. Surprisingly, the simple t test feature ranking typically produced classifiers with the highest accuracy across the 4 datasets. This finding led to additional experiments to incorporate an upfront preselection of the top 200 genes based on their p values on each dataset. Then we applied IFP and SVM-RFE on the resulting smaller datasets. The accuracy results after preselection showed up to 3% performance improvement for both IFP and SVM-RFE across the 4 datasets; they got closer to the t test accuracy and outperformed it at some points. An AUC analysis and a statistical analysis (using the Friedman/Holm test) of the accuracy curves under both scenarios corroborated the superiority of the t test on experiments without gene preselection, and the performance improvement of IFP and SVM-RFE with gene preselection. Also, we investigated the percentage of intersection between the gene sets selected by the 3 methods across the 4 datasets, and found that it was low even at points where any 2 methods reached very similar accuracies. For example on the colon cancer dataset using the entire set of features, both IFP and SVM-RFE showed around 82% accuracy in the range of 70 through 200 features, while their feature intersection in this range was 60%. We observed similar patterns across our 4 datasets and concluded that the same or similar accuracies can be obtained with different sets of genes.

Index Terms—Microarray data, classification, support vector machines, embedded methods, feature ranking, feature selection

1 Introduction

Gene expression microarray datasets tend to be small in sample size due to the cost associated with the assays. Typically, there are many more gene expression measurements (e.g. 54,000 transcripts) available than samples. Hence, the selection of a subset of genes/features is crucial before building a classifier. Identifying a small number of genes that are good predictors is important from a biological standpoint, as expression experiments are typically performed to generate hypotheses for further experimentation in the lab. For clinical applications, identifying a small number of genes that are important in predicting patient survival time or diagnosing cancer can speed the translation of expression signatures into cost-effective tests for clinical practice. From a machine learning viewpoint, too many features/genes in a dataset can negatively influence the classification performance as they increase the possibility of overfitting. Therefore, the feature selection process plays a vital role for the building of a successful classifier from microarray datasets.

An initial version of the feature perturbation method (FP) was introduced in [1]. In this paper, we introduce the iterative feature perturbation method (IFP), which overcomes FP esentially by including a self tune mechanism to determine the amount of noise needed to perturb the features to find the most sensitive ones. IFP is an embedded selector with the capability of using any classification algorithm as the base classifier.

Our IFP implementation used a support vector machine (SVM) to allow for performance comparison with SVM-RFE [2]. IFP is more general than SVM-RFE which is limited to the use of SVM as the base classifier. On the other hand, the t test feature ranking which is based on the p values of the features, was compared, in terms of SVM accuracy to IFP and SVM-RFE. The t test has been previously used to select genes [3], [4], but we expected the more sophisticated approaches to result in higher accuracy classifiers. We experimented with 4 microarray datasets: colon cancer (cancer vs. normal), leukemia (subtype classification), Moffitt colon cancer (prognosis predictor) and lung cancer (prognosis predictor). The results of the experiments indicated that IFP results in accuracy comparable and even superior for some sets of genes to that of SVM-RFE, on 3 out of 4 datasets. Surprisingly, the results showed that the t test based feature selector finds better sets of genes than the more sophisticated IFP and SVM-RFE selection methods. The t test feature selector resulted in the highest SVM classifier accuracy for the largest number of gene sets

- *J. Canul-Reich, L. Hall and D. Goldgof are with the Department of Computer Science and Engineering, University of South FLorida, Tampa, Fl, 33620.*
  E-mail: jcanulre,hall,goldgof@cse.usf.edu
- *S. Eschrich is with the Department of Biomedical Informatics, Moffitt Cancer Center and Research Institute, Tampa, Fl, 33612.*
  E-mail: steven.eschrich@moffitt.org

(gene sets consist of different numbers of genes) across all 4 datasets. This finding led to additional experiments to incorporate a preselection of the top 200 genes based on their p values on each dataset, and then apply IFP and SVM-RFE on the resulting smaller datasets.

The accuracy results after preselection showed up to 3% performance improvement for both IFP and SVM-RFE across the 4 datasets; they got closer to the t test accuracy and outperformed it at some points. An AUC analysis of the accuracy curves under both scenarios, without and with gene preselection, corroborated the superiority of the t test on experiments without gene preselection, and the performance improvement of IFP and SVM-RFE with gene preselection. The statistical significance analysis using the Friedman/Holm test with a 5% threshold performed on the top 50 features under both scenarios, indicated that without gene preselection the t test is better in accuracy than both IFP and SVM-RFE across our 4 datasets mostly throughout their entire set of genes. With gene preselection, the statistical analysis showed the significant performance improvement of both IFP and SVM-RFE as an effect of the preselection. We also looked at the statistical significant difference between the two approaches using both IFP and SVM-RFE. The results indicated that the preselection of features was statistically significant better for some feature sets on the colon cancer, leukemia, and lung cancer datasets.

Also, we investigated the percentage of intersection between the gene sets selected by the 3 methods across the 4 datasets, and found that it was low even at points where any 2 methods reached very similar accuracies. For example on the colon cancer dataset using the entire set of features, both IFP and SVM-RFE showed around 82% accuracy in the range of 70 through 200 genes/features, while their intersection in this range was 60%. We observed similar patterns across our 4 datasets.

2 Related methods

Filters, wrappers and embedded methods are three approaches for feature selection [5]. Filters do not incorporate the learning algorithm in the feature subset search, basically they select features based on a measure/score individually obtained on each feature (univariate case). Low-scored features are then removed. With wrappers, the feature subset search incorporates the learning algorithm to asses diverse feature subsets, the subset resulting with the highest assessment gets chosen [6]. Embedded methods incorporate the learning algorithm in the feature subset search [7], so the search is guided by the learning algorithm.

Embedded feature selection methods used in the microarray domain have applied learning algorithms such as random forests, SVM and logistic regression (the latter two use the weights as a selection criterion) [5].

A method using SVM is the recursive feature elimination for support vector machines (SVM-RFE) introduced in [2]. It is an embedded selector that follows a backward elimination approach. It ranks the features according to their weights, which are calculated from the support vectors. The features to be removed next are those with the lowest weights. SVM-RFE works only with support vector machines.

A hybrid huberized support vector machine (HHSVM) was introduced in [8] for both classification and gene selection. HHSVM uses a combination of the huberized hinge loss function to measure misclassification and the elastic-net penalty which allows for automatic variable selection and grouping effect (groups of correlated variables get selected/removed together). The authors show results of the HHSVM on the leukemia dataset [9] compared to those of SVM-RFE. In the context of the original split of the dataset, 38 training samples and 34 testing samples, SVM-RFE got 2/34 errors with 128 genes, and 0/38 errors with 128 genes in a cross validation environment. Similarly, HHSVM got 0/34 errors with 84 genes in the original split and 0/38 errors in the cross validation context. Also, they showed results for experiments conducted under a randomly-splitting approach, where they combined all the original training/testing samples altogether and made random splits into 38 training and 34 testing samples; they repeated this process 50 times and reported average of results. SVM-RFE showed an average testing error of 2.25% with 256 genes on average; while HHSVM showed an average testing error of 1.67% with 87.9 genes on average.

In [10], the authors showed results of HHSVM on the colon cancer dataset [11] compared to those of SVM-RFE. The dataset was randomly split for 100 times into 42 training samples (27 cancer samples and 15 normal tissues) and 20 testing samples (13 cancer samples and 7 normal tissues). HHSVM resulted in 12.69% of test error with 94.5 genes, while SVM-RFE showed a test error of 17.10% with 64 genes.

3 Method: Iterative Feature Perturbation

The iterative feature perturbation method (IFP) inherited some concepts from its ancestor FP, such as the backward elimination approach and the definition of relevant/non-relevant features. Both methods are embedded feature selectors. As such, the base learning algorithm is involved in the process of determining which features are going to be removed in the next step. Both algorithms start with the entire set of features in the dataset; at every iteration the size of the feature set is reduced by removing the least important features. The criterion to determine which features are the least important relies on the impact on the classification performance that each feature has when perturbed. That is, each feature is perturbed by adding noise to it. If as a result, it leads to a big change in the classification performance, then the feature is considered relevant. Correspondingly, non-relevant features will cause little or no impact to the classification performance. Non-relevant features are then removed so that only relevant features remain.

In [12] we concluded that different amounts of noise were needed to adequately perturb feature sets of different sizes. The new IFP algorithm is described in Fig. 1. Fig. 1a describes the main iterative part of IFP, and Fig. 1b describes the binary search called from Fig. 1a for calculation of the noise level, perturbation and ranking of features. IFP receives as input the original dataset which constitutes the training set *X* and a *K* value indicating the number of features to be removed in the current iteration. At the beginning the subset of current surviving features *S* is set to the initial set of features. The method iterates through stages until no features remain; then a ranked feature list is output. For performance evaluation purposes, a classification model can be created based on a selected feature set from the ranked list, and its accuracy can be calculated on the test set.

In the iterative process, the first stage is to train a classification model on *X* with all existing surviving features *S*. Any classification learning algorithm could be used to create the classification model. For the experiments conducted in this research, a support vector machine (SVM) was used as the base classifier. For performance reasons, after training the classification model, the training set was reduced to the subset of samples selected as the support vectors. These samples carry the essential information needed for the classification problem; the rest of the samples are irrelevant [13], [14]. In the second stage the training accuracy for the reduced samples set was calculated. The third stage perturbs and ranks all features in *S* from lowest to most relevant. This phase is performed using a binary search process, which identifies an appropriate amount of noise to be injected in each feature, and identifies exactly the *K* least relevant features desired for removal. The ranking for a feature is determined by the change in accuracy observed on the training samples before and after adding noise. Non-relevant features were defined as either those causing no change in accuracy or a $0-10\%$ range of change. The binary search process returns a ranking of features which is examined for ties in the fourth stage of IFP. Two features are tied when they cause the same accuracy change. If no tie is found, then the top *K* features in the ranking, which are the *K* least relevant features of the current set *S*, are removed. In case of a tie, all tied features are ranked based on a tie-breaking criterion, and only the top feature is removed. Since SVM is the base classifier used, the weight of each feature [15] was chosen as the tie-breaking criterion, which is calculated following (1):

$$\mathbf{w} = \sum_{i=1}^{l} y_i \alpha_i \mathbf{x}_i \tag{1}$$

where l is the number of support vectors in the SVM model, y_i is the label/class (+1/-1) for the i^{th} support vector, α_i is a positive real value given by the SVM model to the i^{th} support vector indicating its contribution to the margin, and x_i is the gene/feature value in the i^{th} support vector. Given that microarray datasets usually have many less examples than gene expressions, the feature removal process would be computationally expensive if features were removed one at a time. Moreover, only a small subset of features is expected to be relevant for classification. This is the reason that an adaptive feature elimination strategy was applied in our experiments.

The strategy consists of determining the number of features to be removed in relation to the current number of surviving features. Specifically, half the surviving features were removed at a time rather than just one when the number of existing features was larger than a given threshold; thereafter one-at-a-time feature removal was initiated. After removing the *K* non-relevant features, the final feature ranking *F* needed to be updated.

An iterative binary search was implemented to find the amount of noise to be added to each feature when perturbing. Noise is generated following (2):

$$Noise_i = (c * sd_i) \tag{2}$$

where sd_i is the standard deviation of the feature being perturbed across all examples in the training set, and *c* is a dynamic factor indicating the noise level being injected in the perturbation process. It is dynamic in the sense that it varies in magnitude for different sizes of the set of surviving features *S*. That is, the *c* factor for a set *S* with 500 features may be different than that for a set *S* with 10 features. The *c* factor impacts the final amount of noise injected into each feature. On the other hand, when a large amount of noise is applied to the set *S*, it may result in no features within the $0-10\%$ accuracy change. The opposite is also true. When a small amount of noise is applied, it may result in getting more features within the $0-10\%$ range than needed. The more noise applied to a feature, the bigger the accuracy change it will cause. Specifically, the binary search looks for a *c* factor such that when applied in (2), it finds exactly the *K* least important features to remove.

Binary search is an algorithm that locates its target value in the middle of a sorted list. In our implementation, the *c* factor can take a minimum value of 1×10^{-6} and can be as large as needed. This fact allows the search space to be sorted in ascending order. The initial value of the maximum boundary of the search space is set to a very large maximum value. As the search advances, the search space will be reduced to either its left half or right half depending on the need to increment or decrement the number of features within the $0-10\%$ accuracy change respectively. Eventually, the *c* factor being sought will be the value in the midpoint of the search space.

In Fig. 1b, line 1 sets *adjusting_variable* to the lowest value that the *c* factor can take. This variable is used when reducing the search space to its left or right half in lines 16 and 18. The minimum value that *adjusting_variable* is assigned allows for a further scan of the search space. Lines 2 and 3 set the minimum and maximun boundary values of the search space respectively. The perturbation of features is coded in lines 6 through 12. The amount of noise that is injected to each feature in line 7 is calculated with (2). In line 10, Acc_i corresponds to the accuracy obtained after perturbing feature *i* of the set *S*. In line 11, the ranking criterion is calculated; it is the difference of the training accuracy and the accuracy calculated in line 10. In line 13, all features in *S* are ranked based on the ranking criterion; that is, the resulting ranking will go from features causing the least change in accuracy down to features changing accuracy the most. In line 14, the features with $0-10\%$ of accuracy change are counted. In lines 15-16, when there are no features within this range of accuracy change or they total less than the *K* features needed, then the search space is reduced to the left half, meaning that the *c* factor and thus the noise level has to be decreased in order to get more features within the desired range of accuracy change. On the other hand, in lines 17-18, when there are more features within the $0-10\%$ of accuracy change than needed, then the search space is reduced to the right half, meaning that the *c* factor and thus the noise level have to be increased in order to get fewer features within the desired range of accuracy change. Finally, in line 20, when exactly the *K* features needed are found, the ranking of all features is returned by the binary search.

There are a few differences between the new IFP method and the old FP. First, a self-tuning mechanism to determine the amount of noise to be applied in the perturbation phase is now incorporated. Second, the randomness in the calculation of the amount of noise was removed. Third, after training a classifier the training set is reduced to the subset of support vectors.

4 Experimental studies

4.1 Data and preprocessing

Experiments were performed on 4 affymetrix-platform 2-class gene expression microarray datasets. All 4 datasets underwent a preprocessing phase as is typical for this type of data. Data preparation allows for the learning algorithm to easily access the information carried by the datasets [16].

- The Colon cancer dataset is a well-studied publicly available microarray benchmark [11]. It is made up of 62 samples including 22 normal and 40 colon cancer tissues. There are 2000 gene expression values for each sample. For data preprocessing, a $\log_2$-transformation was applied to it to normalize the data.
- The leukemia dataset is another publicly available dataset [9]. It contains information on human acute myeloid (AML) and acute lymphoblastic leukemia (ALL) with 25 and 47 samples respectively. There are 7129 gene-expression values for each sample. The dataset was explored and a large number of negative gene-expression values were found. For data preprocessing, all negative gene values was set to 1 and a $\log_2$-transformation was applied to the entire dataset, resulting in a large number of zero values. A set of 2689 genes were preselected under the criteria of genes having $< 25\%$ of zero values and variance $>= 1$.
- The Moffitt colon cancer dataset used in this paper is a superset of the set described in [3]. It contains information on 122 samples, 84 labeled "good prognosis" representing patients with survival time $>= 36$ months, and 38 samples labeled "poor prognosis" representing patients with survival time $<$ 36 months. The original dataset has 54675 genes. For data preprocessing, a $\log_2$-transformation was done and a preselection of genes using the criteria of genes having variance $>= 0.5$ resulted in a subset of 2619 genes.
- The lung cancer dataset used in our experiments is the same as was used in [17]. It is composed of 410 samples, 271 labeled "good prognosis", and 139 labeled "poor prognosis" relative to their survival time as described for the Moffitt colon dataset. The original dataset has 22282 genes, 68 control genes were left out, and a subset of 22214 genes remained. The dataset was explored and a number of genes with values close to zero were found. For data preprocessing, all gene-expression values < 2 were set to 2, so they would not result in negative numbers in the $\log_2$-transformation. A subset of 2428 genes were preselected under the criteria of genes having $< 25\%$ of gene values = 1 and variance $>= 1$.

Finally, gene expressions in all 4 datasets were scaled to a 0 - 1 range.

4.2 Parameters for the SVM

As stated in the description of IFP in section 3, the IFP method could be used with any classification algorithm as the base classifier. In our implementation, we used SVM as the base classifier to be able to compare our results against those of SVM-RFE. The SVM used is a modified version of libSVM [18]. A linear kernel was used with parameter $C = 1$ to reduce training time and the probability of overfitting. The optimization algorithm used was the sequential minimal optimization (SMO).

4.3 Performance measure

All results reported in Section 4.4 are expressed in weighted accuracy rather than total accuracy. Weighted accuracy was preferred as the classifier performance measure due to the unequal distribution of the two classes in all four datasets. In situations like these, weighted accuracy gives a better performance estimate of the learning algorithm [19]. Weighted accuracy for 2-class datasets is defined in Eq. (3),

$$Weighted\ Accuracy = \left(\frac{tp}{tp+fn} + \frac{tn}{fp+tn}\right)/2 \qquad (3)$$

```
Input: A set $X$ of training samples and a
    $K$ value indicating the number of features
    to be removed.
Output: $F$, the final feature ranking.

    $S$ : Subset of surviving features
$S \leftarrow$ all features
while $S \neq \emptyset$ do
    Train a classifier on $X$ with features $S$
    Calculate the training set accuracy $Acc$
    Perturb and rank features by binary search
    Determine if there are tied features
    if no tied features then
        $S \leftarrow S - \{top\ K\ features\}$
    else
        Rank tied features by weight
        $S \leftarrow S - \{top\ feature\}$
    end if
    Update feature ranking $F$
end while
return $R$
```

(a)

```
Input: The current set $X$ of training samples
    with $S$ surviving features, the $K$ value and
    the training accuracy $Acc$ of current set $X$.
Output: $FR$, ranking of current features in $S$

adjusting_variable $\leftarrow 1 \times 10^{-6}$
$min \leftarrow 1 \times 10^{-6}$
$max \leftarrow$ large maximum value
while $min <= max$ do
    c_factor $\leftarrow$ midpoint $\leftarrow (min + max)/2$
    for all features $i$ in $S$ do {Perturbing features}
        add noise to $i$ across all examples
        get a new perturbed dataset $X'$
        predict classes for $X'$
        calculate accuracy $Acc_i$
        calculate ranking criterion $r_i = abs(Acc - Acc_i)$
    end for
    $FR \leftarrow$ ranking of all features in $S$ by $r_i$
    count $\leftarrow$ features with $0 - 10\%$ ranking criterion
    if count $= 0$ or count $< K$ then {c_factor needs to be decreased}
        $max \leftarrow$ midpoint $-$ adjusting_variable
    else if count $> K$ then {c_factor needs to be increased}
        $min \leftarrow$ midpoint $+$ adjusting_variable
    else {right c_factor found}
        return $FR$
    end if
end while
```

(b)

Fig. 1. (a) is the main loop of the iterative feature perturbation algorithm and (b) the binary search called from (a) for calculation of noise level/c_factor, perturbation and ranking of features.

TABLE 1
Confusion matrix

	Predicted Cancer	Predicted Normal Tissue
Actual Cancer	True Positive (TP)	False Negative (FN)
Actual Normal tissue	False Positive (FP)	True Negative (TN)

where *tp*, *fp*, *tn*, and *fn* respectively are the number of true positives, false positives, true negatives and false negatives in a confusion matrix, as shown in Table 1.

4.4 Experiments and results

The adaptive feature elimination strategy described in section 3 was applied in the feature removal process for all experiments conducted that started with the entire set of features in the dataset. Specifically, 50% of the existing features were removed across iterations until a threshold of 25% was reached, except for the colon cancer dataset whose threshold was set to 10%. All experiments were assessed via a 10-fold cross validation process, and each cross-validation experiment was performed 5 times with different random seeds. The average weighted accuracy over the five runs is reported.

Three methods of feature ranking were assessed with all four datasets: IFP, SVM-RFE and the t test. The t test ranks the features according to their p values. The most relevant feature has the lowest p value and the least relevant feature has the highest p value. Results on the four datasets are shown in Fig. 2. The graph of the colon cancer dataset in Fig. 2a shows that IFP resulted in more accurate classifiers than SVM-RFE throughout most of the range of 6 to 139 features (genes), with a bigger difference in favor of IFP in the ranges 81 through 95, 53 through 68 features and 6 through 42 features. SVM-RFE resulted in more accurate classifier than IFP throughout most of the range of 140 to 2000 features. The p value based feature ranking resulted in the highest accuracy classifier for nearly the entire set of features on this dataset.

The leukemia dataset, graph in Fig. 2b shows that the t test was more accurate than IFP and SVM-RFE throughout the range of 1000 to 2000 features. IFP resulted in a more accurate classifier than both SVM-RFE and the t test throughout the range of 100 to 250 features. The three accuracies were comparable and we would say there was a tie in the range of 57 through 100 features. However there were differences in accuracies in the range of 1 through 56 features. IFP alternated with SVM-RFE in accuracy spikes in this range although IFP overall attained higher accuracies. On the other hand, the t test feature ranking resulted in the highest accuracy classifier throughout the range of 3 to 24 features, and it was the least accurate classifier in the range of 31 to 56 features.

The graph of the Moffitt colon dataset in Fig. 2c, shows comparable accuracies between the three methods throughout the range of 600 to 2619 features. IFP spiked at 500 features and thereafter dropped off to the lowest accuracy of the three methods, mostly throughout the range of 33 to 250 features. The t test clearly resulted in a more accurate classifier throughout the range of 116 to 250 features, whereas SVM-RFE and the t test showed similar accuracies in the range of 80 through 115 features. SVM-RFE was better than the t test throughout most of the range of 48 to 79 features, and the t test was better than SVM-RFE in the range of 33 through 46 features. Last, there were interesting accuracies in the top 33 features. SVM-RFE and IFP showed similar accuracies in this range except that IFP accuracy spiked around 25 features. On the other hand, the t test showed its lowest accuracy in this range at 24 features.

The graph of the lung cancer dataset in Fig. 2d, shows similar accuracies between IFP and SVM-RFE throughout the range of 1000 to 2428 features. The t test resulted in the highest accuracy classifier throughout nearly the entire range of 14 through 250 features, except that IFP reached accuracies closer to those of the t test in the range of 43 through 92 features. SVM-RFE and IFP reached very comparable accuracies throughout the range of 180 to 250 features, whereas IFP clearly outperformed SVM-RFE in the range of 33 through 154 features. From 33 downward they showed similar accuracies. Finally, the t test resulted in the lowest accuracy classifier across the top 9 features.

Surprisingly, the t test ranking based on the p value of the features resulted in selecting subsets of features which, in terms of SVM classifier accuracy, tended to outperform both IFP and SVM-RFE across all four datasets. On the other hand, IFP showed accuracy comparable or superior to that of SVM-RFE on the colon, leukemia and lung datasets. IFP contributed to a less accurate classifier on the Moffitt colon dataset, except at the end when less than 33 features remained where IFP accuracy was comparable or superior to that of SVM-RFE at some points. The t test based classifier accuracy was very low in this range of features.

4.4.1 Intersection across entire set of features

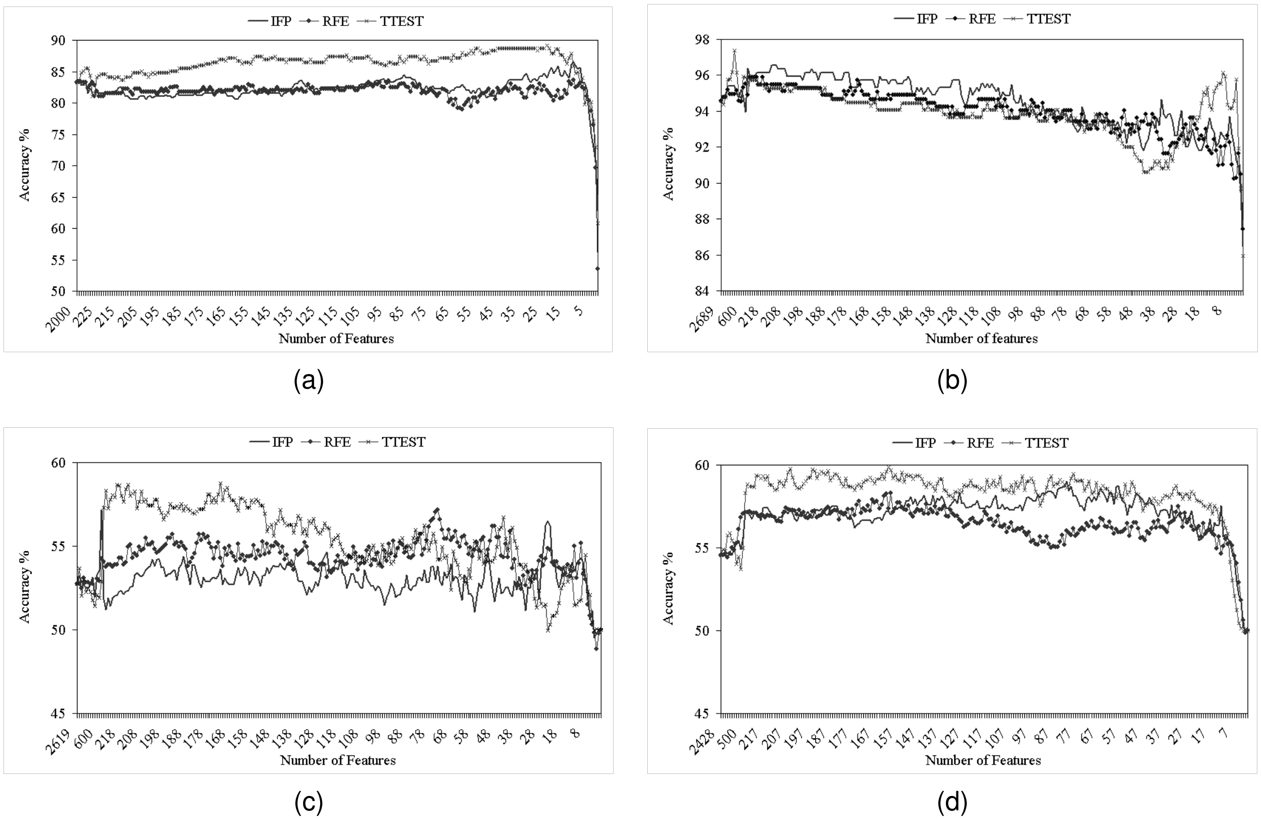
After we observed the results in Fig. 2, a question arose about the degree of similarity of the sets of features at chosen points on the ranked lists resulting from each of the three methods IFP, SVM-RFE and the t test. An analysis of the intersection across the entire set of features on each dataset was done. It consisted of looking at the percentage of intersection of features between any two methods. Given that there were five runs of each method, we ended up with five percentages to average over. Results on the colon cancer dataset are illustrated in Fig. 3, on the leukemia dataset in Fig. 4, on the Moffitt colon dataset in Fig. 5, and on the lung cancer dataset in Fig. 6.

Fig. 3a shows the intersection between IFP and SVM-RFE on the colon cancer dataset. Interestingly, in Fig. 2a accuracies of IFP and SVM-RFE were close to each other at a number of points in the range of 70 through 2000 features, whereas their intersection did not show this same pattern. At 2000 features their intersection started at 100%, and steadily went down as the number of features decreased until around 200 features were left. Thereafter the percentage of intersection was maintained at around 60% all the way until 70 features remained, and their intersection noticeably kept decreasing until 1 feature was left. This fact suggests that different sets of genes can result in similar accuracies.

Fig. 3b, Fig. 3d and Fig. 3f shows intersections of IFP, t test and SVM-RFE respectively with themselves. These three figures show that the t test was more stable than both IFP and SVM-RFE in selecting features.

Fig. 3c shows the intersection between the t test and IFP features which stayed overall very low. It was 100% at 2000 features, and pictorially followed a close-to-linear behavior as the number of features decreased until 250 or less features were left. Thereafter, the intersection was maintained in the range of 15% to 20%. Fig. 3e shows similar behavior on the intersection of the t test and SVM-RFE, except that the percentage of intersection was maintained at around 20% from when 250 features were left.

Fig. 4a shows the intersection between IFP and SVM-RFE on the leukemia dataset. Interestingly, the percentage of features in the intersection between these two methods steadily decreased with the number of features. Also, Fig. 4c and Fig. 4e show the percentage of intersection between the t test and IFP and between the t test and SVM-RFE respectively. The number of features in the intersection of both IFP and SVM-RFE with the t test was low compared to that in the intersection between IFP and SVM-RFE, a result which indicates that the t test selected very different sets of features. However, in terms of accuracies, in Fig. 2b, the three methods did not differ in the same proportion. Particularly, the t test and SVM-RFE showed similar accuracies in the range of 57 through 2689 features. These results highlight the idea that different sets of features (genes) can lead to



(a)

(b)

(c)

(d)

Fig. 2. Comparison of resulting average weighted accuracy of feature ranking given by methods IFP, RFE and t test on (a) colon, (b) leukemia, (c) moffitt colon and (d) lung cancer datasets.

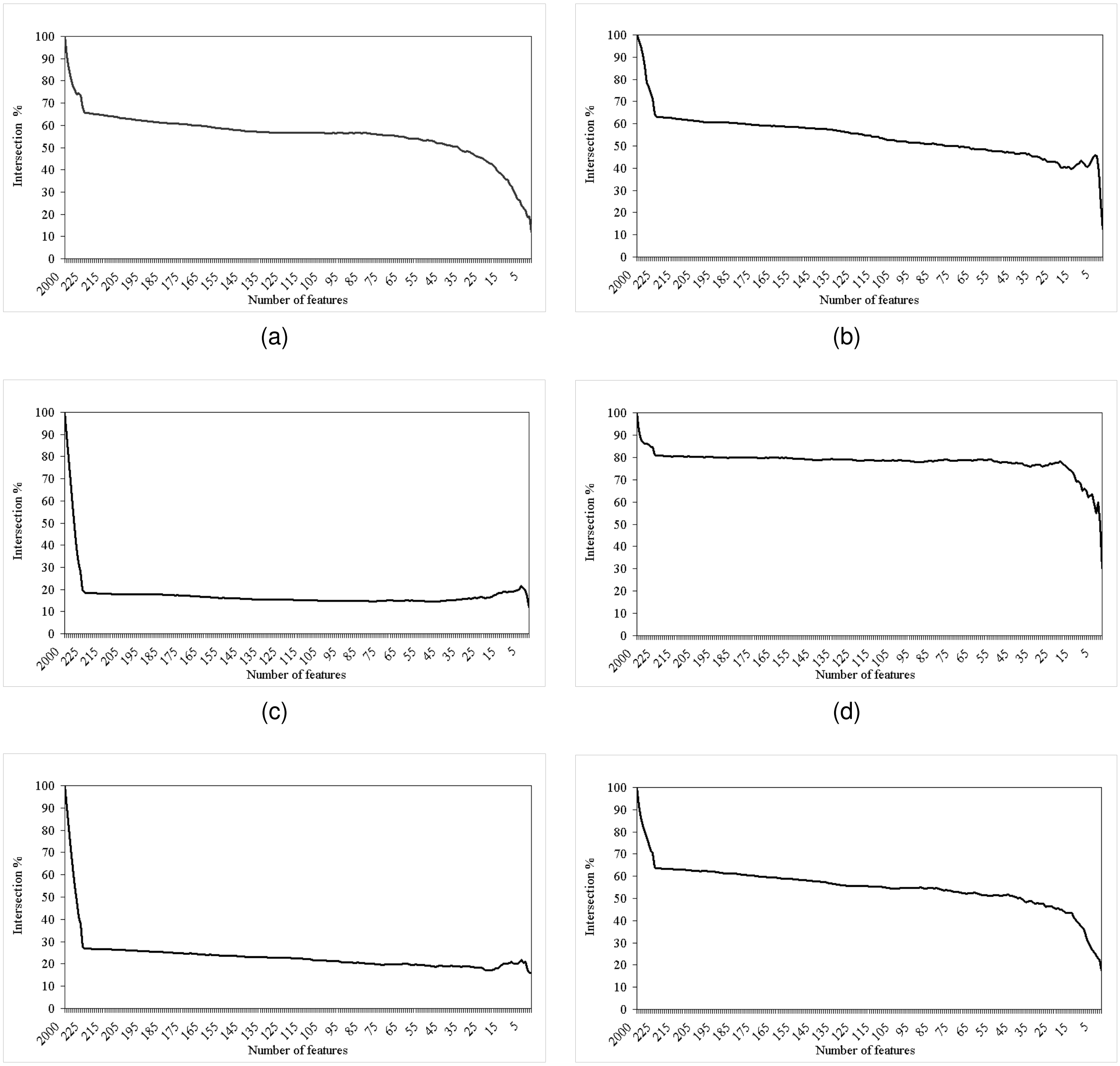
very comparable if not identical accuracies. On the other hand, Fig. 4b, Fig. 4d, and Fig. 4f show, as in the case of the colon cancer dataset in Fig. 3, that the t test was more stable in selecting features across the entire set than IFP and SVM-RFE.

Fig. 5a shows the intersection of features between IFP and RFE on the Moffitt colon cancer dataset. Clearly, it shows the same trend as with the colon and leukemia datasets; that is, the percentage of intersection decreases with the number of features. For this dataset we wanted to focus on the ranges of features where any two methods had similar accuracies. Fig. 2c shows that both the t test and SVM-RFE selection methods resulted in similar accuracies of around 54% in the range of 80 through 114 features; however, their intersection in this range, as shown in Fig. 5e, was around 23%. This means that these methods showed similar base classifier accuracies with different sets of features. The same situation occurred with the top 9 features, where the three methods reached accuracies close to each other, while their percentages of intersections were different. For the t test with IFP as well as for the t test with SVM-RFE the intersection was at less than 20% (2 features) and for IFP with SVM-RFE it was at less than 25%.

Last, the results obtained for lung cancer dataset were not much different from those for the colon cancer, leukemia and Moffitt colon cancer datasets, in the sense that any two methods could reach same or similar accuracies with different sets of features. That was the case for IFP and SVM-RFE whose accuracies shown in Fig. 2d were at 57% in the range of 178 to 225 features, while their intersection shown in Fig. 6a was around 54% in the same range. Another case is the analysis of IFP with the t test, whose accuracies were around 57% in the range of 57 to 71 features, while their intersection shown in Fig. 6c was as low as 6.6% in the same range.

The experiments conducted on four microarray datasets, starting with their entire sets of features, have consistently shown that the feature ranking based on p values resulted in sets of features with very competitive SVM classifier accuracy. Accuracies of SVM classifiers with features chosen by IFP, SVM-RFE and the t test were compared against each other, resulting the t test in either reaching or outperforming the other two in an important range of features.

On the other hand, in [20] it was noted that p values were useful for prioritizing genes for further investigation. Also, in [5] it was advised to pre-reduce the search space upfront via a univariate filter, and secondly apply wrapper or embedded methods. With these arguments in mind, we decided to conduct a series of additional experiments on our four datasets, aiming to see the effect of doing a gene preselection based on their p values, and then examine the performance of IFP and SVM-RFE on



(a) (b)

(c) (d)

(e) (f)

Fig. 3. Intersection across the entire set of features of (a) IFP vs. RFE, (b) IFP vs. IFP, (c) t test vs. IFP, (d) t test vs. t test, (e) t test vs. RFE and (f) RFE vs. RFE on colon cancer dataset.

the resulting dataset comprised of the preselected genes.

4.4.2 Intersection across a subset of features (genes)

The experiments presented in this section were performed on the previously discussed four datasets. They consisted of first preselecting the top 200 genes based on their p values, second, forming a new dataset with this subset, and third, applying IFP and SVM-RFE on it. The gene preselection process was done within a 10-fold cross validation context, each performed five times. The average weighted accuracy over the five runs are reported in Fig. 7. As a reference to results shown in Fig. 2 and for a clearer view of the effect of preselecting the features on each dataset, the t test line of each chart in Fig. 2 was included on its corresponding chart in Fig. 7, as they follow the same aforementioned context. For colon cancer dataset, Fig. 7a shows that with gene preselection both IFP and SVM-RFE reached higher accuracies than those when no preselection was done. Now their accuracies fall on or above 85% for nearly the entire set of 200 genes, in contrast to the below 85% accuracy before. SVM-RFE gained more benefits since it now outperformed the t test in the top 22 features, which did not happen in Fig. 2a.

Fig. 7b shows that for the leukemia dataset the gene preselection process benefited both IFP and SVM-RFE. Their accuracies were higher than when no gene preselection was done; they are now at around 95%. These two methods resulted in more accurate classifiers than the t test in the range of 22 through 200 features on this dataset.

The gene preselection based on p values was beneficial for both IFP and SVM-RFE on the Moffitt colon cancer dataset as Fig. 7c shows. As a positive effect of the preselection, these two methods reached higher accuracies. The IFP average accuracy over the top 200 features without gene preselection was 53.04% while with preselection it went up to 56.27%. The SVM-RFE average accuracy over the top 200 features without gene preselection was 54.43% while with preselection it went up to 56.36%. It is notable in Fig. 7c that both IFP and SVM-RFE reached similar accuracies to those of the t test in the range of 124 to 200 features, while the t test in Fig. 2c, was more accurate in the same range. Also, IFP and SVM-RFE resulted in more accurate classifiers than the t test in the range of 1 to 123 features in Fig. 7c, which did not occurr in Fig. 2c.

Last, Fig. 7d shows once more that both IFP and SVM-RFE gained benefits from gene preselection on the lung cancer dataset. According to the results shown in Fig. 2d, the t test was more accurate than both IFP and SVM-RFE most of the time on the top 200 features, while with gene preselection in Fig. 7d, these two methods improved their accuracies in this range to the point of getting accuracies around or higher than that of the t test.

Overall, results on the four datasets showed that both methods IFP and SVM-RFE improved their accuracies when the dataset underwent a gene preselection process prior to running the learning algorithm under study. In our experiments, the p value criterion was used as the filtering technique. Our results reinforce that it is a useful criterion when preselecting genes for further analysis, as stated by Quackenbush in [20].

In terms of analysis of the intersection across the subset of 200 genes, we proceeded in the same way as described in section 4.4.1. Based on the results obtained in the intersection of genes selected by the three methods, across the entire set of genes of each dataset, the idea now was to focus on points where any two methods reached similar or the same accuracies and look at their intersections of genes at those points. This procedure would help us clarify whether or not it is possible to reach the same accuracy with different sets of features (genes).

Results of percentages of intersection on the colon cancer dataset are shown in Fig. 8, on the leukemia dataset in Fig. 9, on the Moffitt colon cancer dataset in Fig. 10, and on the lung cancer dataset in Fig. 11.

Fig. 8a, Fig. 8c, and Fig. 8e show the average intersection between IFP with SVM-RFE, the t test with IFP, and the t test with SVM-RFE respectively across the 200 genes. In the three cases the amount of intersection decreased with the number of features. The intersection with the t test was lower as fewer features remained. The intersection between the t test and IFP was at less than 50% when 100 features remained and reached its lowest rate of 16.57% at 7 features. The intersection between the t test and SVM-RFE was somewhat similar to that with IFP. These results led us to observe once more that the t test results in selecting very different sets of genes. However, according to accuracies indicated in Fig. 7a, there were points where the three methods reached very similar accuracies, such as at 97 genes where IFP reached 85.57%, SVM-RFE 85.52%, and the t test 85.98%. At this point the intersection with the t test was less than 50% and between IFP and SVM-RFE was 83.96%. Again, it was clear that it is possible to reach similar accuracies with different sets of genes. On the other hand, the intersection of IFP with itself in Fig. 8b resulted in around 59.80% overlap on average, that of SVM-RFE with itself in Fig. 8f resulted in 58.72% overlap on average, and the intersection of the t test with itself in Fig. 8d resulted in 77.28% overlap on average. These numbers indicated that the t test was more stable in selecting genes than IFP and SVM-RFE were.

Fig. 9 shows the average intersections between IFP, SVM-RFE, and the t test on the leukemia dataset. The shape of each of the curves resembles those of the colon dataset in Fig. 8. Similar conclusions were drawn from Fig. 9c and Fig. 9e, as the t test ended up with different sets of selected genes, since the average intersection between the t test with IFP as well as the average intersection between the t test with SVM-RFE, was lower most of the time across the 200 genes than that of IFP with SVM-RFE. Now, the three methods resulted in similar accuracies at 21 genes in Fig. 7b: IFP 93.05%, SVM-RFE 93.37%, and the t test 93.66%, while their average intersections at the same point were, between the t test and IFP 29.43%, betwen the t test and SVM-RFE 30.10%, and between IFP and SVM-RFE 42.57%. Again, as in the colon cancer dataset, results on the leukemia dataset showed that it is possible to reach the same or similar accuracies with different sets of genes. On the other hand, the intersection of IFP with itself in Fig. 9b was 66.84% overlap on average, that of SVM-RFE with itself in Fig. 9f was 67.67% overlap on average, and the intersection of the t test with itself in Fig. 9d resulted in 85.82% overlap on average. Again, these numbers indicated that the t test was more stable in selecting genes than IFP and SVM-RFE.

Fig. 10 shows the average intersections between IFP, SVM-RFE and the t test on the Moffitt colon cancer dataset. The shape of each of the curves resembles those of the colon cancer and leukemia datasets in Fig. 8 and Fig. 9 respectively, except that the t test in Fig. 10d did not maintain the same percentage of intersection across the 200 genes as it did on the colon cancer and leukemia datasets. As for the intersections of IFP and SVM-RFE with the t test in Fig. 10c and Fig. 10e respectively, they showed again that the t test selected very different sets of genes. Also, the three methods reached similar accuracies at 141 features in Fig. 7c: IFP

56.70%, SVM-RFE 56.35%, and t test 56.27%. However their intersections, even though they were high at this point, they were not close in the same proportion among them. Average intersections between the t test with IFP reached 70.74%, the t test with SVM-RFE 71.22%, and IFP with SVM-RFE 91.72%. Once more, the observation was that with different sets of genes, the resulting accuracies can be alike.

Finally, Fig. 11 shows the average intersections between IFP, SVM-RFE, and the t test on the lung cancer dataset. The shape of each of the curves led us to note that similar conclusions can be drawn regarding reaching alike accuracies with different sets of genes. Even though for this dataset, the accuracies of the three methods shown in Fig. 7d are closer to each other across the 200 genes, still their average intersections showed similar behavior to those of the colon cancer, leukemia and Moffitt colon cancer datasets.

4.5 AUC analysis

The accuracy results between methods for all datasets were additionally analyzed using the area under the curve (AUC) [12]. AUC was calculated for all of the accuracy curves in Fig. 2 and Fig. 7 using the trapezoidal method in the student version of the DADiSP software [21]. Fig. 2 shows accuracy results on our 4 datasets using the entire set of features. Fig. 7 shows accuracy results using the top 200 preselected features with base on their p values.

4.5.1 Across the entire set of features

Results are shown in Table 2. Table 3a shows the AUC of each method across the 4 datasets. The t test had the greatest AUC on the colon cancer and the leukemia datasets. While SVM-RFE was highest on the Moffitt colon and the lung cancer datasets. The comparison of any 2 methods indicated in Table 3b that IFP had greater AUC than that of SVM-RFE only on the leukemia dataset. Table 3c shows that the t test had greater AUC than that of IFP on the colon, Moffitt colon, and the lung cancer datasets. Table 3d shows that the t test reached greater AUC than that of SVM-RFE on the colon cancer and the leukemia datasets. Interestingly, the differences in AUC on the datasets where the t test outperformed IFP or SVM-RFE are mostly larger than those where these methods outperformed the t test. Also, differences in AUC between IFP and SVM-RFE are mostly smaller in magnitude than those when these 2 methods compared against the t test.

4.5.2 Across the top 200 subset of features

Results are shown in Table 3. Table 4a shows the AUC of each method across the 4 datasets using only the top 200 features. The t test had the greatest AUC on the colon and the lung cancer datasets. While IFP was highest on the leukemia dataset. SMV-RFE had the greatest AUC on the Moffitt colon cancer dataset. The comparison of any 2 methods indicated in Table 4b that IFP now had greater AUC than SVM-RFE on the colon cancer, the leukemia and the lung cancer datasets. Table 4c shows that IFP had a bigger AUC than that of the t test on the leukemia and the Moffitt colon cancer datasets. Table 4d shows that SVM-RFE had greater AUC than the t test on the leukemia and the Moffitt colon cancer datasets. Interestingly, the differences in AUC on the datasets where either IFP or SVM-RFE outperformed the t test were comparable to those where the t test outperformed the two former. This is true except on the lung cancer dataset where the t test based feature selector resulted in better classifier than both IFP and SVM-RFE with a very minimum difference in AUC. On the other hand, differences in AUC between IFP and SVM-RFE were very small as compared to those between these two methods and the t test. Our perception at this point was that the preselection of genes helped both IFP and SVM-RFE improve their performance. Considering AUC values only, we observed that accuracies of IFP and SVM-RFE got closer to and in some cases exceeded those of the t test as effect of gene preselection.

4.6 Statistical significance analysis

Results described in previous sections were further analyzed for a statistical significance of the differences in accuracies between methods. We used the Friedman-Holm test which has been discussed in [22], [23], [24], [25]. The Friedman test is a non-parametric test that allows the comparison of two or more classifiers. It ranks the methods being compared ranging from 1-3 in this work, 1 and 3 being the highest and the lowest ranks respectively. Ties of 1 are each given 1.5. The null hypothesis states there are no differences between the methods. When the null hypothesis is rejected, a post-hoc test follows to determine the method with better results. In this work we used the Holm procedure as a post-hoc test. It consists of sequentially testing hypotheses starting with the most significant p value. When a hypothesis is rejected the Holm procedure moves on to the next p value and continues until no null hypothesis can be rejected.

We applied the Friedman-Holm test to the top 50 accuracies resulting from two scenarios: a) starting with the entire set of features in the dataset (Fig. 2), and b) starting with a preselected set of n features, n being the number of features with p values <= 0.01. The latter criterion implied that each dataset had a different-in-size initial set of features. The colon cancer dataset started with 201 features, the leukemia dataset with 437 features, the Moffitt colon dataset with 50 features, and the lung cancer dataset with 375 features. Our sample size was 10, that is, for each dataset we conducted the 10-fold cross validation experiment 10 times with different seeds each.

Throughout the following description of the statistical analysis, the terms different and better are used to

TABLE 2
AUC analysis of accuracy curves across all 4 datasets using the entire set of features. (a) AUC (a bold entry represents the highest AUC). Comparison between methods in terms of AUC difference (noted in the column of the method with highest AUC): (b) IFP vs. RFE, (c) IFP vs. t test, and (d) SVM-RFE vs. t test

Dataset/Method	IFP	SVM-RFE	t test
Colon	164654.88	165656.48	**168526.20**
Leukemia	255409.53	254867.40	**256530.08**
Moffitt Colon	138889.91	**139046.06**	138300.87
Lung	134020.46	**134691.81**	134664.98

(a)

Dataset/Method	IFP	SVM-RFE
Colon		1001.6
Leukemia	542.13	
Moffitt colon		156.15
Lung		671.35

(b)

IFP	t test
	3871.32
	1120.08
589.04	
	644.52

(c)

SVM-RFE	t test
	2869.72
	1662.68
745.19	
26.83	

(d)

TABLE 3
AUC analysis of accuracy curves across all 4 datasets using the top 200 features. (a) AUC (a bold entry represents the highest AUC). Comparison between methods in terms of AUC difference (noted in the column of the method with highest AUC): (b) IFP vs. SVM-RFE, (c) IFP vs. t test, and (d) SVM-RFE vs. t test

Dataset/Method	IFP	SVM-RFE	t test
Colon	16962.56	16958.48	**17252.42**
Leukemia	**18940.77**	18877.87	18636.43
Moffitt Colon	11200.31	**11219.15**	10961.38
Lung	11598.31	11593.80	**11600.65**

(a)

Dataset/Method	IFP	SVM-RFE
Colon	4.08	
Leukemia	62.9	
Moffitt colon		18.84
Lung	4.51	

(b)

IFP	t test
	289.86
304.34	
238.93	
	2.34

(c)

SVM-RFE	t test
	293.94
241.44	
257.77	
	6.85

(d)

mean statistically significant different and statistically significant better, respectively. We considered statistical significant results at 95% confidence level (p values <= 0.05).

4.6.1 Starting with the entire set of features

Table 4 shows the results for the colon cancer dataset. Each table entry shows the method that was found better (winner) between the methods at the number of features indicated by the column title. A blank table entry means that no method was found significantly different.

When compared to IFP the t test was better for 35 feature sets, no difference was found for 14 feature sets and IFP was found better for only 1 feature set. When compared to SVM-RFE the t test was better for 38 feature sets, and no difference was found for 12 feature sets. The comparison between IFP and SVM-RFE indicated that both methods were nearly the same except at 1 feature where IFP was a better classifier. Table 5 shows the results for the leukemia dataset. The subtable corresponding to the range of 25-50 features was omitted since no method was found different throughout this range. When compared to IFP the t test was better for 11 feature sets, and no method was found different for 39 feature sets. When compared to SVM-RFE the t test was better for 9 feature

TABLE 4
Statistical analysis of results on the colon cancer dataset between the methods IFP (I), RFE (R) and the t test (T) across features (a) 50 to 31, (b) 30 to 16, and (c) 15 to 1 with no previous preselection of features.

	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31
t test vs IFP	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
t test vs RFE	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
IFP vs RFE																				

(a)

	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16
t test vs IFP	T	T	T	T	T	T	T	T	T	T	T	T	T	T	
t test vs RFE	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
IFP vs RFE															

(b)

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
t test vs IFP												T			I
t test vs RFE			T	T										T	
IFP vs RFE															I

(c)

sets, and no method was found different for 41 feature sets. No difference was found between IFP and SVM-RFE.

The results for the Moffitt colon cancer dataset showed that no difference was found between IFP and the t test for 49 feature sets. IFP was better for 1 feature set (12 features). When compared to SVM-RFE the t test was better for 1 feature set (39 features). No difference was found for 48 feature sets. SVM-RFE was found better for 1 feature set (12 features). No difference was found between IFP and SVM-RFE.

The results for the lung cancer dataset showed that no difference was found between IFP and the t test for 44 feature sets. IFP was better for 3 feature sets (7, 6, and 5 features). The t test was better for 3 feature sets (42, 34, and 23 features). When compared to SVM-RFE no difference was found for 44 feature sets. SVM-RFE was better for 3 feature sets (7, 6, and 5 features). The t test was better for 3 feature sets (42, 37, 34 features). No difference was found between IFP and SVM-RFE.

Interestingly, the statistical analysis showed that the more complicated feature selection algorithms IFP and SVM-RFE did not generally result in better classifiers using a support vector machine as the base classifier on microarray data.

4.6.2 Starting with a preselected set of n features

This section describes the statistical analysis performed on the results derived from experiments under the aforementioned scenario 2. That is, prior to the application of IFP and SVM-RFE each dataset underwent a preselection of the top n features, where n was the number of features with p values <= .01.

Table 6 shows the results on the colon cancer dataset. When compared to IFP the t test was better on 20 feature sets (15 less than without preselection). No difference was found in 29 feature sets (15 more than without preselection). IFP was better in 1 feature set. When compared to SVM-RFE the t test was found better on 15 feature sets (23 less than without preselection). No difference was found in 35 feature sets (23 more than without preselection). IFP was better than SVM-RFE in 1 feature set (6 features).

Results on this dataset indicated that there was a change in the performance of IFP and SVM-RFE by doing upfront a preselection of genes, as follows. The number of feature sets where the t test was found better than IFP and SVM-RFE decreased after preselecting. Also, the number of feature sets where no difference was found between the methods being compared increased after preselecting. So, it was not the case that IFP and SVM-RFE were found better than the t test; however, these methods got enough performance improvement after preselecting, that they showed no difference with the t test in ga reater number of feature sets.

Table 7 shows the results on the leukemia dataset. No difference was found between IFP and the t test on 34 feature sets (5 less than without preselection). Of these 5 feature sets, IFP was better on 3 (3 more than without preselection). The t test was now better on 13 feature sets (2 more than without preselection). When compared to SVM-RFE the t test was found better for 11 feature sets and no difference was found for 39 feature sets. The comparison between IFP and SVM-RFE showed no difference on 46 feature sets (4 less than without preselection). SVM-RFE was found better than IFP on 4 feature sets (4 more than without preselection).

The results for the Moffitt colon cancer dataset indicated that after preselection no difference was found between IFP and the t test throughout all 50 features. Without preselection IFP had resulted better than the t test on 1

TABLE 5
Statistical analysis of results on the leukemia dataset between the methods IFP (I), RFE (R) and the t test (T) across features (a) 24 to 11, and (b) 10 to 1 with no previous preselection of features.

	24	23	22	21	20	19	18	17	16	15	14	13	12	11
t test vs IFP	T	T		T	T	T	T	T					T	T
t test vs RFE				T				T					T	T
IFP vs RFE														

(a)

	10	9	8	7	6	5	4	3	2	1
t test vs IFP						T	T			
t test vs RFE	T	T	T			T	T			
IFP vs RFE										

(b)

TABLE 6
Statistical analysis of results on the colon cancer dataset between the methods IFP (I), RFE (R) and the t test (T) across features (a) 50 to 31, (b) 30 to 16, and (c) 15 to 1 with previous preselection of *n* features (those with p values $<= 0.01$).

	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31
t test vs IFP	T	T	T	T	T	T	T	T	T	T		T	T	T	T			T		
t test vs RFE	T	T	T	T	T	T	T	T	T	T		T	T	T	T			T		
IFP vs RFE																				

(a)

	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16
t test vs IFP	T	T	T	T										T	
t test vs RFE															
IFP vs RFE														R	

(b)

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
t test vs IFP							I								
t test vs RFE															
IFP vs RFE										I					

(c)

feature set. On the other hand, no difference was found between the t test and SVM-RFE on 49 feature sets (1 more than without preselection). That 1 feature set benefited SVM-RFE which improved its performance to reach that of the t test. SVM-RFE was still found better than the t test on 1 feature set. The comparison between IFP and SVM-RFE showed no difference on 47 feature sets (3 less than without preselection). SVM-RFE was found better than IFP on 3 feature sets (3 more than without preselection).

Table 8 shows the results on the lung cancer dataset. The subtable corresponding to the range of 34-50 features was omitted since no method was found different throughout this range. When compared to the t test IFP was better on 5 feature sets (2 more than without preselection). The t test was not found better at all (it used to be better at 3 feature sets without preselection). No difference was found in 45 feature sets (1 more than without preselection). When compared to SVM-RFE the t test was better on 4 feature sets (1 more than without preselection). SVM-RFE was better only on 1 feature set (2 less than without preselection). No difference was found on 45 feature sets (1 more than without preselection). IFP was better than SVM-RFE on 14 feature sets (14 more than without preselection). Interestingly, our results showed that the preselection of genes prior to the application of IFP or SVM-RFE mostly makes a positive impact on the performance of these feature selection methods.

4.6.3 Preselection vs. no preselection of IFP and SVM-RFE.

Previous sections described the statistical significance analysis performed on the top 50 accuracies obtained for each of the methods on each dataset. First, we analyzed the scenario without doing preselection prior to the ap-

TABLE 7
Statistical analysis of results on the leukemia dataset between the methods IFP (I), RFE (R) and the t test (T) across features (a) 50 to 31, (b) 30 to 16, and (c) 15 to 1 with previous preselection of *n* features (those with p values $<= 0.01$).

	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31
t test vs IFP						I	I	I												
t test vs RFE																				
FP vs RFE																				

(a)

	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16
t test vs IFP							T	T			T	T		T	T
t test vs RFE					T		T					T		T	
FP vs RFE											R				

(b)

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
t test vs IFP			T	T	T						T	T	T		T
t test vs RFE			T	T	T	T	T				T	T			
FP vs RFE													R	R	R

(c)

TABLE 8
Statistical analysis of results on the lung dataset between the methods IFP (I), RFE (R) and the t test (T) across features (a) 33 to 18, (b) 17 to 1 with previous preselection of *n* features (those with p values $<= 0.01$).

	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18
t test vs IFP																
t test vs RFE	T							T								
IFP vs RFE								I				I			I	I

(a)

	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
t test vs IFP										I	I	I	I	I			
t test vs RFE				T	T								R				
IFP vs RFE	I	I	I	I	I	I		I	I	I	I						

(b)

plication of IFP and SVM-RFE. Second, we analyzed the scenario where we did a preselection of features/genes prior to the application of these two methods. In this section we statistically describe how each method compared to itself on its two versions: with no preselection and with preselection of genes. The analysis was performed on each dataset. The analysis showed at each feature set which version was statistically significant better or if no statistical significant difference was found between the two versions.

Table 9 shows the results on the colon cancer dataset. On IFP, doing preselection was found better on 26 feature sets. No difference between the two versions was found on 24 feature sets. On SVM-RFE, doing preselection was found better on 23 feature sets. No difference was found on 27 feature sets.

Results on the leukemia dataset showed that for IFP, doing preselection was found better on 4 feature sets (45, 46, 47, and 48 features). No difference was found between the two versions on 46 feature sets. On SVM-RFE, doing preselection was found better in 1 feature set (21 features). No difference was found in 49 feature sets. Results on the Moffitt colon cancer dataset showed that for IFP, not doing preselection was found better on 3 feature sets (9, 13, 14 features). No difference was found between the two versions on 47 feature sets. On SVM-RFE, no difference at all was found across the 50 feature sets.

Table 10 shows the results on the lung cancer dataset. The subtable corresponding to the range 50-33 was omitted since no difference difference between the two scenarios was found throughout this range. On IFP, doing preselection was found better on 13 feature sets. No difference between the two versions was found on 37 feature sets. On SVM-RFE, not doing preselection was found better on 1 feature set. No difference was found on 49 feature sets.

Our results showed that the preselection of features

TABLE 9
Statistical comparison between doing preselection (P) and not doing preselection (NP) with the methods IFP and SVM-RFE across features (a) 50 to 31, (b) 30 to 16, and (c) 15 to 1 on the colon cancer dataset.

	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31
IFP	P	P	P	P	P	P	P	P							P	P		P		P
RFE																P	P	P	P	P

(a)

	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16
IFP	P	P	P	P	P	P	P	P	P	P					
RFE	P	P	P	P	P	P	P		P	P	P	P	P	P	P

(b)

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
IFP						P	P	P					P		
RFE	P	P	P												P

(c)

made a statistically significant difference on some feature sets on the colon cancer, leukemia and lung cancer datasets.

5 Conclusions

The IFP algorithm was introduced. It includes a self-tuning mechanism, via binary search, to determine the amount of noise needed to perturb any number of features (genes). We compared the performance of three feature selection methods: IFP, SVM-RFE and the t test, in terms of average SVM classifier accuracy. Four microarray datasets were preprocessed and used in our experiments: colon cancer, leukemia, Moffitt colon cancer and lung cancer datasets. Overall, IFP resulted in a classifier comparable or superior in accuracy to SVM-RFE on the colon, leukemia and lung datasets. IFP resulted in a less accurate classifier on the Moffitt colon dataset.

Surprisingly, the t test feature ranking, which is based on the p values of the genes, turned out to be the best gene selector explored. It found better sets of genes than the more complicated IFP and SVM-RFE did, in the sense that the genes selected by the t test led to SVM accuracies higher than those of IFP and SVM-RFE, in many gene subsets across all 4 datasets. This suggests that perhaps the more complex algorithms for feature selection increase the risk of overfitting for such small sample problems.

Based on the good performance of the t test as a gene selector, we investigated the effect of doing a preselection of genes/features across our 4 datasets before the application of IFP and SVM-RFE. We used the p values of each gene as our filter criterion and analyzed the accuracy results statistically using the Friedman/Holm test and using the AUC criterion. Both scenarios of experiments were analyzed, with and without gene preselection. Our results confirmed the superiority of the t test on experiments without gene preselection as well as the performance improvement of IFP and SVM-RFE on experiments where gene preselection with the t test was incorporated.

We also looked at the similarity of the sets of genes selected by each of the methods, with particular emphasis on points where any two methods reached alike accuracies. Our findings indicate that similar accuracies can be reached with different sets of genes.

While the t test can be an accurate technique for feature selection, it is limited to two-class problems. However, the use of ANOVA could provide a similar method in the case of 3 or more classes.

Acknowledgments

The authors would like to thank...

References

[1] L. Chen, D. Goldgof, L. Hall, and S. Eschrich, "Noise-Based Feature Perturbation as a Selection Method for Microarray Data," in *Bioinformatics Research and Applications: Third International Symposium, Isbra 2007, Atlanta, Ga, USA, May 7-10, 2007, Proceedings*. Springer-Verlag New York Inc, 2007, p. 237.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[3] S. Eschrich, I. Yang, G. Bloom, K. Kwong, D. Boulware, A. Cantor, D. Coppola, M. Kruhoffer, L. Aaltonen, T. Orntoft *et al.*, "Molecular staging for survival prediction of colorectal cancer patients," *Journal of Clinical Oncology*, vol. 23, no. 15, p. 3526, 2005.

[4] J. Jäger, R. Sengupta, and W. Ruzzo, "Improved gene selection for classification of microarrays," *Biocomputing 2003*, p. 53, 2003.

[5] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, p. 2507, 2007.

[6] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[7] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006.

TABLE 10
Statistical comparison between doing preselection (P) and not doing preselection (NP) with the methods IFP and SVM-RFE across features (a) 32 to 13, (b) 12 to 1 on the lung cancer dataset.

	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13
IFP	P				P			P			P		P	P		P	P	P		
RFE																				

(a)

	12	11	10	9	8	7	6	5	4	3	2	1
IFP			P			P		P	P			
RFE						NP						

(b)

[8] L. Wang, J. Zhu, and H. Zou, "Hybrid huberized support vector machines for microarray classification," in *Proceedings of the 24th international conference on Machine learning*. ACM New York, NY, USA, 2007, pp. 983–990.

[9] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, p. 531, 1999.

[10] L. Wang, J. Zhu, and H. Zou, "Hybrid huberized support vector machines for microarray classification and gene selection," *Bioinformatics*, vol. 24, no. 3, p. 412, 2008.

[11] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[12] J. Canul-Reich, L. Hall, D. Goldgof, and S. Eschrich, "Feature selection for microarray data by auc analysis," in *2008 IEEE International Conference on Systems, Man, and Cybernetics, Singapore, October 12-15, Proceedings*, 2008, pp. 768–773.

[13] B. Scholkopf, C. Burges, and V. Vapnik, "Extracting support data for a given task," in *Knowledge Discovery and Data Mining*, 1995, pp. 252–257. [Online]. Available: http://www.aaai.org/Papers/KDD/1995/KDD95-030.pdf

[14] M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent systems*, vol. 13, no. 4, pp. 18–28, 1998.

[15] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, March 2000.

[16] D. Pyle, *Data preparation for data mining*. Morgan Kaufmann Pub, 1999.

[17] K. Shedden, J. Taylor, S. Enkemann, M. Tsao, T. Yeatman, W. Gerald, S. Eschrich, I. Jurisica, T. Giordano, D. Misek *et al.*, "Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," *Nature Medicine*, vol. 14, no. 8, pp. 822–827, 2008.

[18] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[19] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Pub, 2005.

[20] J. Quackenbush, "Microarray analysis and tumor classification," *The New England journal of medicine*, vol. 354, no. 23, p. 2463, 2006.

[21] *DADiSP/SE 2002: The ultimate engineering spreadsheet*, version 6.0. [Online]. Available: http://www.dadisp.com

[22] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[23] Lowry, R., *Concepts and applications of inferential statistics*, Vassar College, Poughkeepsie, NY, 2009. [Online]. Available: http://faculty.vassar.edu/lowry/webtext.html

[24] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[25] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

PLACE
PHOTO
HERE

Juana Canul-Reich Biography text here.

PLACE
PHOTO
HERE

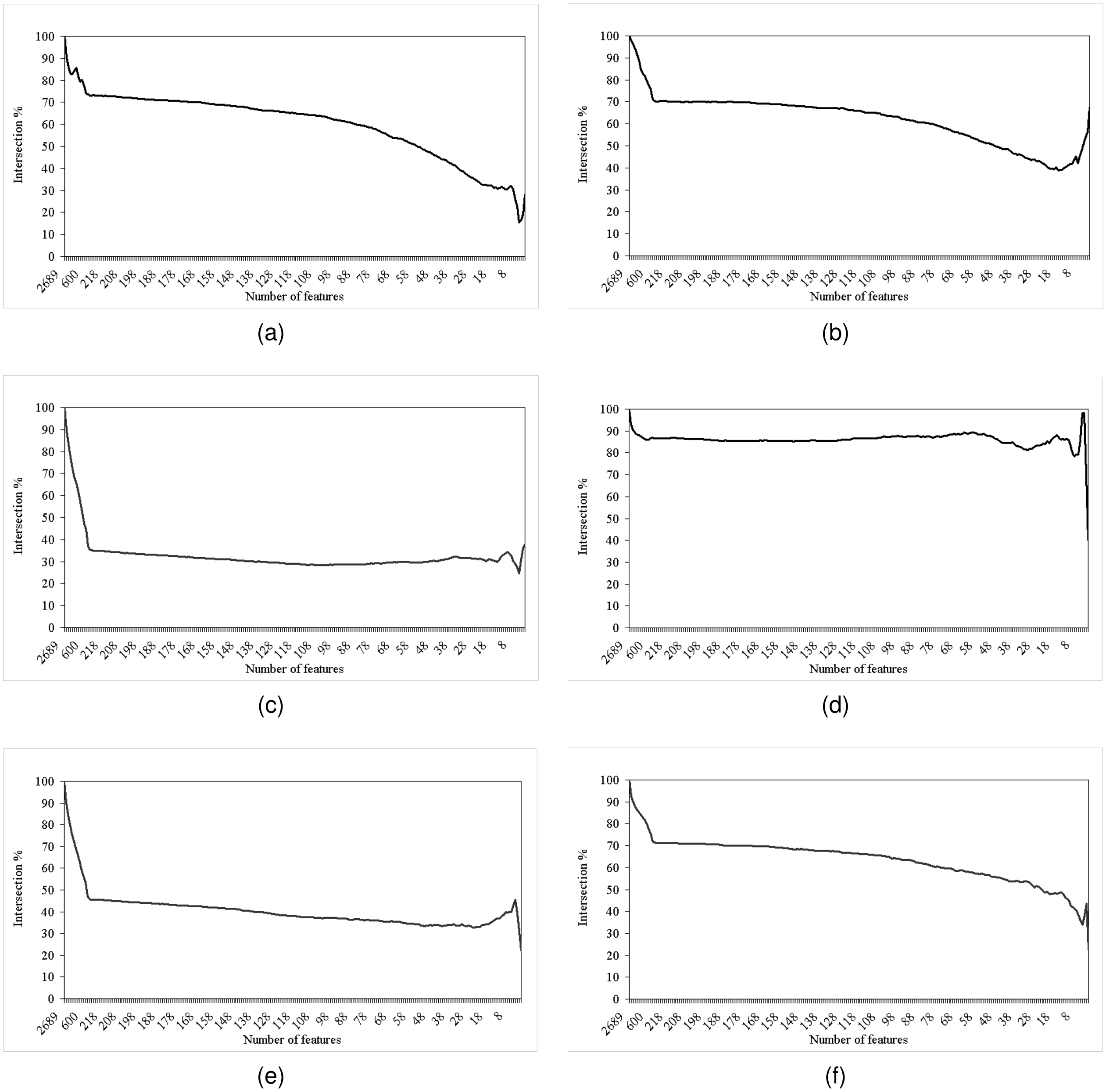
Lawrence O. Hall Biography text here.

PLACE
PHOTO
HERE

Dmitry Goldgof Biography text here.

PLACE
PHOTO
HERE

Steven Eschrich Biography text here.



(a)

(b)

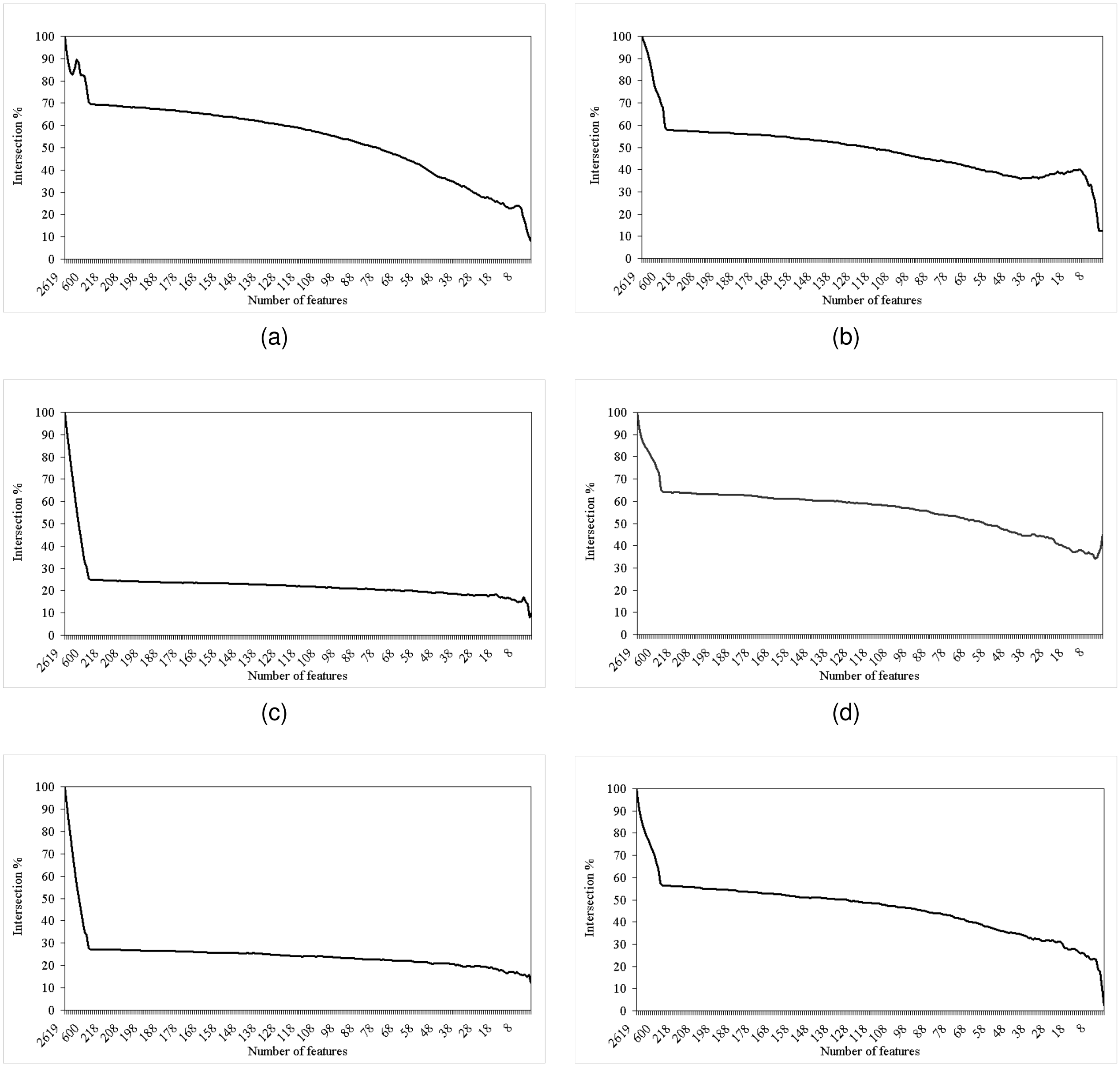
(c)

(d)

(e)

(f)

Fig. 4. Intersection across the entire set of features of (a) IFP vs. RFE, (b) IFP vs. IFP, (c) t test vs. IFP, (d) t test vs. t test, (e) t test vs. RFE and (f) RFE vs. RFE on leukemia dataset.



(a)

(b)

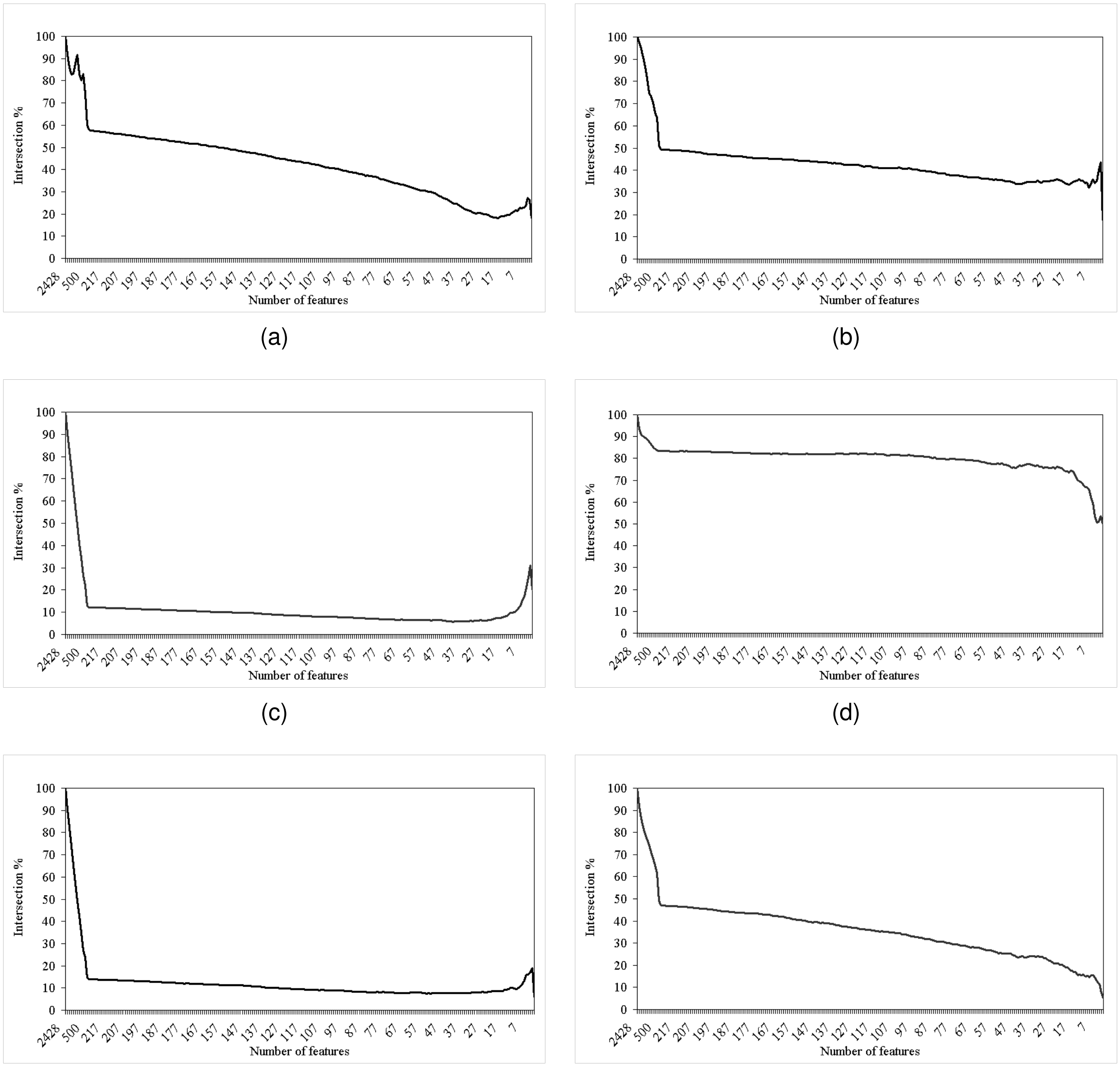
(c)

(d)

(e)

(f)

Fig. 5. Intersection across the entire set of features of (a) IFP vs. RFE, (b) IFP vs. IFP, (c) t test vs. IFP, (d) t test vs. t test, (e) t test vs. RFE and (f) RFE vs. RFE on Moffitt colon cancer dataset.



(a)

(b)

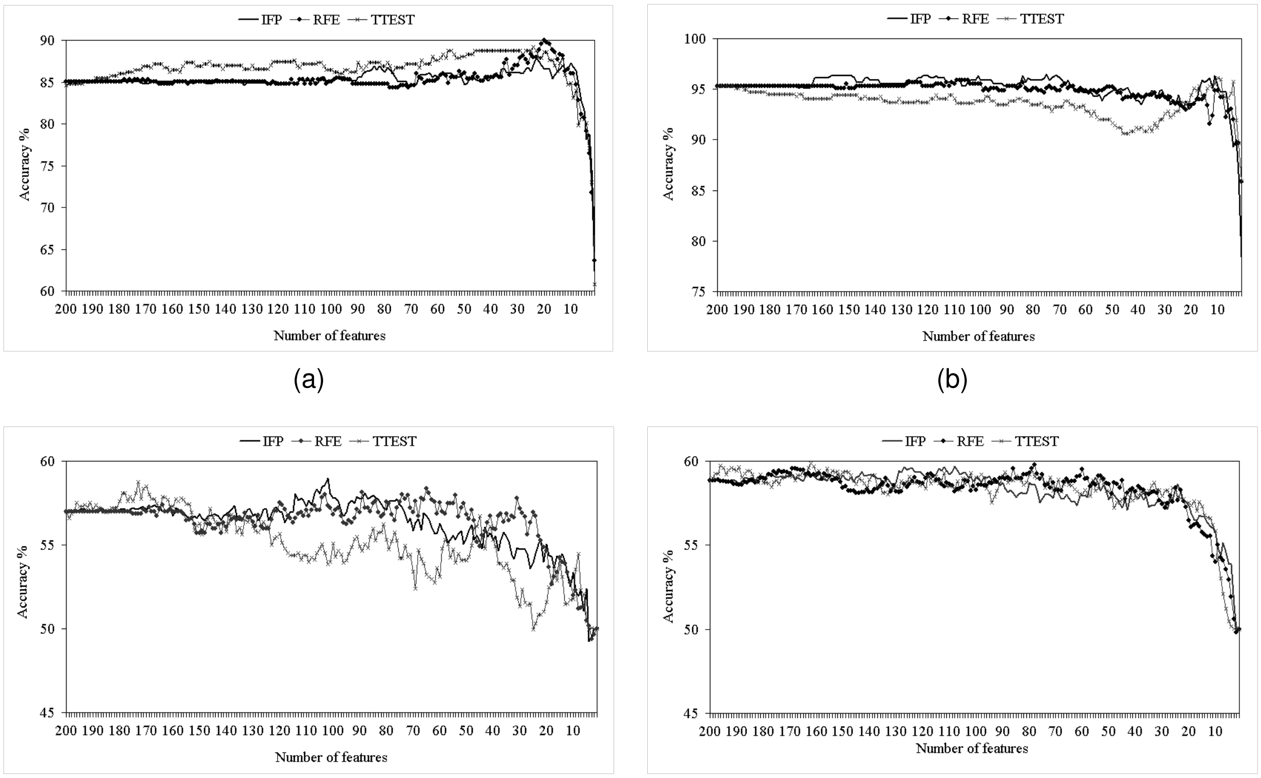
(c)

(d)

(e)

(f)

Fig. 6. Intersection across the entire set of features of (a) IFP vs. RFE, (b) IFP vs. IFP, (c) t test vs. IFP, (d) t test vs. t test, (e) t test vs. RFE and (f) RFE vs. RFE on lung cancer dataset.



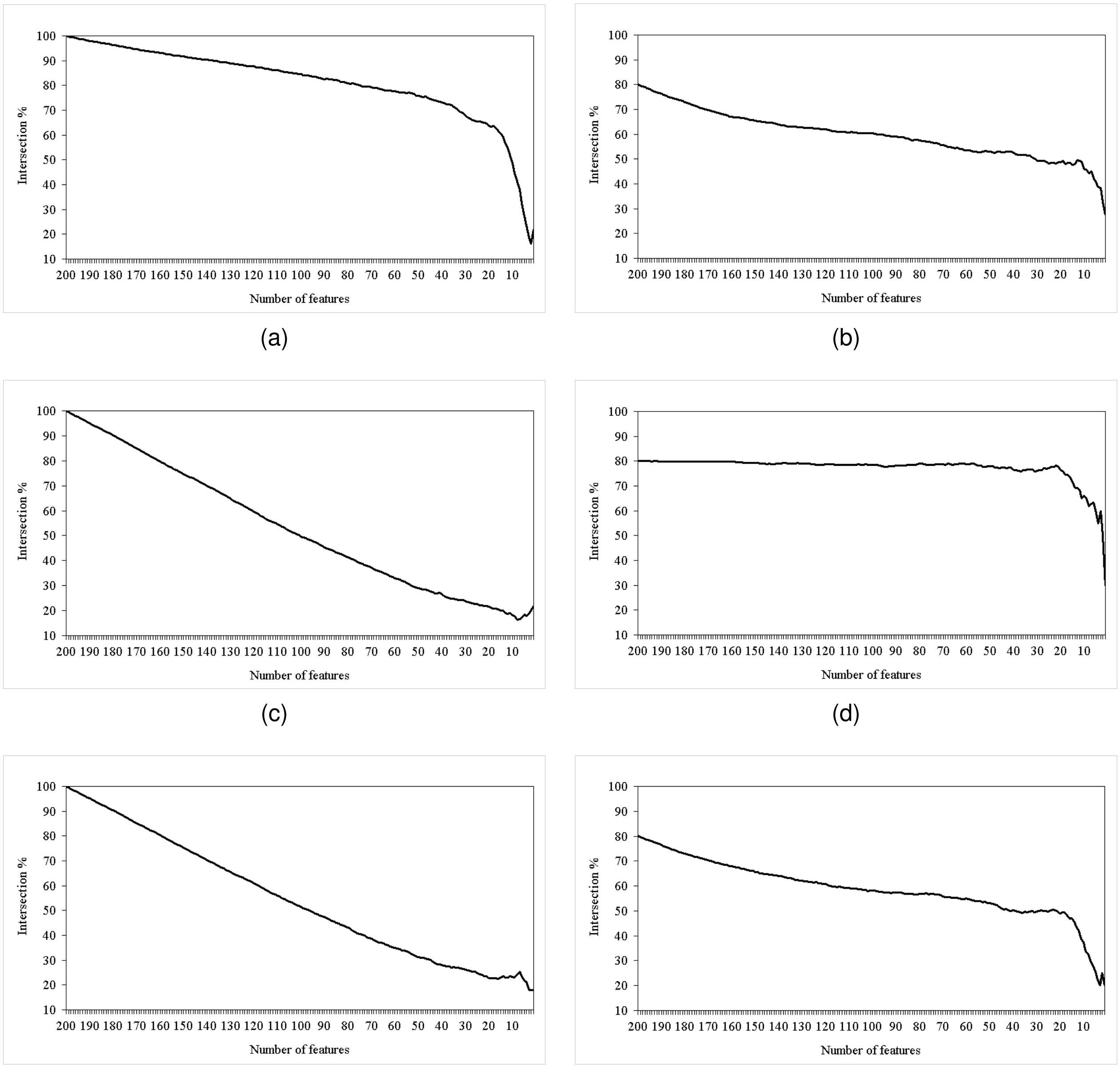
(a)

(b)

(c)

(d)

Fig. 7. Comparison of resulting average weighted accuracy of feature ranking given by methods IFP, RFE and t test on (a) colon, (b) leukemia, (c) moffitt colon and (d) lung cancer datasets across top 200 features as filtered by p values.



(a)

(b)

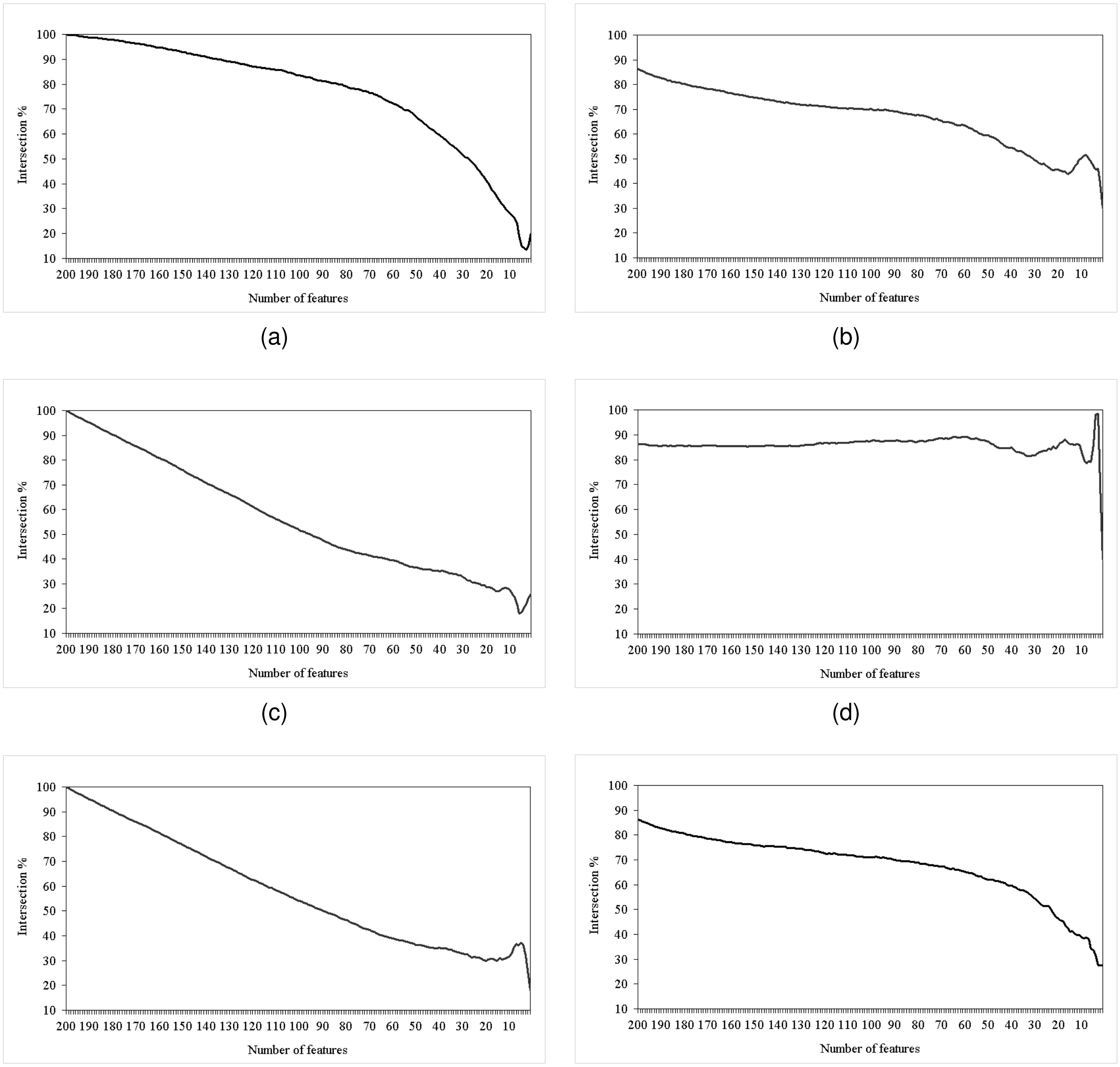
(c)

(d)

(e)

(f)

Fig. 8. Intersection across the top 200 subset of features of (a) IFP vs. RFE, (b) IFP vs. IFP, (c) t test vs. IFP, (d) t test vs. t test, (e) t test vs. RFE and (f) RFE vs. RFE on colon cancer dataset.



(a)

(b)

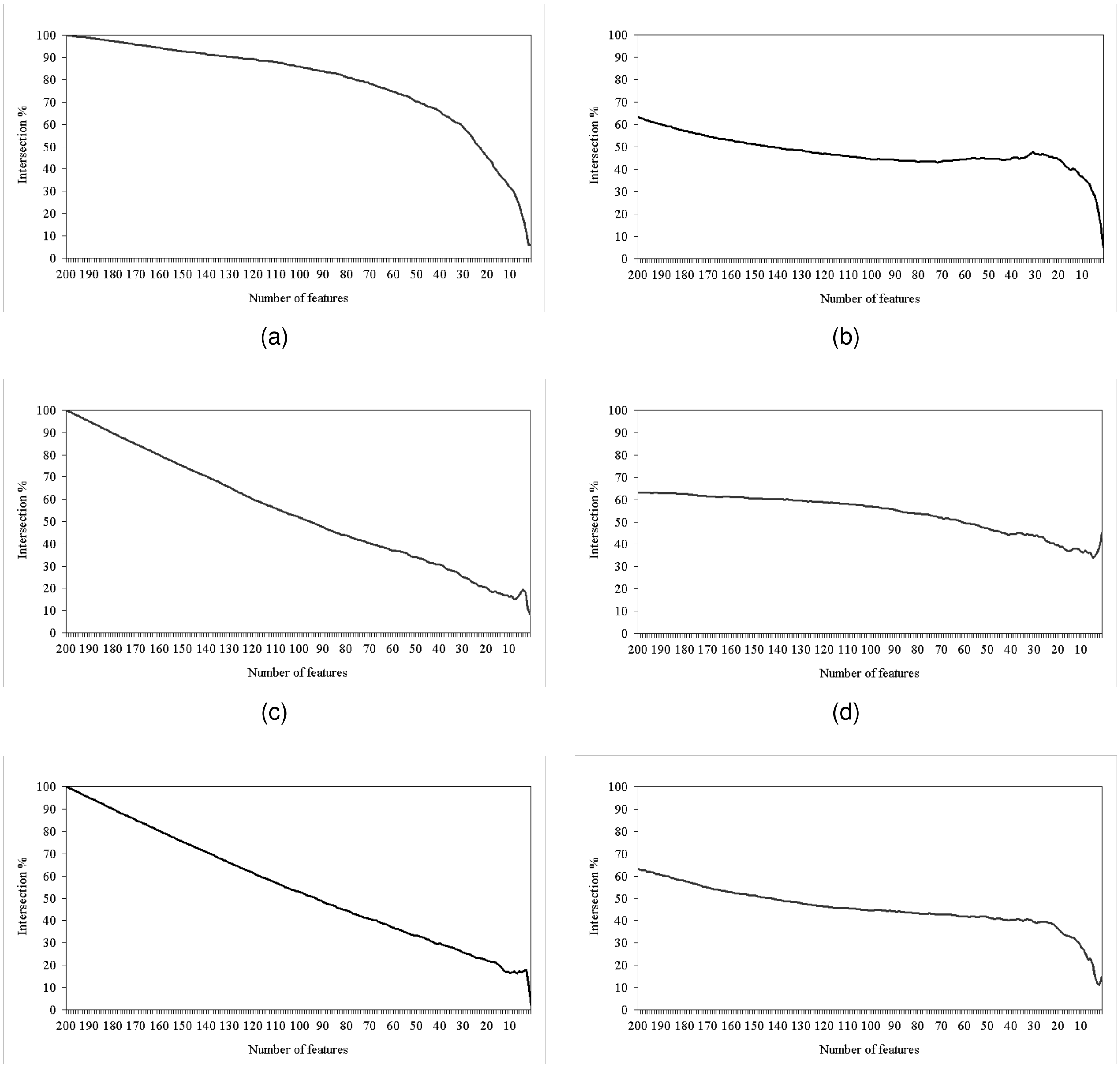
(c)

(d)

(e)

(f)

Fig. 9. Intersection across the top 200 subset of features of (a) IFP vs. RFE, (b) IFP vs. IFP, (c) t test vs. IFP, (d) t test vs. t test, (e) t test vs. RFE and (f) RFE vs. RFE on leukemia cancer dataset.



(a)

(b)

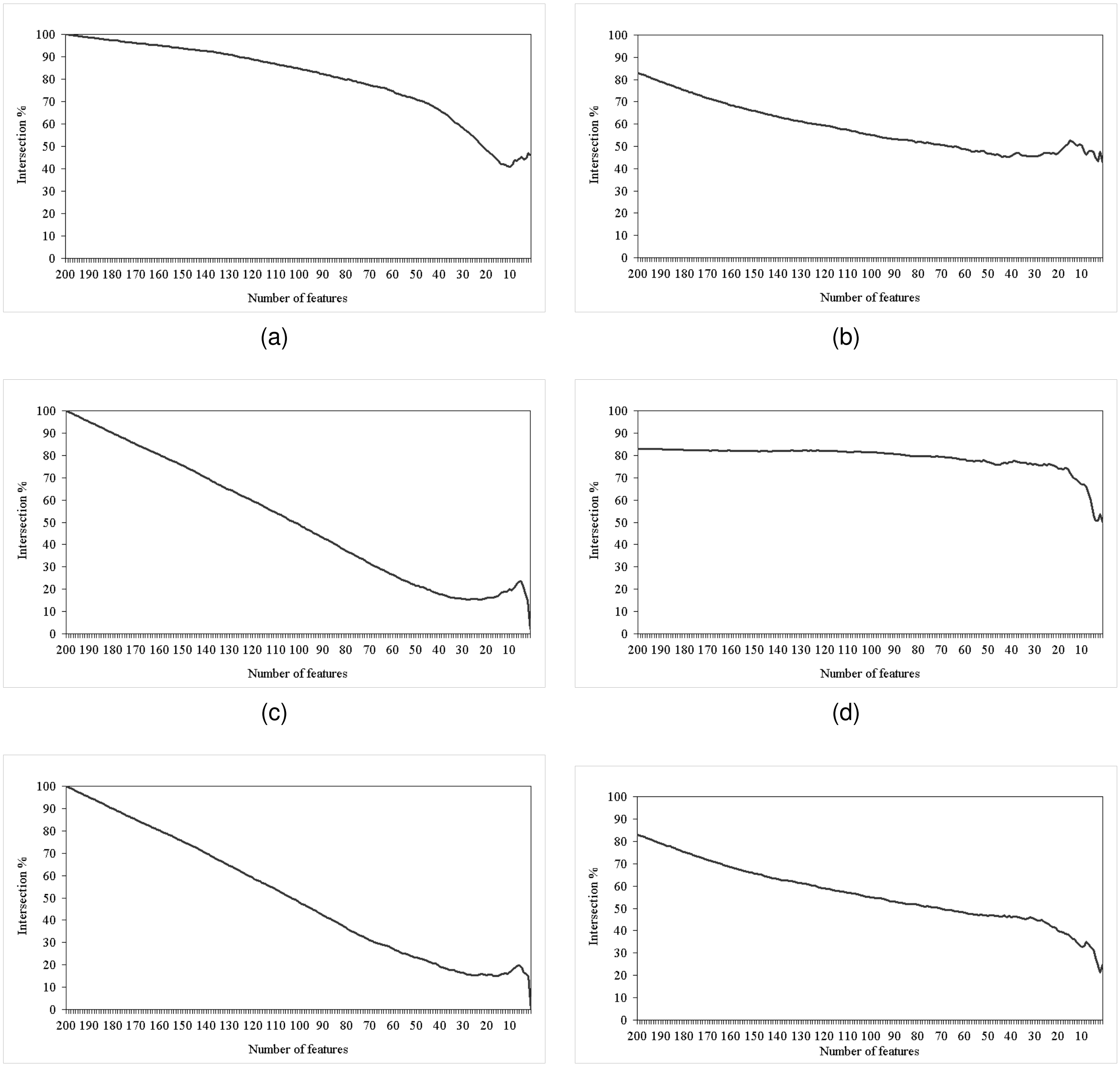
(c)

(d)

(e)

(f)

Fig. 10. Intersection across the top 200 subset of features of (a) IFP vs. RFE, (b) IFP vs. IFP, (c) t test vs. IFP, (d) t test vs. t test, (e) t test vs. RFE and (f) RFE vs. RFE on Moffitt colon cancer dataset.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 11. Intersection across the top 200 subset of features of (a) IFP vs. RFE, (b) IFP vs. IFP, (c) t test vs. IFP, (d) t test vs. t test, (e) t test vs. RFE and (f) RFE vs. RFE on lung cancer dataset.