

*University of Texas, MD Anderson Cancer  
Center*

UT MD Anderson Cancer Center Department of Biostatistics  
Working Paper Series

---

*Year 2004*

*Paper 3*

---

An Empirical Study of Optimism and  
Selection Bias in Binary Classification with  
Microarray Data

Michael L. Lecoche\*

Kenneth Hess<sup>†</sup>

\*Department of Statistics, Rice University, mlecoche@stat.rice.edu

<sup>†</sup>Department of Biostatistics and Applied Mathematics, UT MD Anderson Cancer Center, khess@mdanderson.org

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/mdandersonbiostat/paper3>

Copyright ©2004 by the authors.

# An Empirical Study of Optimism and Selection Bias in Binary Classification with Microarray Data

Michael L. Lecoche and Kenneth Hess

## Abstract

Motivation: Feature subset selection is a very important aspect of performing binary classification using gene expression data. Once feature subsets are obtained, there is the need to evaluate the various models that are formed. This paper considers the problem of how best to evaluate prediction rules formed from the models such that the effects of both optimism and selection bias (i.e., overly optimistic misclassification error rates) are properly taken into account.

Results: An empirical study is presented, in which a 10-fold cross-validation is applied a) internally and b) externally to the feature selection process. These procedures are applied with respect to three supervised learning algorithms and six published two-class microarray datasets. We find that when no cross-validation is performed, optimism bias is present, but it is generally small. Also, we find that when the feature selection is not performed during each stage of the cross-validation process, selection bias is present, but again, it is generally small. Considering all datasets, classifiers, and gene subset sizes together, the average optimism, selection, and total (optimism plus selection) bias estimates are only 4%, 3%, and 7%, respectively. For five of the six datasets, the misclassification rates and bias estimates were very consistent, suggesting that these results should generalize well to other clinical microarray datasets. The same should hold with respect to classifiers, since the three classifiers used in this study behave in different ways, and since there is no clear reason to suspect that the results are connected to the method of classification.

Availability: Datasets are available from the authors upon request.

Contact: [mlecocke@stat.rice.edu](mailto:mlecocke@stat.rice.edu) and [khess@mdanderson.org](mailto:khess@mdanderson.org)

# An Empirical Study of Optimism and Selection Bias in Binary Classification with Microarray Data

Mike Lecocke\* and Kenneth Hess<sup>†</sup>

14th December 2004

## Abstract

**Motivation:** Feature subset selection is a very important aspect of performing binary classification using gene expression data. Once feature subsets are obtained, there is the need to evaluate the various models that are formed. This paper considers the problem of how best to evaluate prediction rules formed from the models such that the effects of both optimism and selection bias (i.e., overly optimistic misclassification error rates) are properly taken into account. **Results:** An empirical study is presented, in which a 10-fold cross-validation is applied a) internally and b) externally to the feature selection process. These procedures are applied with respect to three supervised learning algorithms and six published two-class microarray datasets. We find that when no cross-validation is performed, optimism bias is present, but it is generally small. Also, we find that when the feature selection is not performed during each stage of the cross-validation process, selection bias is present, but again, it is generally small. Considering all datasets, classifiers, and gene subset sizes together, the average optimism, selection, and total (optimism plus selection) bias estimates are only 4%, 3%, and 7%, respectively. For five of the six datasets, the misclassification rates and bias estimates were very consistent, suggesting that these results should generalize well to other clinical microarray datasets. The same should hold with respect to classifiers, since the three classifiers used in this study behave in different ways, and since there is no clear reason to suspect that the results are connected to the method of classification. **Availability:** Datasets are available from the authors upon request. **Contact:** mlecocke@stat.rice.edu and khess@mdanderson.org

## 1 Introduction

### 1.1 Motivation

DNA microarray technology has greatly influenced the realms of biomedical research, with the hopes of significantly impacting the diagnosis and treatment of diseases. Microarrays have the ability to measure the expression levels of thousands of genes simultaneously. They measure how much a

---

\*Department of Statistics, Rice University, Houston, Texas 77005

<sup>†</sup>Department of Biostatistics and Applied Mathematics, UT MD Anderson Cancer Center, Houston, Texas 77030

<sup>‡</sup>To whom correspondence should be addressed.

given type of messenger RNA (mRNA) is present in a tissue sample at a given moment. The wealth of gene expression data that has become available for microarray data analysis has introduced a number of statistical questions to tackle. Some questions are targeted towards various preprocessing stages of a microarray experiment such as RNA hybridization to arrays, image processing, and normalization, while others are geared towards assessing differential expression and identifying profiles for classification and prediction. Within the framework of tumor classification, the types of goals that have been explored include discovering or identifying previously unknown tumor classes, classifying tumors into previously known classes, and identifying “marker genes” that characterize various tumor classes.

In standard discrimination problems, the number of training observations  $N$  is usually much larger than the number of feature variables  $p$ . However, in the context of microarrays, the number of tissue samples  $N$  is usually between 10 and 100, significantly smaller than the thousands of genes considered in a typical microarray analysis. This presents a number of problems to a prediction rule in a discriminant analysis setting. The prediction rule may not even be able to be formed using *all*  $p$  variables (e.g., if using Fisher’s linear discriminant analysis, as discussed in Ambrose and McLachlan (2002)), and even if all the variables could be taken into account in forming the prediction rule, some of them may possess minimal (individual) discriminatory power, potentially inhibiting the performance of the prediction rule when applied to new (unclassified) tumors. Ultimately, with a collection of genes that have high discriminatory power, an effective prediction rule can be developed based on these genes and used to allocate subsequent unclassified tissue samples as one of two classes (e.g., cancer or normal, or perhaps one of two subtypes of a particular cancer).

## 1.2 Supervised Learning

Gene expression data for  $p$  genes over each of  $N$  mRNA samples can be expressed as an  $N \times p$  matrix  $X = (x_{ij})$  ( $i = 1, \dots, N$  and  $j = 1, \dots, p$ ). Each value  $x_{ij}$  corresponds to the expression level for gene  $j$  in sample  $i$ . Each sample would have associated with it a gene expression profile  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in R^p$ , along with its class designation  $y_i$  (response, or dependent variable), which is one of two predefined values among  $\{0, 1\}$ . Using the observed measurements  $X$ , a classifier for two classes is thus a mapping  $G : R^p \rightarrow \{1, 2\}$ , where  $G(\mathbf{x})$  denotes the predicted class,  $\hat{y} = c$ ,  $c \in \{0, 1\}$ , for a sample with feature vector  $\mathbf{x}$ .

The samples already known to belong to certain classes,  $\mathcal{L} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_{n_L}, y_{n_L})\}$ , constitute the training (or learning) set. The training set is used to construct a classifier, which is then used to predict the classes of an independent set of samples (the test set  $\mathcal{T} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_T}\}$ ). This way, the class  $\hat{y}_i$ , ( $i = 1, 2, \dots, n_T$ ) predictions for each test set expression profile  $\mathbf{x}_i$  can be made. Of course, with the true classes  $y_i$ , ( $i = 1, 2, \dots, n_T$ ) of the test set known, a misclassification error rate (*MER*) can then be computed.

### 1.3 Feature Subset Selection

In general, feature (variable) selection is an extremely important aspect of classification problems, since the features selected are used to build the classifier. Careful consideration should be given to the problem of feature subset selection with high-dimensional data. With respect to microarray data, this of course amounts to reducing the number of genes used to construct a prediction rule for a given learning algorithm. There are several reasons for performing feature reduction. Whereas two variables could be considered good predictors individually, there could be little to gain by combining more than one variable together in a feature vector. It has been reported that as model complexity is increased with more genes added to a given model, the proportion of training samples (tissues) misclassified may decrease, but the misclassification rate of new samples (generalization error) would eventually begin to increase; this latter effect being the product of overfitting the model with the training data (Hastie et al., 2001; McLachlan, 1992; Theodoridis, 1999; Xing, 2002; Xiong et al., 2001). Further, if another technology will be used to implement the gene classifier in practice (e.g., to develop diagnostic assays for selected subsets of genes), the cost incurred is often a function of the number of genes. Finally, there is the obvious issue of increased computational cost and complexity as more and more features are included in a model.

### 1.4 Assessing the Performance of a Prediction Rule: Cross-validation

One approach to estimate the error rate of the prediction rule would be to apply the rule to a “held-out” test set randomly selected from among the training set samples. As an alternative to the “hold-out” approach, cross-validation (CV) is very often used, especially when one does not have the luxury of withholding part of a dataset as an independent test set and possibly even another part as a validation set (usually the case with microarray data). Further, the repeatability of results on new data can be assessed with this approach. In general, all CV approaches can fall under the “ $K$ -fold CV” heading. Here, the training set of samples is divided into  $K$  non-overlapping subsets of (roughly) the same size. One of the  $K$  subsets is “held-out” for testing, the prediction rule is trained on the remaining  $K - 1$  subsets, and an estimate of the error rate can then be obtained from applying each stage’s prediction rule to its corresponding test set. This process repeats  $K$  times, such that each subset is treated once as the test set, and the average of the resulting  $K$  error rate estimates forms the  $K$ -fold CV error rate. The whole  $K$ -fold CV process could be repeated multiple times, using different partitions of the data each run and averaging the results, to obtain more reliable estimates. At the expense of increased computation cost, repeated- (10-) run CV has been recommended as the procedure of choice for assessing predictive accuracy of the classification of microarray data (Braga-Neto and Dougherty, 2004; Kohavi, 1995).

## 1.5 Two Approaches to CV: External and Internal CV

With microarray classification problems, the practice has generally been to perform CV only on the classifier construction process, not taking into account feature selection. The feature selection process is applied to the entire set of data (“internal” cross-validation) (Ambroise and McLachlan, 2002; Dudoit and Fridlyand, 2003; McLachlan, 1992). Although the intention of CV is to provide accurate estimates of classification error rates, using CV in this manner means that any inference would be made with respect to the classifier building process only. Leaving out feature selection from the cross-validation process will inevitably lead to a problem with selection bias (i.e., with overly optimistic error rates), as the feature selection would not be based on the particular training set for each CV run. To prevent this selection bias from occurring, an “external” cross-validation process (Ambroise and McLachlan, 2002; Dudoit and Fridlyand, 2003; McLachlan, 1992) should be implemented following the feature selection at each CV stage. That is, the feature selection is performed based only on those samples set aside as training samples at each stage of the CV process, external to the test samples at each stage.

## 1.6 Optimism Bias, Selection Bias, and Total Bias

Several measures of bias are considered in this study. First of all, the difference between the internal CV misclassification error rate (MER) and the resubstitution (training) MER, for any given subset size of genes, is referred to as the optimism bias:

$$\widehat{ob} = MER_{IntCV} - MER_{Resub}. \quad (1)$$

This estimate represents the bias incurred from using the same data to both train the classifier and estimate the performance of the classifier. Feature subset selection for both MER’s used in this computation is based on using all samples of each dataset. Positive values of this estimate signify that the MER’s based on internal CV were higher than those based on using all the data to both train and test the classification rule, and vice-versa for negative values.

Taking this a step further, another bias estimate used in this study is the selection bias, given by:

$$\widehat{sb} = MER_{ExtCV} - MER_{IntCV}. \quad (2)$$

This estimate represents the bias incurred from using the same data to both select the gene subsets and estimate the performance of the classification rule based on these subsets. Positive values of this estimate signify that the MER’s based on including the feature subset selection in the CV process were higher than those based on performing the feature selection outside the CV process using all the samples of a given dataset, and vice-versa for any negative values.

Finally, consider the selection and optimism bias estimates of Eqns. 1 and 2, respectively, as the two components that comprise a third bias estimate – measure of total bias:

$$\widehat{tb} = \widehat{sb} + \widehat{ob} \quad (3)$$

$$= MER_{ExtCV} - MER_{Resub} \quad (4)$$

This estimate represents the bias incurred from using the same data to select gene subsets, train the classifier, and estimate the performance of the classifier. It also takes into account the bias from using the same data to select the gene subsets and estimate the performance of the classification rule based on these subsets. In a sense then, computing this difference between the resubstitution error and the test error (10-fold external CV error) can also provide one with a measure of the degree of overfitting.

## 1.7 Further Thoughts

Careful consideration should be given to the feature subset selection problem when constructing a prediction rule within the framework of a supervised classification problem. Moreover, for a given model based on a selected subset of genes, it is very important to assess the performance of the resulting prediction rule in a way that takes into account both optimism and selection bias. This paper focuses on the implementation of internal and external cross-validation. Empirical results for both of these procedures, for three different learning algorithms and six published microarray datasets, are presented in this paper.

## 2 Methods

### 2.1 Supervised learning methods

In this study, three well known and widely used choices of supervised learning algorithms were implemented: support vector machines (SVM's), DLDA, and  $k$ -NN ( $k=3$  in this study). For more details on each of these classifiers, the reader should refer to Dudoit and Fridlyand (2003) and Dudoit et al. (2000).

### 2.2 Feature subset selection

A univariate-based means of feature subset selection was used to perform gene selection. Rank-based, unequal variance T-tests were performed on each of the genes from the designated training sets of samples among each of the six datasets. In each training set, this resulted in an ordered (by increasing p-value) list of “top genes” for use in generating various “top gene subset size” models.

### 2.3 Repeated-run, external and internal 10-fold CV

To obtain a Monte Carlo type of estimate of both the external and internal 10-fold CV misclassification error rates, the standard 10-fold process is run 10 separate times, and the average of the resulting ten 10-fold CV *MER* estimates is recorded. For a given dataset, for each of the classifiers implemented for a given dataset, the same ten training and test set partitions for a given iteration were used to maintain consistency in interpreting the repeated-run 10-fold CV results.

## 3 Datasets

The following datasets are analyzed in this paper, all of which are from Affymetrix microarrays (Affymetrix, 1999, 2000a,b, 2002). The only preprocessing that was done on each dataset was to standardize the arrays such that they each have zero mean and unit variance (an approach also used in the comparative gene expression classification study of Dudoit et al. (2000)). Standardization of microarray data in this manner achieves a location and scale normalization of the arrays. This was done to ensure that all the arrays of a given dataset were independent of the particular technology used (i.e., reduce the effect of processing artifacts, such as longer hybridization periods, less post-hybridization washing of the arrays, and greater laser power, to name a few). This way, for a given dataset, the values corresponding to individual genes can be compared directly from one array to another. Further, it's been shown that this type of normalization has been effective in preventing the expression values of one array from dominating the average expression measures across arrays (Yang et al., 2001). Currently there is no universally accepted means of normalizing microarray data.

#### Alon et al. (1999) colon cancer dataset

This dataset consists of gene expression levels measured from Affymetrix oligonucleotide arrays (HU6000; quantization software uncertain) for 2000 genes across 62 samples. The binary classes used for analysis are normal (22 samples) and tumor (40 colon tumor samples). As discussed in Li et al. (2001), five colon samples previously identified as being contaminated were omitted (N34, N36, T30, T33, and T36), leaving the total sample size for analysis at 57. See Alon et al. (1999) for more details on this dataset.

#### Golub et al. (1999) leukemia dataset

This dataset consists of gene expression levels (presumably measured from the GeneChip software) from Affymetrix chips (HuGeneFl). The oligonucleotide arrays have 7129 probe sets over 72 samples. The binary classes used for analysis are acute myeloid leukemia (AML; 25 samples) and acute

lymphoblastic leukemia (ALL; 47 samples). See Golub et al. (1999) for more details on this dataset.

#### **Nutt et al. (2003) brain cancer dataset**

This dataset consists of gene expression levels measured from Affymetrix high-density oligonucleotide chips (U95Av2) using the GeneChip software. Each array contains 12625 probe sets over 50 samples. The binary classes used for analysis are glioblastoma (28 samples) and anaplastic oligodendroglioma (22 samples). The downloaded raw expression values were previously normalized by linear scaling such that the mean array intensity for active (“present”) genes was identical for all the scans. See Nutt et al. (2003) for more details on this dataset.

#### **Pomeroy et al. (2002) brain cancer dataset**

This dataset consists of gene expression levels measured from Affymetrix high-density oligonucleotide chips (HuGeneF1) using the GeneChip software. Each chip contains 7129 probe sets. To facilitate the binary classification framework, dataset 'A2' from the project website was used, in which 60 medulloblastoma (MD) samples formed one class and the remaining 30 samples classified as “Other” for the second class (Note: of these 30, there were 10 malignant gliomas (MG), 10 atypical teratoid/rhabdoid tumor (AT/RT), 6 supratentorial primitive neuroectodermal tumors (PNET), and 4 normal cerebellum samples). See Pomeroy et al. (2002) for more details on this dataset.

#### **Shipp et al. (2002) lymphoma dataset**

This dataset consists of gene expression levels measured from Affymetrix chips (HuGeneFL) using the GeneChip software. Each oligonucleotide array contained 7129 probe sets over 77 samples. The two classes used for analysis are diffuse large B-cell lymphoma (DLBCL; 58 samples) and follicular lymphoma (FL; 19 samples). See Shipp et al. (2002) for more details on this dataset.

#### **Singh et al. (2002) prostate cancer dataset**

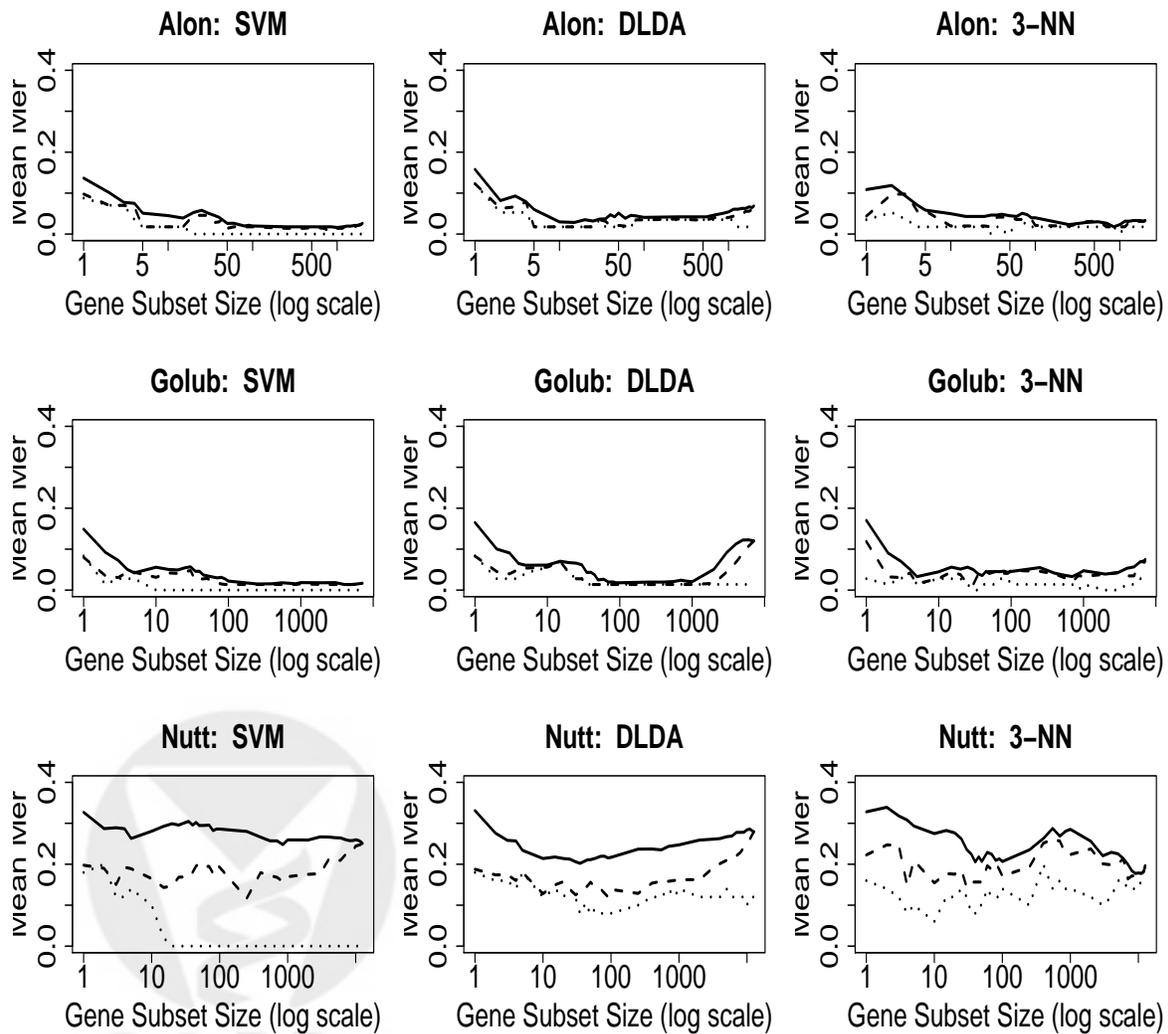
This dataset consists of gene expression levels measured from Affymetrix chips (HU95Av2) using the GeneChip software. The number of arrays available for analysis was 102, with each containing 12600 probe sets. The two classes used for analysis are normal (50 samples) and prostate cancer (52 samples). See Singh et al. (2002) for more details on this dataset.

## 4 Results

### 4.1 Internal and External Cross-Validation

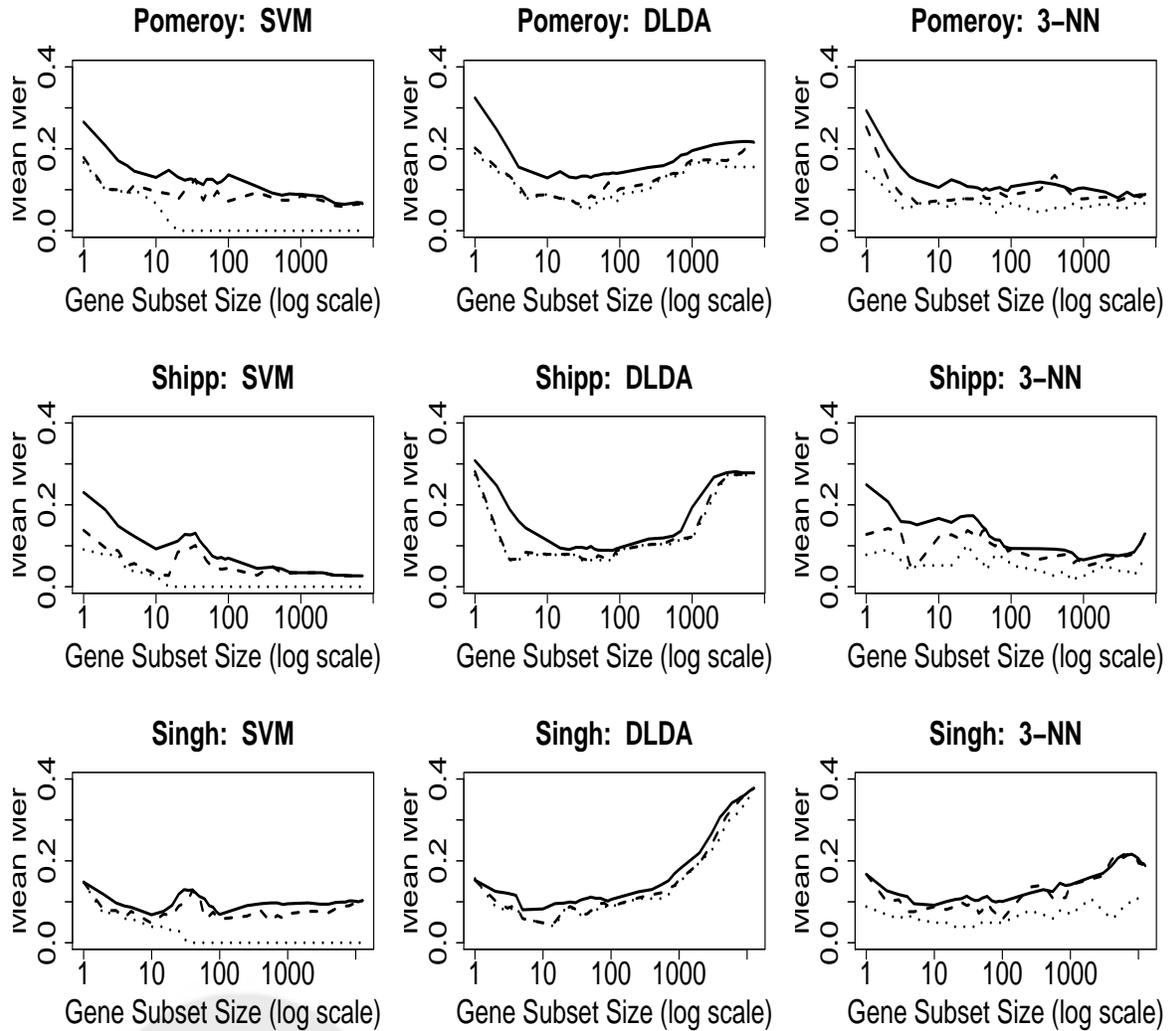
The 10-run external and internal 10-fold CV results for each dataset across a number of gene subset sizes, using each of the classifiers, are shown in Figures 1 and 2. Also included in each plot is the resubstitution error rate.

Figure 1: 10x10-Fold CV; MER vs. Gene Subset Size: Alon, Golub, Nutt Datasets



(a) Solid: ExtCV, Dashed: IntCV, Dotted: Resub

Figure 2: 10x10-Fold CV; MER vs. Gene Subset Size: Pomeroy, Shipp, Singh Datasets



(a) Solid: ExtCV, Dashed: IntCV, Dotted: Resub



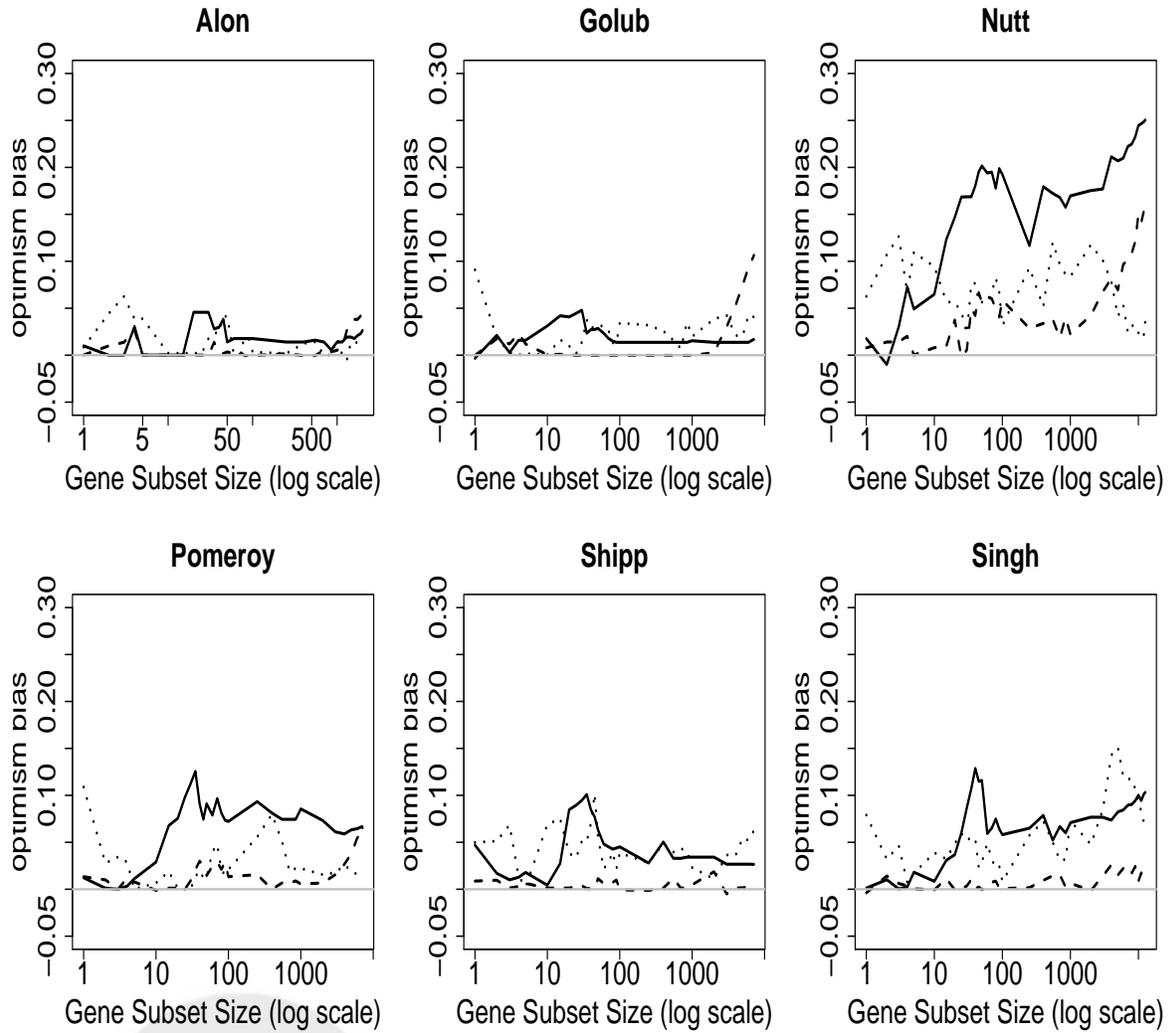
Several observations should be noted from Figures 1 and 2. First, with respect to the resubstitution error curves, in general most of the classifiers' resubstitution error rates across all datasets were slightly lower than the CV curves, as expected. It should also be noted that the Nutt dataset was marked by more variable resubstitution curves than the other datasets for all classifiers. Also, the external CV error rates were larger than the internal ones, as expected. It should be observed, though, that for all datasets except the Nutt one, the difference between the external internal CV results for all classifiers was small. In comparing the datasets, the Alon and Golub datasets had the lowest MER curves for all the classifiers across the gene subset sizes for both the external and internal CV analyses, followed by the Pomeroy, Shipp, Singh, and Nutt datasets. Also, the MER curves from the Shipp, Singh, and especially Nutt datasets were generally more variable than those of the other three datasets, for each of the classifiers. Regarding the effect of gene set size, some general classifier-specific trends can be seen. For SVM and 3-NN, the MER's generally decreased with increasing subset size. For DLDA, the MER's were higher at the smallest and largest gene set sizes and lower at the middle gene set sizes. In general, though, for all but the Nutt dataset, the misclassification rates (whether external or internal) were not drastically different among the datasets. The same holds with respect to classifiers, since the error rates among the three classifiers were not significantly different for a given dataset and subset size.

## 4.2 Optimism Bias, Selection Bias, and Total Bias

Figures 3, 4, and 5 illustrate how the optimism bias, selection bias, and "total bias" (selection bias + optimism bias) estimates, respectively, vary by classifier within dataset across gene subset sizes.



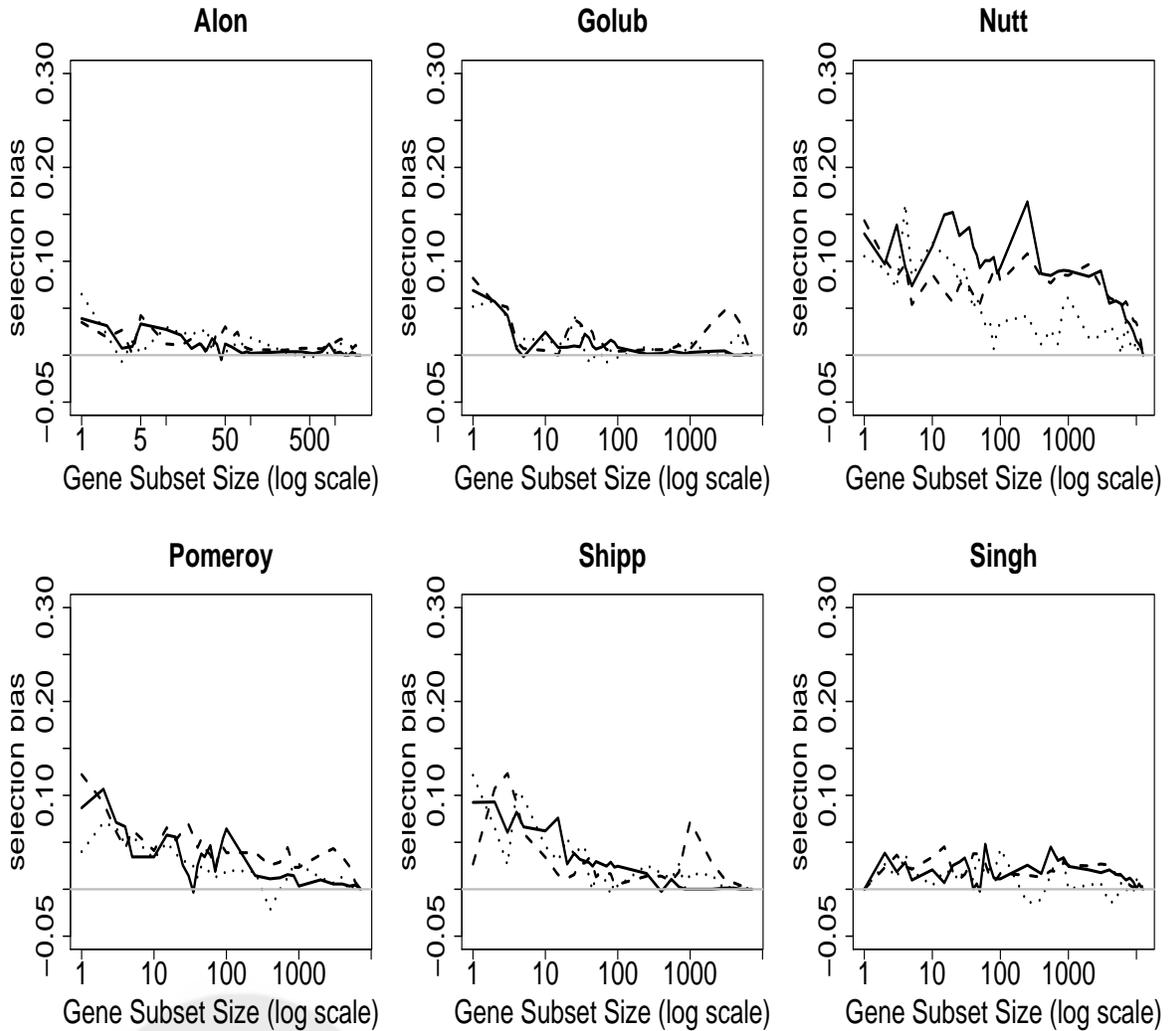
Figure 3: 10x10-Fold CV; Optimism Bias vs. Gene Subset Size



(a) Solid: SVM, Dashed: DLDA, Dotted: 3-NN



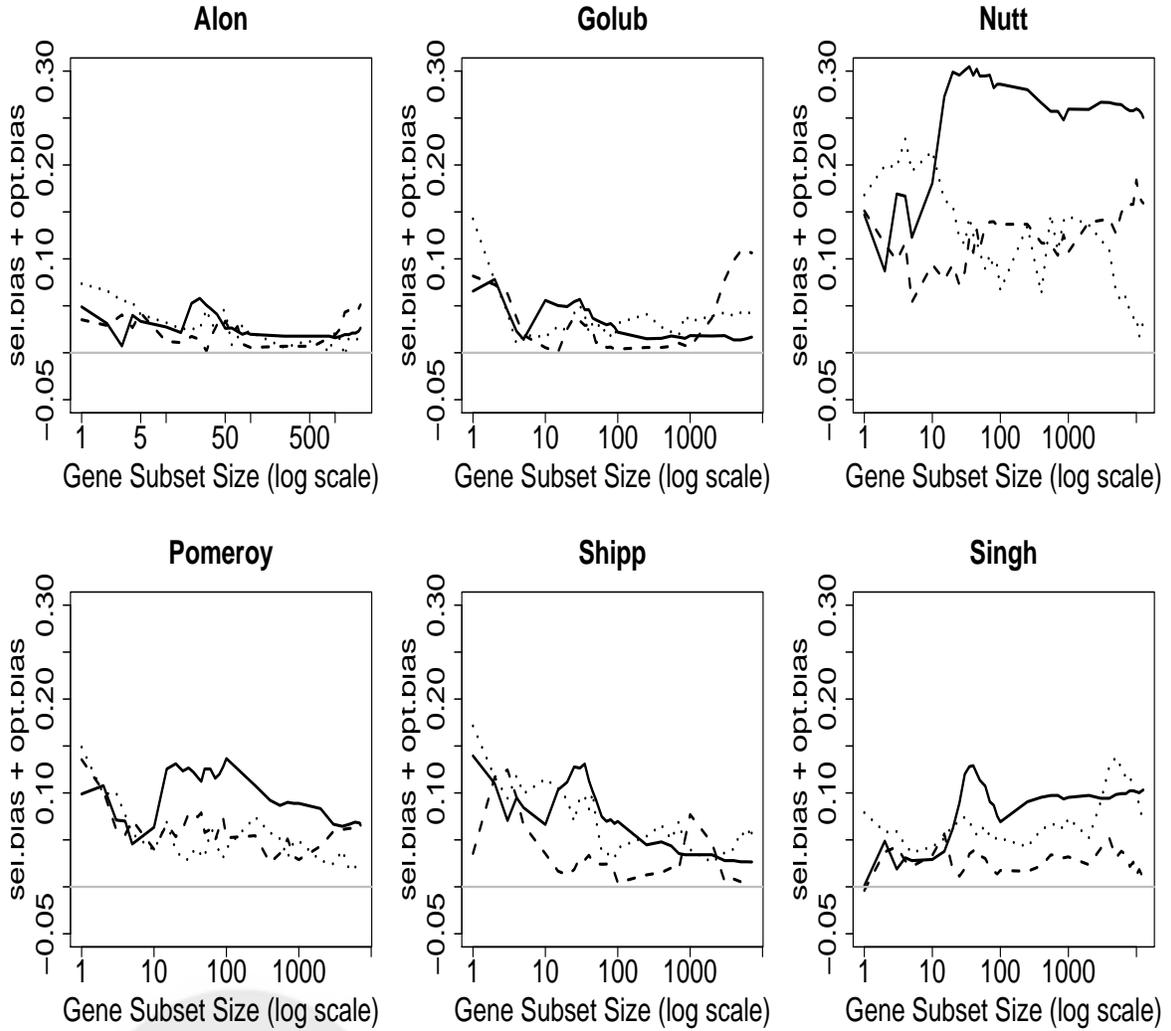
Figure 4: 10x10-Fold CV; Selection Bias vs. Gene Subset Size



(a) Solid: SVM, Dashed: DLDA, Dotted: 3-NN



Figure 5: 10x10-Fold CV; “Total Bias” (Selection + Optimism Bias) vs. Gene Subset Size



(a) Solid: SVM, Dashed: DLDA, Dotted: 3-NN



It should be observed from the plots in Figure 3 that the Alon and Golub datasets generally had smaller optimism bias across the subset sizes than the other datasets, for all classifiers, while the Nutt dataset generally had the highest amount of optimism bias among the three classifiers, across subset sizes. Among the classifiers, DLDA yielded the lowest bias curves for all six datasets. SVM generally led to the highest bias curves for subset sizes larger than 10 across all the datasets (except the Golub one for sizes  $\geq 100$  and the Singh one for sizes beyond 1000). With respect to the selection bias plots in Figure 4, it should be observed that the Alon, Singh, and Golub datasets all demonstrated very little selection bias across all gene subset sizes. The Nutt, Shipp, and Pomeroy datasets all had slightly higher selection bias, especially the Nutt dataset, whose selection bias curves for all classifiers were also much more variable than those of the other datasets. For all dataset and classifier combinations, it should be observed that the selection bias estimates across the majority of the subset sizes were positive, indicating there was at least some penalty in terms of higher MER when performing external 10-fold CV instead of the internal CV approach. In addition, there seemed to be a bigger selection bias for the smaller sized gene subsets (i.e., especially sizes  $\leq 10$ ) than for the bigger sized subsets – an observation consistent among all the classifiers. As was the case with optimism bias, the fact that all the curves were predominantly positive across subset sizes showed that there was clearly selection bias present for all the classification rules shown, across each of the six datasets. Looking at the “total bias” plots in Figure 5, the Nutt dataset again led to the largest total bias values. Overall, there was clearly some bias present for all the classification rules and across all the datasets. To the extent that “total bias” measures overfit, the results indicate that overfitting is not a consistent function of the number of genes included. A final note to keep in mind is that the MER estimates and the bias estimates were not closely related (results not shown in this report).

Finally, boxplots of optimism, selection, and “total” bias are presented in Figures 6, 7, and 8, respectively. These boxplots are collapsed over gene subset sizes.



Figure 6: 10x10-Fold CV; Boxplots of Optimism Bias Over All Subset Sizes

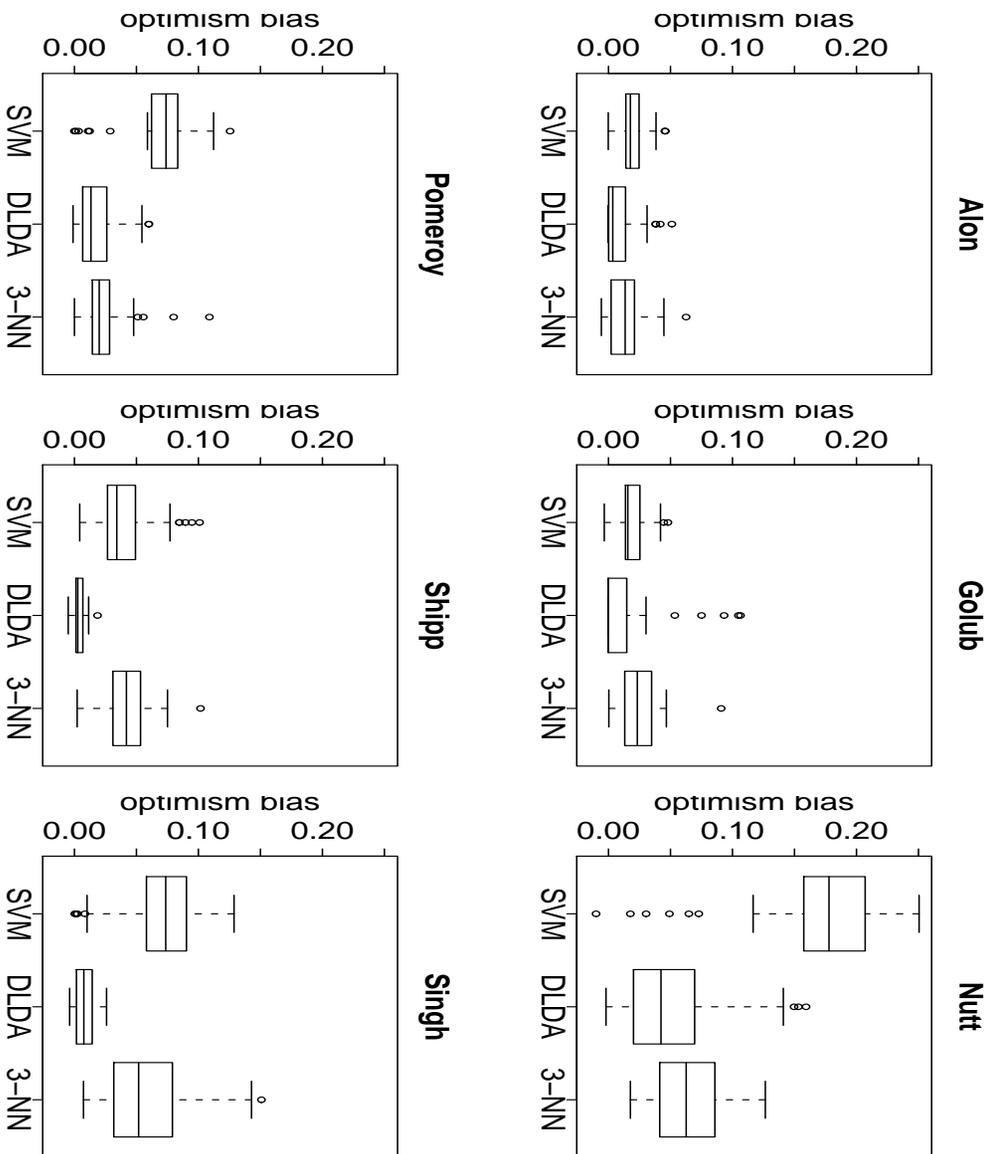


Figure 7: 10x10-Fold CV; Boxplots of Selection Bias Over All Subset Sizes

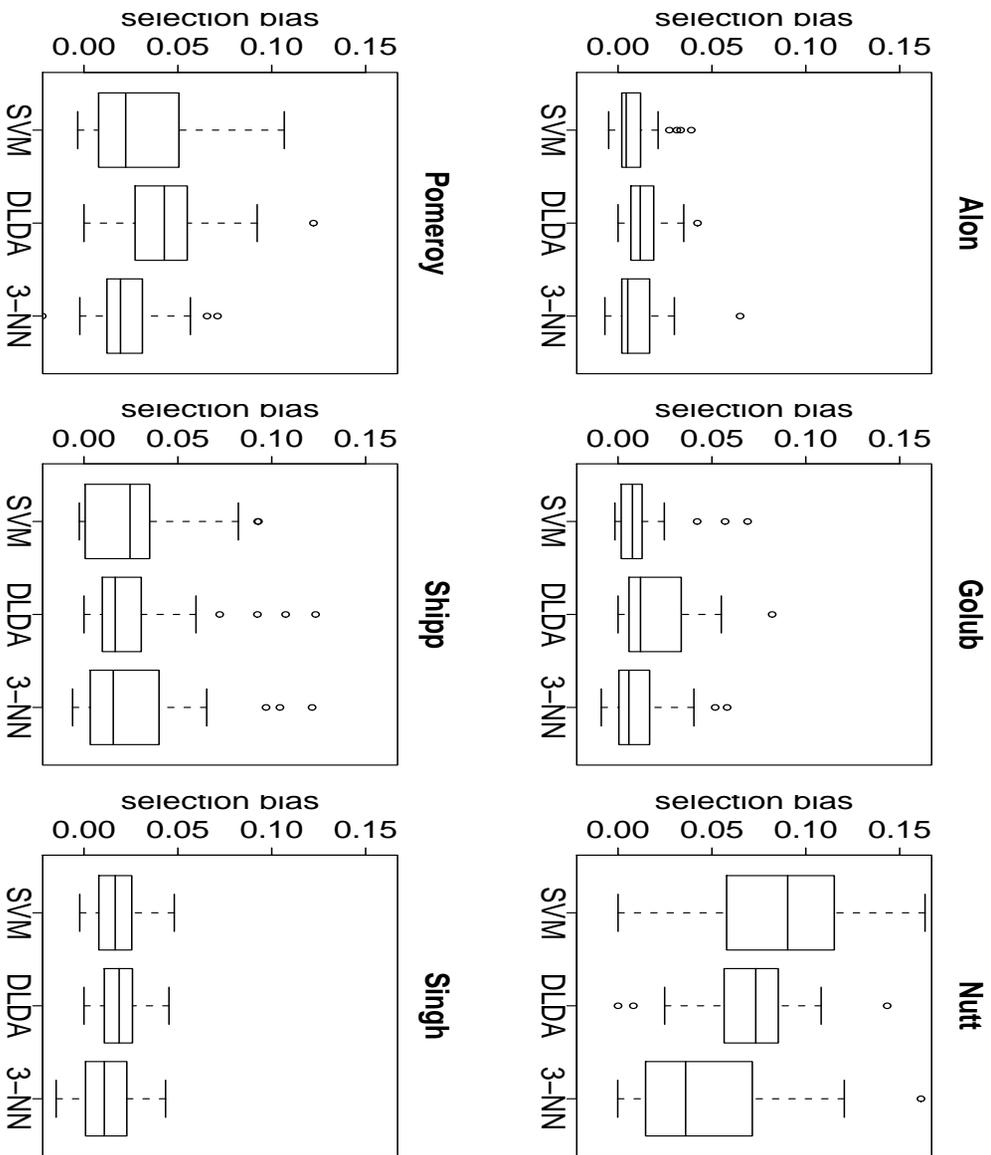
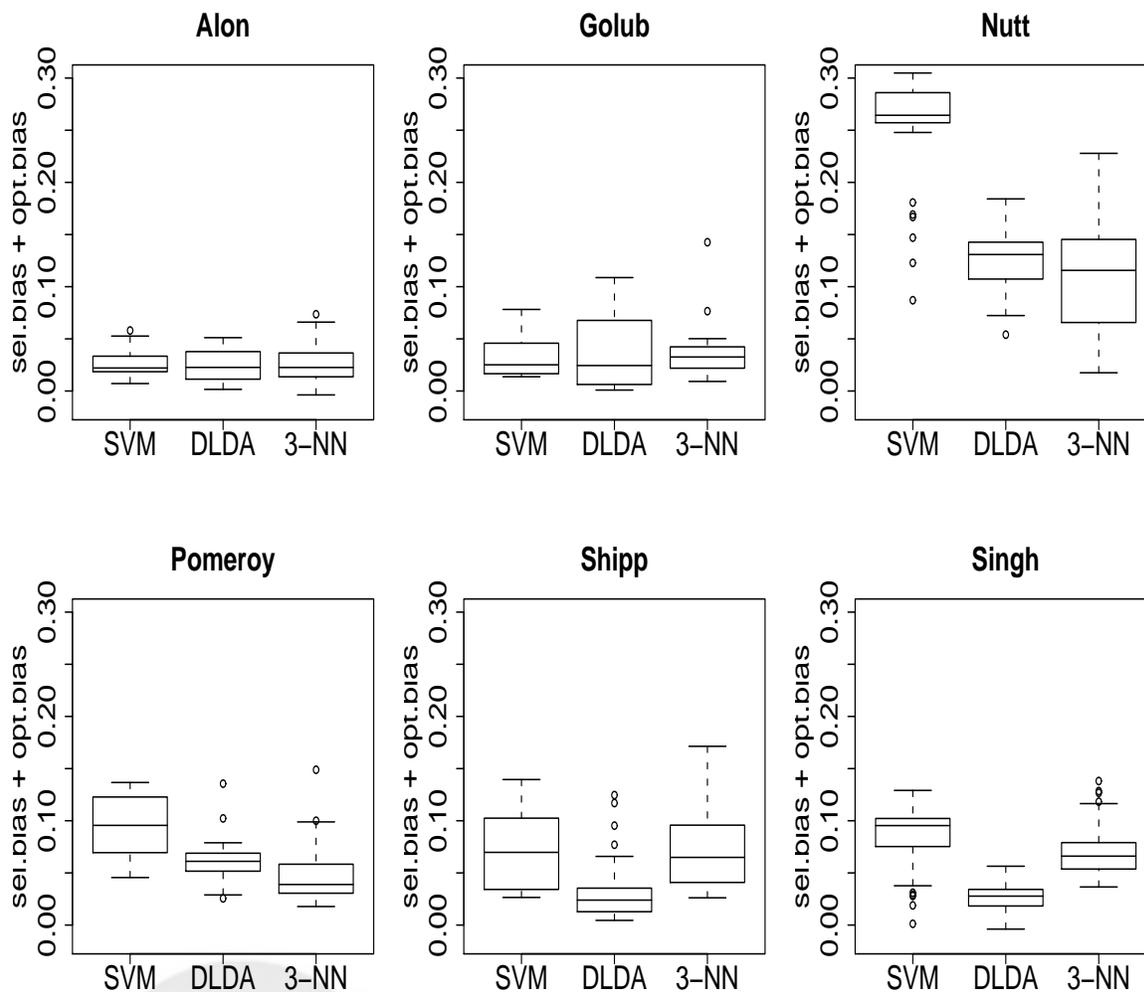


Figure 8: 10x10-Fold CV; Boxplots of “Total Bias” (Selection + Optimism Bias) Over All Subset Sizes



Concluding this section is a table summarizing the means and standard deviations of each of the three bias estimates across gene subset sizes, for each dataset and classifier combination. The empirical grand means (across all gene subset sizes, datasets, and classifiers) for each of the three bias estimates’ means and standard deviations are provided in the last row of Table 1. Overall, considering all datasets, classifiers, and gene subset sizes together, the average optimism, selection, and total bias estimates are only 4%, 3%, and 7%, respectively. It should be noted that if the Nutt data were excluded, these averages become 3%, 2%, and 5%, respectively.

Table 1: Optimism, Selection, and Total Bias Estimates Across All Gene Set Sizes

Dataset	Classifier	Opt.Bias: Mean (SD)	Sel.Bias: Mean (SD)	Total Bias: Mean (SD)
Alon	SVM	0.019 (0.013)	0.009 (0.011)	0.027 (0.013)
	DLDA	0.010 (0.015)	0.014 (0.010)	0.024 (0.015)
Golub	3-NN	0.016 (0.016)	0.011 (0.014)	0.026 (0.019)
	SVM	0.020 (0.011)	0.012 (0.016)	0.032 (0.018)
Nutt	DLDA	0.017 (0.033)	0.021 (0.020)	0.037 (0.037)
	3-NN	0.025 (0.018)	0.012 (0.017)	0.036 (0.024)
	SVM	0.166 (0.066)	0.087 (0.042)	0.253 (0.052)
Pomeroy	DLDA	0.056 (0.045)	0.070 (0.027)	0.126 (0.029)
	3-NN	0.066 (0.031)	0.047 (0.040)	0.113 (0.056)
	SVM	0.067 (0.032)	0.030 (0.028)	0.097 (0.026)
Shipp	DLDA	0.018 (0.017)	0.043 (0.024)	0.061 (0.021)
	3-NN	0.026 (0.023)	0.043 (0.020)	0.048 (0.028)
	SVM	0.043 (0.026)	0.028 (0.029)	0.071 (0.037)
Singh	DLDA	0.004 (0.005)	0.028 (0.031)	0.032 (0.031)
	3-NN	0.044 (0.021)	0.027 (0.032)	0.071 (0.034)
	SVM	0.068 (0.034)	0.017 (0.013)	0.085 (0.032)
Grand Avg	DLDA	0.008 (0.008)	0.019 (0.012)	0.027 (0.012)
	3-NN	0.060 (0.036)	0.012 (0.015)	0.072 (0.026)
		<b>0.041 (0.025)</b>	<b>0.029 (0.022)</b>	<b>0.069 (0.028)</b>

## 5 Discussion

Dudoit et al. (2000) provide an in-depth comparative study of several supervised learning methods for tumor classification based on filtered sets of genes from several published microarray datasets. The learning algorithms used in their study were linear discriminant analysis (LDA), diagonal LDA (DLDA), quadratic LDA (DQDA), classification trees, and  $k$ -NN. The gene selection method implemented was to select the  $p$  genes with largest ratio of between to within-sum-of-squares. In this study, repeated (150) runs of training/test set partitions were performed, with feature selection done only on each training set. The ratio of training to test set samples was 2:1. No cross-validation study was performed. More recently, Dudoit and Fridlyand (2003) applied univariate screening with both a simple t-test and a rank-based t-test (Wilcoxon Test) to analyze a couple of published two-class microarray datasets. The classification schemes they used were  $k$ -NN, DLDA, boosting with trees, random forests, and SVM's. In this study, they applied external and internal CV, but only using leave-one-out (LOO) CV. For both studies, the general conclusion was that the simpler classification methods (e.g., DLDA) performed better than the more complicated ones (e.g.,  $k$ -NN, SVM). In the more recent study, the authors found that the internal LOO CV led to misclassification error rates that were severely biased downward compared to the external CV approach. External and internal CV was also implemented on two published datasets in Ambrose and McLachlan (2002). The samples were randomly divided into 50 different training and test set partitions, with the CV performed only on the training data. They used two schemes for feature selection and classification – backward selection with SVM and forward selection with LDA. No univariate-based approach to perform the feature selection was implemented. They considered the effect of selection bias by performing external 10-fold CV and internal LOO CV (although

unfortunately no internal 10-fold and external LOO results were provided). The average values of the error rate estimates across the multiple runs were obtained for both approaches for each dataset. They found that the internal LOO CV led to overly optimistic error rates compared to the external 10-fold CV process, for both classification schemes and datasets. All these findings stressed the importance of taking into account selection bias when estimating the misclassification error of a classification rule based on microarray data. The current research builds on the findings of these studies, as both internal and external 10-fold CV was implemented using univariate-based feature selection to assess the performance of various prediction rules across multiple two-class microarray datasets.

For each of the six datasets used in this study, we have demonstrated the importance of considering both selection bias and optimism bias in estimating the prediction error for a classification rule constructed from a selected subset of genes from microarrays. To illustrate this, multiple (10) runs of 10-fold internal and external cross-validation were applied to each of the six datasets, using each of three different learning algorithms (SVM, DLDA, 3-NN). For the internal CV approach, the rank-based, unequal variance T-test feature selection process was performed on the entire set of samples for each dataset. With the internal CV approach, the test sets of each partition of the CV process were not external to the feature selection. With the external CV approach, the feature selection was performed at each stage of the CV process, based only on the training set partitions of each stage (i.e., external to the test sets used to evaluate the models). Reinforcing the results from Ambroise and McLachlan (2002), using the external approach, we found that the misclassification errors were only slightly higher than those obtained with internal CV, suggesting that not having the feature subset selection process built into the CV process caused selection bias. In fact, the selection bias estimates across all the classifiers and gene set sizes were small. Only the Nutt dataset had noticeably higher selection bias estimates than those of the other datasets, but even then, on average they were all below 10%. Also, the external CV MER curves were more stable than those from the internal CV approach. Similarly, it was found for all datasets that there was a fair amount of optimism bias present among the classification rules used, as a result of using the same samples to both build the classifier as well as estimate the performance. Again though, the optimism bias estimates across all the classifiers and subset sizes were very small. Only the Nutt dataset had higher optimism bias estimates, especially with the SVM classifier (on average, below 20%). To the extent that the “total bias” estimates measure overfit, the results indicate that overfitting is not a consistent function of the number of genes included in a given model. Overall, it was found in this study that since both the selection and optimism bias estimates across the majority of the gene subset sizes were positive, there was at least some penalty (albeit not substantial) in terms of higher error rates a) when performing external 10-fold CV instead of internal CV, and b) when performing internal CV instead of using all samples for both building the classifiers and evaluating them. However, because the magnitude of the bias estimates in general across all the datasets, classifiers, and gene set sizes were small, and given the computational burdens for correcting these biases, perhaps there are some situations when the resubstitution estimate may

be adequate. Ultimately, the misclassification rates and bias estimates did not vary significantly by dataset, for five of the six datasets at least (the Nutt one being the exception), suggesting that these results should generalize well to other clinical microarray datasets. The same generalization ability should hold with respect to classifiers, since the three classifiers used function in different ways, and since there is no clear reason to suspect that the results are connected to the method of classification.

## **Acknowledgment**

The authors sincerely thank Jeff Morris for careful reading of the manuscript and useful suggestions.



## References

- Affymetrix (1999). Genechip analysis suite. User guide, version 3.3, Affymetrix.
- Affymetrix (2000a). Expression analysis technical manual. Technical report, Affymetrix.
- Affymetrix (2000b). Genechip expression analysis. Technical manual, Affymetrix.
- Affymetrix (2002). Statistical algorithms description document. Technical report, Affymetrix.
- Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.*, pages 6745–6750.
- Ambrose, C. and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the USA*, 99(10):6562–6566.
- Braga-Neto, U. and Dougherty, E. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380.
- Dudoit, S. and Fridlyand, J. (2003). *Classification in Microarray Experiments*, chapter 3, pages 93–158. Chapman and Hall/CRC. Appearing in 'Statistical Analysis of Gene Expression Microarray Data' (ed. Terry Speed).
- Dudoit, S., Fridlyand, J., and Speed, T. (2000). Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576, University of California, Berkeley, Dept. of Statistics.
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143, Montreal, Canada. in 'Proceedings of the 14th International Joint Conference on Artificial Intelligence' (IJCAI-95).
- Li, L., Weinberg, C., Darden, T., and Pedersen, L. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. *Bioinformatics*, 17(12):1131–1142.
- McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

- Nutt, C., Mani, D., Betensky, R., Tamayo, P., Cairncross, J., Ladd, C., Pohl, U., Hartmann, C., McLauhlin, M., Batchelor, T., Black, P., von Deimling, A., Pomeroy, S., TR, T. G., and Louis, D. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, 63:1602–1607.
- Pomeroy, S., Tamayo, P., Gaasebeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E., and Golub, T. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442.
- Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutor, J., Aguiar, R., Gaasenbeer, M., Angelo, M., Reich, M., Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, A., Mesirov, J., Neuberg, D., Lander, E., Aster, J., and Golub, T. (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74.
- Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D’Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T., and Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209.
- Theodoridis, S. (1999). *Pattern Recognition*. Academic Press, San Diego.
- Xing, E. (2002). *Feature Selection in Microarray Analysis*, chapter 6, pages 110–131. Kluwer Academic Publishers. Appearing in ‘A Practical Approach to Microarray Data Analysis’ (eds. D. Berrar, W. Dubitzky, and M. Granzow).
- Xiong, M., Fang, X., and Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Research*, 11:1878–1887.
- Yang, Y., Speed, T., Dudoit, S., and Luu, P. (2001). Normalization for cdna microarrya data. In Bittner, M. et al., editors, *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proc. SPIE*, pages 141–152.

