Appendix E. Supplementary Information

1. Classification without gene selection using Design II

Table E1. *Performance results (**accuracies**) without gene selection obtained using LOOCV with 10-fold cross-validation for parameter selection (Design II).*

		Multicategory classification					
Method		**9_Tumors**	**11_Tumors**	**14_Tumors**	**Brain_Tumor1**	**Brain_Tumor2**	**Leukemia1**
MC-SVM	**OVR**	**63.33%**	**95.40%**	74.35%	**91.11%**	76.00%	**97.22%**
	OVO	55.00%	89.66%	44.48%	**91.11%**	**78.00%**	**97.22%**
	DAGSVM	55.00%	89.66%	46.10%	**91.11%**	**78.00%**	95.83%
	WW	61.67%	94.25%	64.61%	**91.11%**	**78.00%**	**97.22%**
	CS	**63.33%**	94.83%	**75.32%**	90.00%	76.00%	**97.22%**
non-SVM	**KNN**	38.33%	75.29%	50.00%	86.67%	60.00%	80.56%

		Multicategory classification			*Binary classifcation*		**Averages**
Method		**Leukemia2**	**Lung_Cancer**	**SRBCT**	**Prostate_Tumor**	**DLBCL**	
MC-SVM	**OVR**	**98.61%**	96.06%	**100.00%**	**92.16%**	**97.40%**	**89.24%**
	OVO	95.83%	95.07%	**100.00%**	**92.16%**	**97.40%**	85.08%
	DAGSVM	95.83%	95.57%	**100.00%**	**92.16%**	**97.40%**	85.15%
	WW	97.22%	95.07%	**100.00%**	**92.16%**	**97.40%**	88.06%
	CS	97.22%	**96.55%**	**100.00%**	**92.16%**	**97.40%**	89.09%
non-SVM	**KNN**	86.11%	91.13%	84.34%	79.41%	88.31%	74.56%

2. Statistical comparison among classifiers

Table E2. *P-values of the statistical test that compares accuracies of all algorithms with ones of the best MC-SVM methods (CS, OVR, and WW) using classifications obtained by nested stratified 10-fold cross-validation design (Design I) for all 11 datasets <u>without gene selection</u>. Bold p-values correspond to cases when we cannot reject null hypothesis at the 0.05 level.*

Method		**Comparison: all algorithms versus**		
		OVR	**WW**	**CS**
MC-SVM	**OVR**	-	**0.981**	**0.326**
	OVO	0.016	0.009	0.012
	DAGSVM	0.009	0.011	0.009
	WW	**0.052**	-	**0.064**
	CS	**0.875**	**0.990**	-
non-SVM	**KNN**	0.006	0.005	0.004
	NN	0.001	0.002	0.001
	PNN	0.003	0.005	0.006

Table E3. *P-values of the statistical test that compares accuracies of all algorithms with ones of the best MC-SVM methods (CS, OVR, and WW) and NN using classifications obtained by nested stratified 10-fold cross-validation design (Design I) for 4 datasets (9_Tumors, 14_Tumors, Brain_Tumor1, Brain_Tumor2) <u>with gene selection</u>. Bold p-values correspond to cases when we cannot reject null hypothesis at the 0.05 level.*

Method		**Comparison: all algorithms versus**			
		OVR	**WW**	**CS**	**NN**
MC-SVM	**OVR**	-	**0.965**	**0.415**	**0.956**
	OVO	0.017	0.021	0.011	0.024
	DAGSVM	0.016	0.031	0.014	0.043
	WW	**0.071**	-	**0.052**	**0.502**
	CS	**0.751**	**0.972**	-	**0.977**
non-SVM	**KNN**	0.027	0.048	0.028	0.043
	NN	**0.083**	**0.612**	**0.072**	-
	PNN	0.018	0.028	0.020	0.027

3. Using inverse power-law curves to explain observed classification accuracy

We were interested to what extent the number of samples, categories, and variables (prior to gene selection) can explain classification accuracy observed in the datasets studied. We explored the relations between classification accuracy and the following possible performance predictors: Θ_a = number of samples; Θ_b = number of categories; Θ_c = number of variables; Θ_d = number of samples divided by number of categories; Θ_e = number of samples divided by number of variables; and Θ_f = number of samples divided by the product of number of variables times number of categories. We fitted inverse power-law curves of the type $p = 100\% - \alpha \cdot \Theta^{-\beta}$, where p is the classification performance (accuracy), Θ is one of the performance predictors, α and β are model parameters. The choice of inverse power-law curve is motivated by prior machine learning research in learning curves [Cortes1993]. The curve was fitted using the nonlinear Levenberg-Marquardt least squares iterative method available in the Matlab Optimization Toolbox [Venkataraman2002].

The inverse power-law curve for explaining performance of the best MC-SVM methods OVR, WW, and CS as a function of Θ_f (number of samples divided by the product of number of variables prior to gene selection[1] times number of categories) fitted best according to Euclidian norm metric (**Figure E1**). The resulting curve is $p = 100\% - 0.001021 \cdot \Theta_f^{-1.431305}$. According to this formula, whenever $\Theta_f > 1.6 \cdot 10^{-3}$, MC-SVM techniques CS, WW, and OVR produce multicategory cancer diagnoses with accuracies > 90%. We note that this heuristic rule did not hold in our experiments neither when gene selection was employed[2] nor when RCI performance metric was used instead of accuracy.

The analysis presented above in an initial step towards this research. It is important to note that curve fitting procedure used in this study is very simplistic since it does not incorporate predictors describing degree of biological difficulty and assumes that datasets and learning tasks used in this study are representative. More complex approaches to modeling performance of the classifier as a function of dataset characteristics may also applicable for this domain (e.g., [Mukherjee2003]).

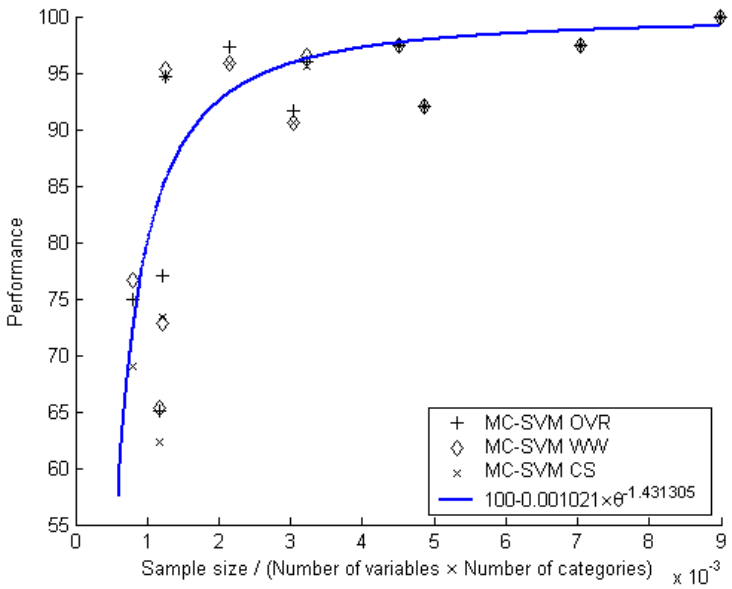

Figure E1. *Number of samples divided by the product of the number of categories times the number of variables (horizontal axis) explains observed classification accuracy of the best MC-SVM classifiers, OVR, WW, and CS (vertical axis).*

[1] The original *SRBCT* data from cDNA array with 6567 genes is not publicly available. We used a public version of this dataset with 2308 genes with red intensity greater than 20 across all samples [Khan2001]. While fitting inverse power-law curves, we treat this version of *SRBCT* dataset as if no gene selection has been performed.

[2] Given an arbitrary gene expression dataset with more samples than categories, it is always possible to select 1 best gene according to some gene scoring criterion. Hence, Θ_f will be > 1. Based on our heuristic rule, predictive accuracy should be greater than 99.99%, which obviously may not be the case.

4. Ensemble classification

Table E4. *Performance results (**accuracies**) of ensemble classification applied to outputs of learners that were used without gene selection. Ensembles were constructed both based on outputs of only MC-SVM classifiers and based on outputs of all (MC-SVM and non-SVM) classifiers. These results were obtained using nested stratified 10-fold cross-validation design (Design I). MC-SVM ensemble methods indicated with "*" were applied using extended ranges for optimization of SVM parameters: costs = {1e-5, 1e-4, 1e-3, 1e-2, 0.1, 1, 10, 100, 1000, 10000} and degrees = {1,2,3,4}. MC-SVM ensemble methods indicated with "+" were used with RBF kernel and the following ranges for optimization of SVM parameters: cost = {0.0001, 0.01, 1, 100} and values of σ (RBF kernel parameter) = {0.0001, 0.001, 0.01, 1}.*

			Multicategory classification					
Input data of ensemble learner	**Method**		**9_Tumors**	**11_Tumors**	**14_Tumors**	**Brain_Tumor1**	**Brain_Tumor2**	**Leukemia1**
Outputs of MC-SVM methods	**DTs**		56.43%	93.62%	70.04%	**91.67%**	**79.50%**	**97.50%**
	MC-SVM	**OVR**	19.10%	34.02%	21.89%	78.36%	56.17%	88.93%
		OVO	32.38%	84.05%	40.04%	90.56%	72.50%	96.07%
		DAGSVM	34.05%	84.61%	39.70%	90.56%	72.50%	96.07%
		OVR*	17.19%	30.49%	23.20%	81.47%	63.83%	88.93%
		OVO*	35.48%	86.87%	50.76%	90.56%	72.50%	96.07%
		DAGSVM*	41.67%	87.37%	48.81%	90.56%	70.50%	96.07%
		OVR+	45.33%	90.27%	51.08%	90.56%	68.83%	**97.50%**
		OVO+	59.00%	92.04%	57.14%	90.56%	70.33%	96.07%
		DAGSVM+	59.00%	92.04%	54.90%	90.56%	68.33%	96.07%
	Majority voting		63.90%	94.68%	70.57%	90.56%	75.33%	**97.50%**
Outputs of all methods	**Majority voting**		61.90%	93.04%	67.06%	**91.67%**	77.83%	**97.50%**
Best results of non-ensemble classifiers			**65.33%**	**95.30%**	**76.60%**	**91.67%**	77.83%	**97.50%**

			Multicategory classification			***Binary classifcation***	
Input data of ensemble learner	**Method**		**Leukemia2**	**Lung_Cancer**	**SRBCT**	**Prostate_Tumor**	**DLBCL**
Outputs of MC-SVM methods	**DTs**		**97.32%**	96.05%	**100.00%**	**92.00%**	**97.50%**
	MC-SVM	**OVR**	76.43%	84.76%	86.63%	**92.00%**	**97.50%**
		OVO	94.46%	95.55%	**100.00%**	**92.00%**	**97.50%**
		DAGSVM	94.46%	95.55%	**100.00%**	**92.00%**	**97.50%**
		OVR*	82.14%	85.21%	86.63%	**92.00%**	**97.50%**
		OVO*	94.46%	96.05%	**100.00%**	**92.00%**	**97.50%**
		DAGSVM*	94.46%	96.05%	**100.00%**	**92.00%**	**97.50%**
		OVR+	**97.32%**	96.05%	**100.00%**	**92.00%**	**97.50%**
		OVO+	94.46%	96.05%	**100.00%**	**92.00%**	**97.50%**
		DAGSVM+	94.46%	96.05%	**100.00%**	**92.00%**	**97.50%**
	Majority voting		95.89%	96.05%	**100.00%**	**92.00%**	**97.50%**
Outputs of all methods	**Majority voting**		95.89%	95.09%	**100.00%**	**92.00%**	**97.50%**
Best results of non-ensemble classifiers			**97.32%**	**96.55%**	**100.00%**	**92.00%**	**97.50%**

Table E5. *Performance results (**accuracies**) of ensemble classification applied to outputs of learners that were used with gene selection on three datasets: 9_Tumors*, *Brain_Tumor1*, and *Brain_Tumor2. Ensembles were constructed both based on outputs of only MC-SVM classifiers and based on outputs of all (MC-SVM and non-SVM) classifiers. These results were obtained using nested stratified 10-fold cross-validation design (Design I).*

Input data of ensemble learner	Method	9_Tumors	Brain_Tumor1	Brain_Tumor2
Outputs of MC-SVM methods	DTs	66.19%	88.31%	79.17%
	Majority voting	71.52%	90.42%	84.00%
Outputs of all methods	DTs	59.52%	88.44%	73.50%
	Majority voting	68.43%	90.42%	82.00%
Best results of non-ensemble classifiers		*74.86%*	*92.67%*	*85.67%*
FS method and number of features		*BW, 1000*	*KW, 500*	*KW, 500*

5. Comparison with previously published results

Table E6. *Comparison of classification results obtained in the present study with previously published studies on the same datasets. If multiple studies were present for a dataset, we selected one that employed the most similar learning task and followed the most similar experimental design compared to our study.*

	Our study			Published studies			
Dataset	**Accuracy without gene selection**		**Accuracy with gene selection**	**Major differences in dataset preparation between our study and published studies**	**Experimental design**	**Methods and Results**	**Reference**
	Design I	**Design II**	**Design I**				
9_Tumors	**65.33%**	**63.33%**	**74.86%**	This dataset was used for a different task - prediction of chemosensitivity			[Staunton2001]
11_Tumors	**95.30%**	**95.40%**	**Not analyzed**	The dataset used by [Su2001] contains 1 more sample compared to our study. That sample was not included in the publicly available version of the dataset.	The study performed LOOCV on 100 samples (training set) and then tested the classifier on 75 samples (testing set).	**Methods:** Gene selection procedure involved the following three major steps: (1) minimal thresholding of gene expression data; (2) selection of genes with small p-values according to Wilcoxon test; (3) further gene selection by use of SVMs. The study applied a variant of MC-SVM OVR method for classification. **Results:** The study achieved **97%** accuracy for LOOCV on the training set (100 samples) and **95%** accuracy on the testing set (75 samples).	[Su2001]
14_Tumors	**76.60%**	**75.32%**	**76.60%**	Compared to our study, the dataset used in [Ramaswamy2001] did not contain 90 samples from normal tissues and spanned only over 14 diagnostic categories (not 26 as used in our study). Furthermore, [Ramaswamy2001] excluded 20 "*poorly differentiated samples*" from the main testing set, although their technical quality was "indistinguishable from the other samples in the study". Unlike [Ramaswamy2001], we treated all testing samples equally and included these 20 samples in the testing set.	The study performed LOOCV on 144 samples (training set) and then tested the classifier on two testing sets - 54 samples (main testing set) and 20 samples ("*poorly differentiated*" testing set).	**Methods:** Genes were ranked and selected according to their contribution to the solution of classification problem by SVMs. The study employed MC-SVM methods (OVO and OVR) and variants of KNN and WV algorithms for classification. **Results:** Best classification results were obtained without gene selection and using MC-SVM OVR classifier: **78%** accuracy for LOOCV on the training set (144 samples), **78%** accuracy on the main testing set (54 samples), and **30%** accuracy on the "*poorly differentiated*" testing set (20 samples). Classification results with gene selection were inferior compared to results obtained by utilizing all genes.	[Ramaswamy2001]
Brain_Tumor1	**91.67%**	**91.11%**	**92.67%**	Unlike [Pomeroy2002], where researchers experimented only with a 42 sample-dataset, our study used a version of this dataset with 90 samples.	The study performed LOOCV on 42 samples.	**Methods:** Genes were selected with S2N metric. KNN algorithm was used for classification. **Results:** The study obtained **83.33%** accuracy.	[Pomeroy2002]
Brain_Tumor2	**77.83%**	**78.00%**	**85.67%**	Unlike our study which solved this classification problem with 4 specific diagnostic categories, [Nutt2003] solved this problem with 2 more general categories.	The study performed LOOCV on 21 samples (training set) and then tested the classifier on 29 samples (testing set).	**Methods:** Gene selection procedure involved the following two steps: (1) minimal thresholding of gene expression data; (2) selection of genes with S2N metric. KNN algorithms was used for classification. **Results:** The study achieved **86%** accuracy for LOOCV on the training set (21 samples) and **59%** accuracy on the testing set (29 samples).	[Nutt2003]

This table is continued on the next page.

This table begins on the previous page.

	Our study			Published studies			
Dataset	**Accuracy without gene selection**		**Accuracy with gene selection**	**Major differences in dataset preparation between our study and published studies**	**Experimental design**	**Methods and Results**	**Reference**
	Design I	**Design II**	**Design I**				
DLBCL	**97.50%**	**97.40%**	**Not analyzed**	No differences	The study performed LOOCV on 77 samples.	**Methods:** Genes were selected with S2N metric. Classification was performed with WV algorithm. **Results:** The study obtained **92%** accuracy.	[Shipp2002]
Leukemia1	**97.50%**	**97.22%**		No differences	The study performed LOOCV on 34 out of 72 samples.	**Methods:** Gene selection procedure involved the following two steps: (1) minimal thresholding of gene expression data; (2) gene selection with BW metric. [Lee2003] used MC-SVM MSVM algorithm for classification. **Results:** The study obtained **97%** accuracy.	[Lee2003]
Leukemia2	**97.32%**	**98.61%**		The dataset used by [Armstrong2002] contains 5 less samples compared to our study (and publicly available version of the data).	The study performed LOOCV on 57 samples (training set) and then tested the classifier on 10 samples (testing set).	**Methods:** Gene selection was performed with S2N metric. KNN algorithm was used for classification. **Results:** The study achieved **95%** accuracy for LOOCV on the training set (57 samples) and **90%** accuracy on the testing set (10 samples).	[Armstrong2002]
Lung_Cancer	**96.55%**	**96.55%**		[Aliferis2003a] solved the following three binary classification tasks: *Task 1* - normal versus cancerous (203 samples); *Task 2* - adeno versus squamous (160 samples); and *Task 3* - metastatic adeno versus non-metastatic adeno (139 samples).	The study used nested stratified N-fold cross-validation design (N = 5, 5, and 7 for tasks 1, 2, and 3, respectively).	**Methods:** Gene selection was performed by SVM-based recursive feature elimination and univariate attribute filtering (UAF). Classification algorithms SVM, Neural Networks, and KNN were used. **Results:** The study achieved **99.64%**, **99.07%**, and **96.83%** area under ROC curve (AUC) for experiments without gene selection for tasks 1, 2, and 3, respectively. When gene selection was performed, the study achieved **99.80%**, **99.63%**, and **97.62%** AUC for tasks 1, 2, and 3, respectively.	[Aliferis2003a]
Prostate_Tumor	**92.00%**	**92.16%**		No differences	The study performed LOOCV on 102 samples.	**Methods:** Gene selection was performed with S2N metric. KNN algorithm was used for classification. **Results:** The study achieved **94%** accuracy.	[Singh2002]
SRBCT	**100.00%**	**100.00%**		The dataset used by [Khan2001] contains 5 more non-SRBCT samples compared to our study.	The study performed LOOCV on 63 samples (training set) and then tested the classifier on 25 samples, including 5 non-SRBCT samples (testing set).	**Methods:** Minimal thresholding was applied to gene expression data. PCA was used to reduce dimensionality to 10 first principal components. Gene selection was performed by measuring sensitivity of the outputs with respect to inputs. Neural Networks were used for classification. **Results:** The study achieved **100%** accuracy for LOOCV on the training set (63 samples) and **100%** accuracy on the testing set (25 samples).	[Khan2001]

6. Classification results with Decision Trees and Weighted Voting

Table E7. *Performance results (**accuracies**) without gene selection obtained using a nested stratified 10-fold cross-validation design (Design I) for DT, WV OVR, and WV OVO classifiers. These results are further improved by gene selection (see **Table E8**). The last column in the bottom table reports average performance computed over datasets.*

		Multicategory classification					
Method		**9_Tumors**	**11_Tumors**	**14_Tumors**	**Brain_Tumor1**	**Brain_Tumor2**	**Leukemia1**
non-SVM	**DT**	28.29%	68.82%	53.23%	70.42%	63.17%	83.39%
	WV OVR	30.14%	30.52%	25.69%	22.14%	47.00%	18.04%
	WV OVO	25.29%	34.22%	28.74%	26.86%	36.83%	63.57%

		Multicategory classification			*Binary classifcation*		**Averages**
Method		**Leukemia2**	**Lung_Cancer**	**SRBCT**	**Prostate_Tumor**	**DLBCL**	
non-SVM	**DT**	85.00%	86.14%	82.74%	89.00%	88.21%	72.58%
	WV OVR	41.79%	5.45%	23.65%	59.73%	75.36%	34.50%
	WV OVO	34.64%	40.24%	53.83%	59.73%	75.36%	43.57%

Table E8. *Performance results (**accuracies**) with gene selection obtained using a nested stratified 10-fold cross-validation design (Design I) for DT, WV OVR, and WV OVO classifiers and two datasets, Brain_Tumor1 and Brain_Tumor2.*

Method		**Brain_Tumor1**	**Brain_Tumor2**
non-SVM	**DT**	75.89%	72.67%
	WV OVR	34.44%	47.00%
	WV OVO	47.31%	36.83%

References

[Aliferis2003a] Aliferis C.F., I. Tsamardinos, P. Massion, A. Statnikov, N. Fananapazir, D. Hardin. "Machine Learning Models For Classification Of Lung Cancer and Selection of Genomic Markers Using Array Gene Expression Data" In Proceedings of the 16th International Florida Artificial Intelligence Research Society (FLAIRS) Conference, 2003.

[Armstrong2002] Armstrong, S., et al. "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia", Nature Genetics, volume 30, January 2002.

[Cortes1993] Cortes, C., Lawrence D. Jackel, Sara A. Solla, Vladimir Vapnik, John S. Denker: Learning Curves: Asymptotic Values and Rate of Convergence. NIPS 1993: 327-334.

[Khan2001] Khan, J., et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature Medicine, volume 7, Number 6, June 2001.

[Lee2003] Lee, Yoonkyung and Cheol-Koo Lee. "Classification of multiple cancer types by multicategory support vector machines using gene expression data", Bioinformatics 2003 19: 1132-1139.

[Mukherjee2003] Mukherjee, S., Pablo Tamayo, Simon Rogers, Ryan Rifkin, Anna Engle, Colin Campbell, Todd R Golub, Jill P Mesirov, "Estimating dataset size requirements for classifying DNA microarray data", Journal of Computational Biology, 10(2):119-142, 2003.

[Nutt2003] Nutt, C., et al. "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification", Cancer Res. 2003 Apr 1;63(7):1602-7.

[Pomeroy2002] Pomeroy, L., et al. "Prediction of central nervous system embryonal tumour outcome based on gene expression", Nature, vol 415, 24 January 2002.

[Ramaswamy2001] Ramaswamy, S., et al. "Multiclass cancer diagnosis using tumor gene expression signatures", Proc Natl Acad Sci U S A Dec 11, 2001.

[Singh2002] Singh, D., et al. "Gene expression correlates of clinical prostate cancer behavior", Cancer Cell : March 2002, Vol. 1

[Shipp2002] Shipp, M., et al. "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning", Nature Medicine, Volume 8, Number 1, January 2002.

[Staunton2001] Staunton, J., et al. "Chemosensitivity prediction by transcriptional profiling", PNAS, September 11, 2001, vol. 98, no. 19, 10787-10792.

[Su2001] Su, A.I., et al. "Molecular classification of human carcinomas by use of gene expression signatures", Cancer Res. 2001 Oct 15;61(20):7388-93.

[Venkataraman2002] Venkataraman, P. Applied Optimization with MATLAB Programming. John Wiley & Sons, Inc., 2002.