Diagnosis of Ovarian Cancer Based on Mass Spectra of Blood Samples*

Hong Tang
Computer Science and Eng.
University of South Florida
Tampa, FL 33620
htang2@csee.usf.edu

Yelena Mukomel
Computer Science and Eng.
University of South Florida
Tampa, FL 33620
mukomel@csee.usf.edu

Eugene Fink
Language Technologies
Carnegie Mellon University
Pittsburgh, PA 15213
e.fink@cs.cmu.edu

Abstract – *The early detection of cancer is crucial for successful treatment, and medical researchers have investigated a number of early-diagnosis techniques. Recently, they have discovered that some cancers affect the concentration of certain molecules in the blood, which allows early diagnosis by analyzing the blood mass spectrum. Researchers have developed several techniques for the analysis of the mass-spectrum curve, and used them for the detection of prostate, ovarian, breast, bladder, pancreatic, kidney, liver, and colon cancers.*

We have continued this work and applied data mining to the diagnosis of ovarian cancer. We have identified the most informative points of the mass-spectrum curve, and then used decision trees, support vector machines, and neural networks to determine the differences between the curves of cancer patients and healthy people.

Keywords: Data mining, medical application, decision trees, support vector machines, neural networks.

1 Introduction

The development of tools for the early cancer diagnosis is a major open problem, and clinicians have investigated a variety of diagnosis techniques. Recently, they have discovered that cancer may affect the blood mass spectrum, and studied diagnosis methods based on the analysis of mass-spectrum data, which provide information about proteins and their fragments [3, 4, 24]. The blood mass spectrum is a curve (Figure 1), where the x-axis shows the ratio of the weight of a specific molecule to its electric charge, and the y-axis is the signal intensity for the same molecule. The mass-spectrum analysis is a fast inexpensive procedure based on a sample of a patient's blood, and it may potentially allow cancer screening with little discomfort to a patient.

Medical researchers have developed several techniques for analyzing the mass-spectrum data, which allow the diagnosis of various cancers, including ovarian, breast, prostate, bladder, pancreatic, kidney, liver, and colon cancers. The effectiveness of these techniques varies across cancer types, methods for generating mass spectra, and algorithms for analyzing the resulting data. Clinicians use three standard measures of the effectiveness of diagnosis techniques: sensitivity, specificity, and accuracy. The *sensitivity* is the probability of the correct diagnosis for a patient with cancer, the *specificity* is the chances of the correct diagnosis for a healthy person, and the *accuracy* is the chances of the correct diagnosis for the overall population of healthy and sick people. The sensitivity of the mass-spectrum diagnosis techniques has varied from 64% to 99%, the specificity has been between 66% and 98%, and the overall accuracy has been between 73% and 98%.

Table 1: Data sets used in the experiments.

Data set	Number of cases	
	Cancer	Healthy
1	100	116
2	100	116
3	162	91

We have continued this work and investigated techniques for the diagnosis of early-stage ovarian cancer. Specifically, we have applied decision-tree learning, support vector machines, and neural networks to identify the differences between the mass spectra of ovarian-cancer patients and those of healthy people. We have used three data sets (Table 1), available at *http://clinicalproteomics.steem.com*. Sets 1 and 2 include the mass spectra of 100 cancer patients and 116 healthy people, whereas Set 3 includes the data of 162 cancer patients and 91 healthy people. Each mass-spectrum curve consists of 15,155 points.

The experiments have confirmed that the mass spectra allow the diagnosis of ovarian cancer. The sensitivity of the developed technique varies from 85% to 99%, depending on the data set, its specificity is between 81% and 99%, and its accuracy is between 82% and 99%.

2 Previous work

Medical researchers have developed techniques for the detection of early cancer based on *protein markers,* which are certain molecules in body tissues and fluids [15], but these techniques are often inaccurate.

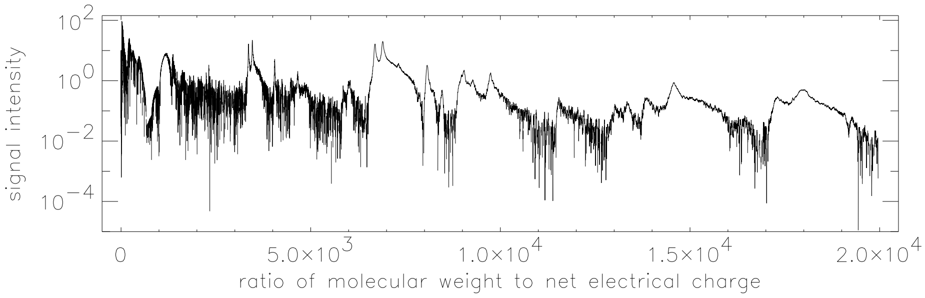

Figure 1: Example of a mass-spectrum curve.

For example, the specificity of an antigen method for the prostate-cancer detection is only 25–30%, although its sensitivity is high [2]; as another example, the sensitivity of a similar method for breast cancer is 23%, and its specificity is 69% [10]. Recently, researchers have developed a new cancer-detection method based on the application of data mining to the mass spectra of patients' tissue cells, blood, serum, and other body fluids [12, 14, 23].

Peaks in mass spectra. Some researchers have analyzed mass-spectrum curves using the Ciphergen System software, which helped to identify major peaks. Hlavaty *et al.* found that a 50.8k Dalton protein peak was present in all prostate-cancer samples, and absent in all samples of healthy people [9]. Watkins *et al.* used the same method to detect breast, colon, and prostate cancers [21]. They correctly identified 100% of breast cancer cases and ruled out 96% of noncancer cases. For colon cancer, they correctly identified 100% of cancer cases and ruled out 86% of noncancer cases. For prostate cancer, their results were 100% correct for both cancer and noncancer cases. Sauter *et al.* analyzed mass-spectrum curves of the nipple aspirate fluid over the 5–40k Dalton range, and identified five relevant peaks [19]. The most relevant peaks were 6.5k Dalton and 15.9k Dalton, and their use gave 84% sensitivity and 100% specificity.

Decision trees. Adam *et al.* applied decision-tree learning to the blood mass spectra of prostate-cancer patients [1]. They used the Ciphergen System software for peak detection, and decision trees for classification based on the intensity of nine highest peaks, which gave 96% accuracy, 83% sensitivity, and 97% specificity. They also experimented with biostatistical algorithms, genetic clustering, and support vector machines, which gave accuracy between 83% and 90%. Qu *et al.* applied a boosted decision tree method [17] using the same data and features as Adam. They developed two new classifiers, called AdaBoost and Boosted Decision Stump Feature Selection. For AdaBoost, the sensitivity was 98.5% with the 95% confidence interval of 96.5–99.7%, and the specificity was 97.9% with the 95% confidence interval of 95.5–99.4%. For Boosted Decision Stump Feature Selection, the sensitivity was 91.1% with the 95% confidence interval of 86.9–94.6%, and the specificity was 94.3% with the 95% confidence interval of 90.7–97.1%.

Neural networks. Ball *et al.* applied back-propagation neural networks to determine astroglial tumor grade (1 or 2), which gave 100% accuracy [5]. Poon *et al.* used neural networks to distinguish hepatocellular carcinoma from chronic liver disease, which gave 92% sensitivity and 90% specificity [16].

Clustering. Petricoin *et al.* combined a genetic algorithm with self-organizing cluster analysis for identifying ovarian cancer [11]. The sensitivity of their technique was 100% with the 95% confidence interval of 93–100%, and the specificity was 95% with the 95% confidence interval of 87–99%. They also applied their technique to diagnose prostate cancer [13], which gave 95% sensitivity with the 95% confidence interval of 82–99%, and 78% specificity with the 95% confidence interval of 72–83%. Poon *et al.* applied two-way hierarchical clustering to distinguish hepatocellular carcinoma from chronic liver disease [16]; however, they did not report its sensitivity, specificity, or accuracy.

Other methods. Valerio *et al.* applied the statistical χ^2 test to the mass spectra of thirteen pancreatic cancer patients, nine chronic pancreatitis patients, and ten healthy people, and found unique protein peaks for each of the three groups [20]; however, they did not report the sensitivity, specificity, or accuracy of their method. Cazares *et al.* analyzed mass spectra of prostate cancer [6]; they used the Ciphergen System software for peak detection, and logistic regression for classification, which gave 93% sensitivity and 94% specificity. Wu *et al.* compared several methods for classification of ovarian cancer, including linear discriminant analysis, quadratic discriminant analysis, nearest neighbors, bagging classification trees, boosting classification trees, support vector machines, and random forests [22]; they concluded that the random-forest classification was the most effective.

3 New results

We describe a technique for selecting relevant points of the mass-spectrum curve, and give results of detecting ovarian cancer based on the values of these points.

Feature selection. We view each point of a mass-spectrum curve as a feature, and the corresponding signal intensity as its value. To select relevant features, we calculate the mean intensity values for each point in the mass spectra of the cancer and non-cancer groups, μ_1 and μ_2, and the corresponding standard deviations, σ_1 and σ_2. The mean difference of these intensities is $|\mu_1 - \mu_2|$, and the standard deviation of this difference is $\sqrt{\sigma_1^2 + \sigma_2^2}$. For each point, we determine the ratio of the mean difference to its standard deviation, $|\mu_1 - \mu_2| / \sqrt{\sigma_1^2 + \sigma_2^2}$, and select a given number of points with the greatest ratios.

We impose a lower bound on the distance between selected points, which prevents the choice of points with correlated values. After selecting the point with the greatest ratio, we discard all points within the distance bound from it and choose the second greatest-ratio point among the remaining points. Then, we discard the points within the distance bound from the second selected point, choose the third greatest-ratio point among the remaining points, and so on.

Experiments. We have experimented with the use of decision trees, support vector machines, and neural networks for identifying cancer patients based on the selected points. We have used the C4.5 package (*www.cse.unsw.edu.au/∼quinlan*) for learning decision trees [18], the SVMFu package (*five-percent-nation.mit.edu/SvmFu*) for constructing support vector machines with linear kernel functions [7], and the Cascor 1.2 package (*www.cs.cmu.edu/afs/cs/project/connect/code/supported*) for generating neural networks using the cascade-correlation algorithm [8].

We have implemented an experimental setup that allows control over the number of features and minimal distance between selected features. We have varied the number of features from 1 to 64, and the minimal distance from 1 to 1024. For each combination of settings, we have used eighteen-fold cross-validation to evaluate the three learning techniques. In Figures 2–8, we show the dependency of the accuracy on the control variables. In Table 2, we give the minimal and maximal sensitivity, specificity, and accuracy for decision trees, support vector machines, and neural networks.

We have determined the number of features and minimal distance between features that lead to the highest accuracy (Table 3). The optimal number of features varies from four to thirty-two, depending on the learning technique and data set. We have also constructed the learning curves for the optimal choice of parameters (Figures 9–11); these curves show the dependency of the accuracy on the training-set size. The results show that all three techniques reach the maximal accuracy after processing about one hundred learning examples.

4 Concluding remarks

We have considered the problem of diagnosing ovarian cancer based on the blood mass-spectrum curve, and identified the relevant points of the curve. We have then applied decision trees, support vector machines, and neural networks to determine the values of these points that indicate ovarian cancer. The effectiveness of these techniques varies across the available data sets; the accuracy of decision trees is between 82% and 99%, the accuracy of support vector machines is between 83% and 99%, and the accuracy of neural networks is between 82% and 99%.

References

[1] Bao-Ling Adam, Yinsheng Qu, John W. Davis, Michael D. Ward, Mary Ann Clements, Lisa H. Cazares, Oliver John Semmes, Paul F. Schellhammer, Yutaka Yasui, Ziding Feng, and George L. Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, 62:3609–3614, 2002.

[2] Bao-Ling Adam, Antonia Vlahou, Oliver John Semmes, and George L. Wright. Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics*, 1(10):1264–1270, 2001.

[3] Ray Bakhtiar and Randall W. Nelson. Mass spectrometry of the proteome. *Molecular Pharmacology*, 60(3):405–415, 2001.

[4] Ray Bakhtiar and F. L. S. Tse. Biological mass spectrometry: A primer. *Mutagenesis*, 15(5):415–430, 2000.

[5] Graham Ball, Saira Mian, F. Holding, R. O. Allibone, J. Lowe, Selman Ali, Gui-Ru Li, S. McCardle, Ian O. Ellis, Colin Creaser, and Robert C. Rees. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*, 18(3):395–404, 2002.

[6] Lisa H. Cazares, Bao-Ling Adam, Michael D. Ward, Suhail Nasim, Paul F. Schellhammer, Oliver John Semmes, and George L. Wright. Normal, benign, preneoplastic, and malignant

Table 2: Effectiveness of ovarian-cancer diagnosis. We show the minimal (Min) and maximal (Max) accuracy, sensitivity, and specificity for decision trees, support vector machines, and neural networks.

		Decision trees		SVM		Neural nets	
		Min	Max	Min	Max	Min	Max
Data Set 1	Accuracy	75%	82%	76%	83%	67%	82%
	Sensitivity	68%	86%	75%	85%	66%	85%
	Specificity	72%	81%	74%	85%	67%	84%
Data Set 2	Accuracy	81%	94%	78%	94%	75%	96%
	Sensitivity	81%	95%	70%	98%	74%	94%
	Specificity	77%	96%	79%	94%	76%	97%
Data Set 3	Accuracy	96%	99%	88%	99%	94%	99%
	Sensitivity	96%	99%	85%	100%	94%	100%
	Specificity	91%	100%	95%	99%	92%	99%

Table 3: Control-variable values that lead to the maximal accuracy.

		Num. of features	Minimal distance	Accuracy	Sensitivity	Specificity
Data Set 1	Decision trees	4	1	82%	86%	78%
	SVM	32	16	83%	82%	84%
	Neural nets	32	256	82%	80%	84%
Data Set 2	Decision trees	8	4	94%	92%	96%
	SVM	4	2	94%	96%	93%
	Neural nets	32	1	96%	93%	98%
Data Set 3	Decision trees	8	64	99%	98%	100%
	SVM	16	8	99%	100%	99%
	Neural nets	16	2	99%	100%	99%

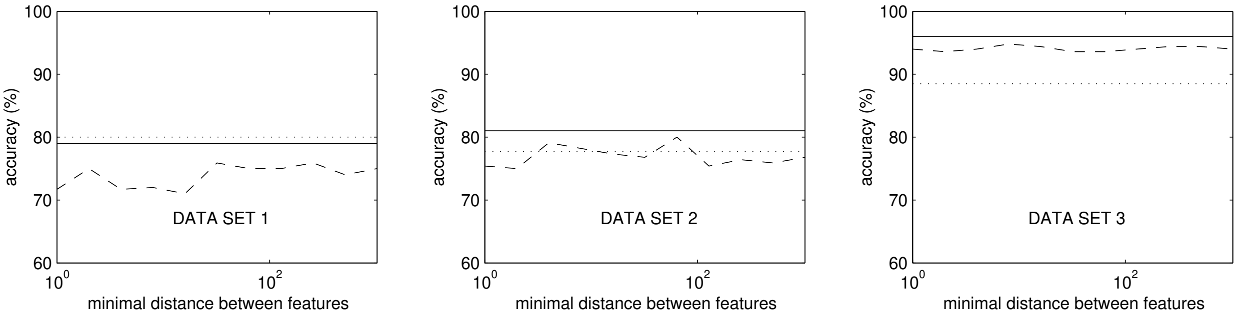

Figure 2: Experiments with *one feature.* We show the results for Data Set 1 (left), Data Set 2 (middle), and Data Set 3 (right). For each set, we plot the accuracy of decision trees (solid lines), support vector machines (dotted lines), and neural networks (dashed lines).

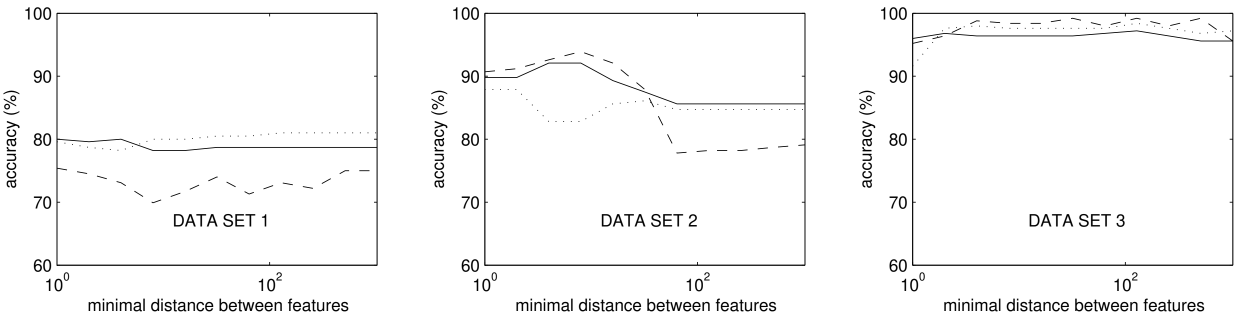

Figure 3: Experiments with *two features;* the legend is the same as in Figure 2.

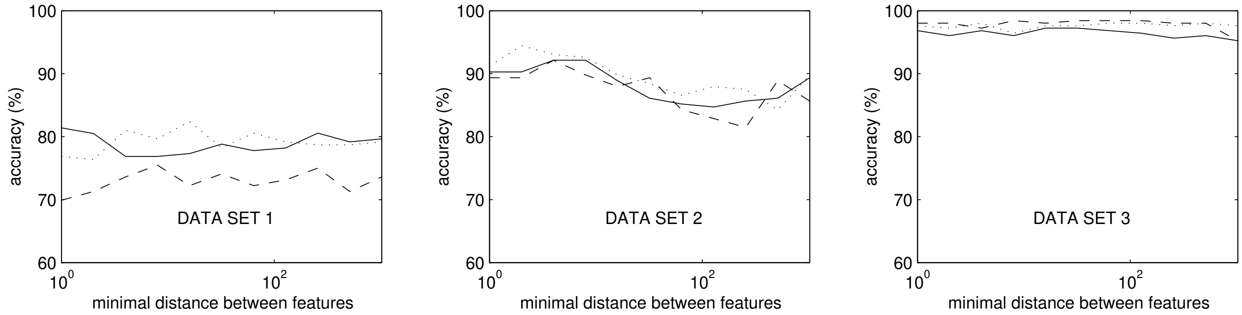

Figure 4: Experiments with *four features;* the legend is the same as in Figure 2.

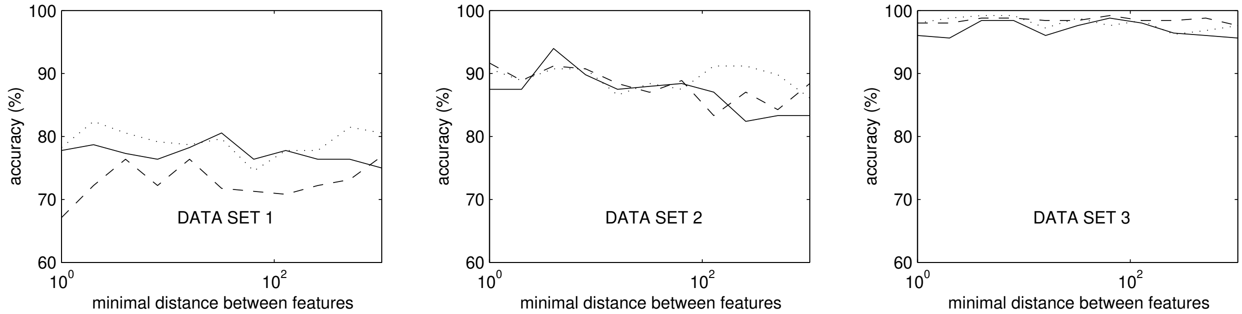

Figure 5: Experiments with *eight features;* the legend is the same as in Figure 2.

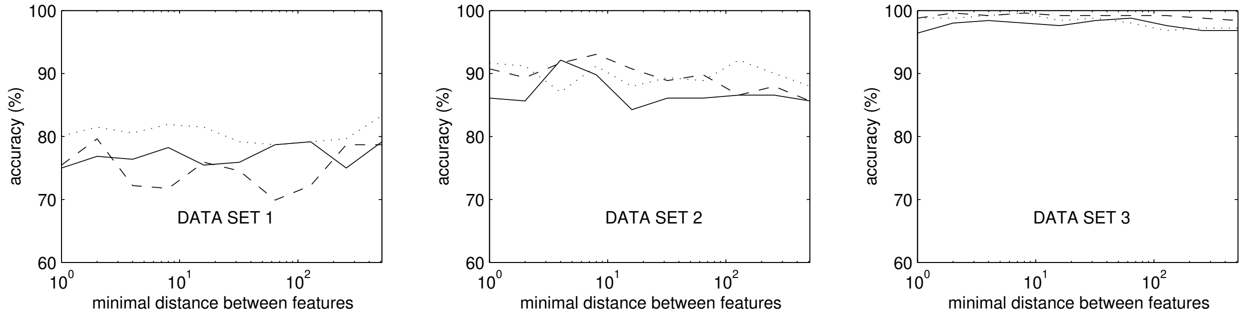

Figure 6: Experiments with *sixteen features;* the legend is the same as in Figure 2.

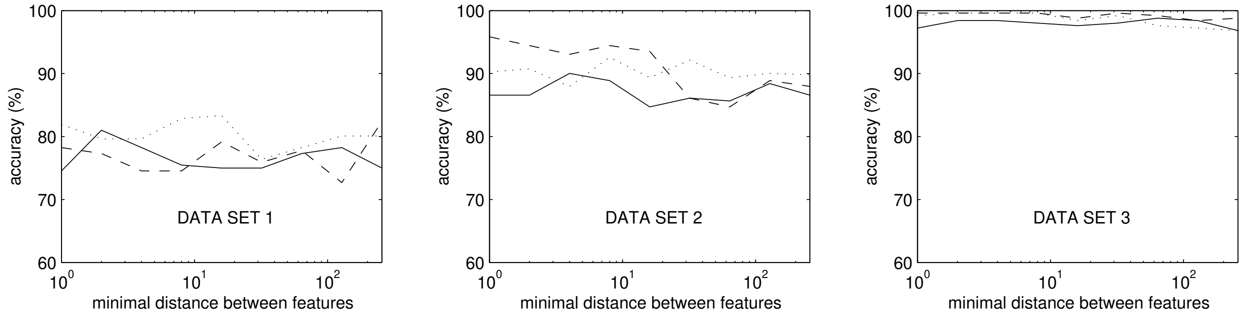

Figure 7: Experiments with *thirty-two features;* the legend is the same as in Figure 2.

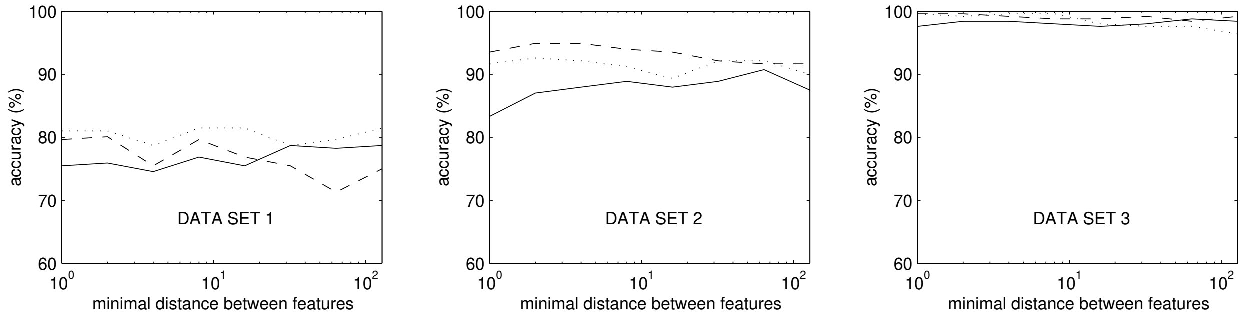

Figure 8: Experiments with *sixty-four features;* the legend is the same as in Figure 2.

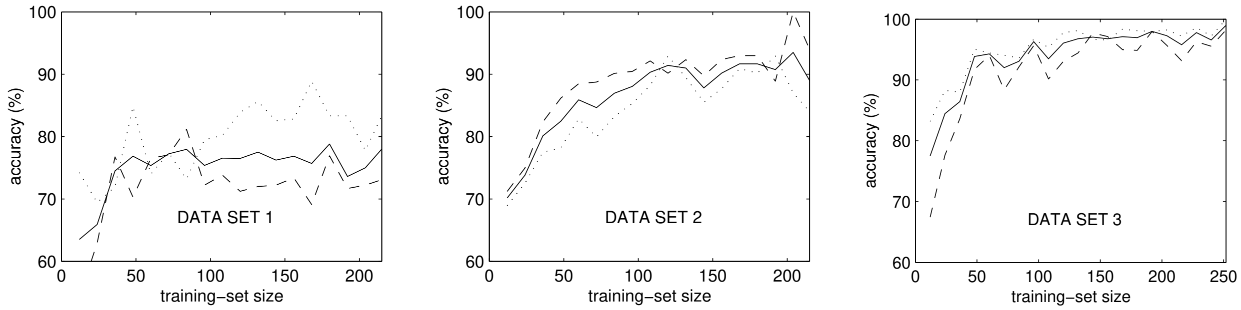

Figure 9: Learning curves for *decision trees.* We show the results of experiments with Data Set 1 (left), Data Set 2 (middle), and Data Set 3 (right). For each set, we plot the accuracy (solid lines), sensitivity (dotted lines), and specificity (dashed lines).

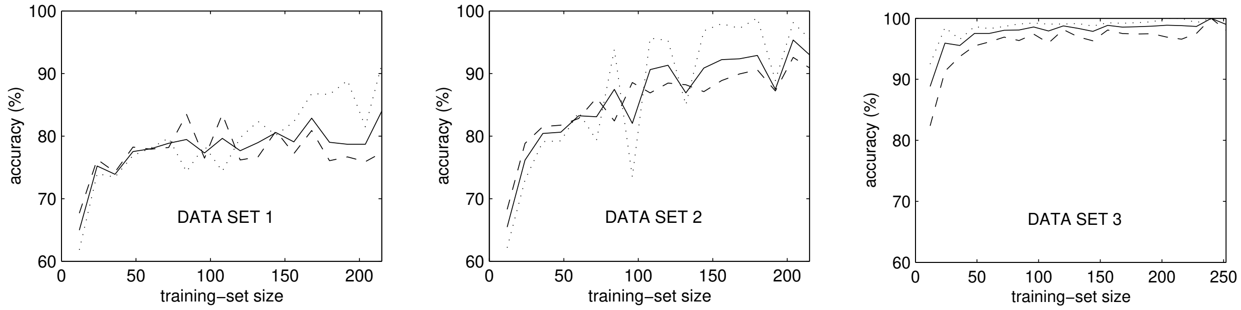

Figure 10: Learning curves for *support vector machines;* the legend is the same as in Figure 9.

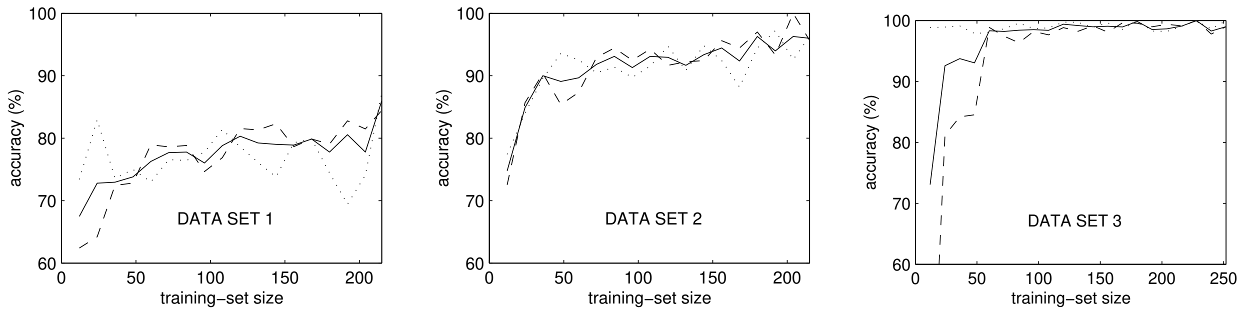

Figure 11: Learning curves for *neural networks;* the legend is the same as in Figure 9.

prostate cells have distinct protein expression profiles resolved by surface enhanced laser desorption/ionization mass spectrometry. *Clinical Cancer Research*, 8(8):2541–2552, 2002.

[7] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK, 2000.

[8] Scott E. Fahlman and Christian Lebiere. The cascade-correlation learning architecture. In David S. Touretzky, editor, *Advances in Neural Information Proceeding Systems 2*, pages 524–532. Morgan Kaufmann, Los Altos, CA, 1990.

[9] John J. Hlavaty, Alan W. Partin, Felicity Kusinitz, Matthew J. Shue, Adam Stieg, Kate Bennett, and Joseph V. Briggman. Mass spectroscopy as a discovery tool for identifying serum markers for prostate cancer. *Clinical Chemistry*, 47(10):1924–1926, 2001.

[10] Jinong Li, Zhen Zhang, Jason Rosenzweig, Young Y. Wang, and Daniel W. Chan. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry*, 48(8):1296–1304, 2002.

[11] Emanuel F. Petricoin, Ali M. Ardekani, Ben A. Hitt, Peter J. Levine, Vincent A. Fusaro, Seth M. Steinberg, Gordon B. Mills, Charles Simone, David A. Fishman, Elise C. Kohn, and Lance A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359:572–577, 2002.

[12] Emanuel F. Petricoin and Lance A. Liotta. Proteomic analysis at the bedside: Early detection of cancer. *Trends in Biotechnology*, 20(12)Suppl.:S30–S34, 2002.

[13] Emanuel F. Petricoin, David K. Ornstein, Cloud P. Paweletz, Ali Ardekani, Paul S. Hackett, Ben A. Hitt, Alfredo Velassco, Christian Trucco, Laura Wiegand, Kamillah Wood, Charles B. Simone, Peter J. Levine, W. Marston Linehan, Michael R. Emmert-Buck, Seth M. Steinberg, Elise C. Kohn, and Lance A. Liotta. Serum proteomic patterns for detection of prostate cancer. *Journal of the National Cancer Institute*, 94(20):1576–1578, 2002.

[14] Emanuel F. Petricoin, Kathryn C. Zoon, Elise C. Kohn, J. Carl Barrett, and Lance A. Liotta. Clinical proteomics: Translating benchside promise into bedside reality. *Nature Reviews Drug Discovery*, 1:683–695, 2002.

[15] Terence C. W. Poon and Philip J. Johnson. Proteome analysis and its impact on the discovery of serological tumor markers. *Clinica Chimica Acta*, 313:231–239, 2001.

[16] Terence C. W. Poon, Tai-Tung Yip, Anthony T. C. Chan, Christine Yip, Victor Yip, Tony S. Mok, Conrad C. Y. Lee, Thomas W. T. Leung, Stephen K. W. Ho, and Philip J. Johnson. Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. *Clinical Chemistry*, 49(5):752–760, 2003.

[17] Yinsheng Qu, Bao-Ling Adam, Yutaka Yasui, Michael D. Ward, Lisa H. Cazares, Paul F. Schellhammer, Ziding Feng, Oliver John Semmes, and George L. Wright. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry*, 48(10):1835–1843, 2002.

[18] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[19] Edward R. Sauter, Weizhu Zhu, X. J. Fan, R. P. Wassell, Inna Chervoneva, and Garrett C. Du Bois. Proteomic analysis of nipple aspirate fluid to detect biologic markers of breast cancer. *British Journal of Cancer*, 86:1440–1443, 2002.

[20] A. Valerio, Daniela Basso, S. Mazza, G. Baldo, A. Tiengo, S. Pedrazzoli, R. Seraglia, and M. Plebani. Serum protein profiles of patients with pancreatic cancer and chronic pancreatitis: Searching for a diagnostic protein pattern. *Rapid Communications in Mass Spectrometry*, 15(24):2420–2425, 2001.

[21] Brynmor Watkins, Robert Szaro, Shannon Ball, Tatyana Knubovets, Joseph Briggman, John J. Hlavaty, Felicity Kusinitz, Adam Stieg, and Ying-Jye Wu. Detection of early-stage cancer by serum protein analysis. *American Laboratory*, 33:32–36, 2001.

[22] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003.

[23] Julia D. Wulfkuhle, Lance A. Liotta, and Emanuel F. Petricoin. Proteomic applications for the early detection of cancer. *Nature Reviews Cancer*, 3:267–275, 2003.

[24] John R. Yates. Mass spectrometry: From genomics to proteomics. *Trends in Genetics*, 16(1):5–8, 2000.