2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Wavelet-Based Feature Extraction for Microarray Data Classification

Shutao Li, Chen Liao, James T. Kwok

Abstract— **Microarray data typically have thousands of genes, and thus feature extraction is a critical problem for accurate cancer classification. In this paper, a feature extraction method based on the discrete wavelet transform (DWT) is proposed. The approximation coefficients of DWT, together with some useful features from the high-frequency coefficients selected by the maximum modulus method, are used as features. The combined coefficients are then forwarded to a SVM classifier. Experiments are performed on two standard benchmark data sets: ALL/AML Leukemia and Colon tumor. Experimental results show that the proposed method can achieve state-of-the-art performance on cancer classification.**

I. Introduction

Cancer is usually caused by abnormal cells that grow and spread unconventionally. Because of its terrible influence to humans, it has become one of the top life threats. In recent years, the study of DNA microarray has become a very important technology for cancer classification. In a microarray experiment, the genes are monitored many times under different conditions and for different tissue types. However, since the data usually include a few samples but thousands, or even tens of thousands, of genes, it poses significant difficulty to traditional classifiers. Hence, gene selection is an important issue in cancer classification with microarray data. A good gene selection method should eliminate irrelevant, redundant, or noisy genes for classification, while at the same time keeps highly discriminative genes [1].

In general, there are three approaches to gene (feature) extraction: filter, wrapper and embedded approaches. In the filter approach, genes are selected according to the intrinsic characteristics. The filter approach works as a preprocessing step without the incorporation of any learning algorithm. Examples include the nearest shrunken centroid method, TNoM-score based method, and the T-statistics method [2]. In the wrapper approach, a learning algorithm is used to score the feature subsets based on the resultant predictive power, and an optimal feature subset is searched for a specific classifier [3]. Examples include Recursive Feature Elimination, and genetic algorithm-based algorithms.

On the other hand, multi-resolution wavelet transform can process both stationary and non-stationary signals and has good multi-resolution capabilities. Because of these advantages, it has been effectively used in many bioinformatics applications such as DNA sequence analysis [4], [5] and genomic data analysis [6], [7]. In [8], the authors proposed a gene selection method using mutual information and wavelet transforms. They use the wavelet transform to represent the microarray data. Each feature (wavelet coefficient) of the wavelet transform is related to several genes of the original gene microarray data. Then, a mutual information-based feature selection method is adopted to select the strongest features from among the wavelet coefficients. In [9], the authors pointed out that wavelet transforms may also be useful in the realm of gene expression. The expression signal given by the genes in clustered order can be wavelet transformed, which then compresses the signal from many genes into a few components, possibly aiding in the development of new tumor classifiers.

In this paper, a new gene selection method using the discrete wavelet transform (DWT) is proposed for cancer classification using microarray data. The rest of this paper is organized as follows. Section II briefly introduces the one-dimensional DWT. In Section III, the DWT-based feature selection method is proposed. Experimental results are presented in Section IV, and the last section gives some concluding remarks.

II. 1D-DWT

The wavelet transform [10], [11] represents any arbitrary function as a superposition of wavelets, which are functions generated from a mother wavelet by dilations and translations. It has been used as a significant mathematical tool for decomposing a function in terms of its time and frequency components. It outperforms the classical Fourier transform on the condition that the localization must be both in time and the frequency domain for non-stationary signals. The DWT is efficiently computed using Mallat's pyramid algorithm.

One property of the DWT is the decorrelating property of the wavelet coefficients. Given a function $\{y(t) : t = 0, 1, \ldots, N-1\}$ (with N being a power of 2), the first level of coefficients from the one-dimensional DWT of $y(t)$, for $t = 0, 1, \ldots, \frac{N}{2} - 1$, can be expressed as:

$$s_{t,1} = \sum_{m=0}^{M} l_m y(m + 2t) \Rightarrow s_1 = Ly(t), \quad (1)$$

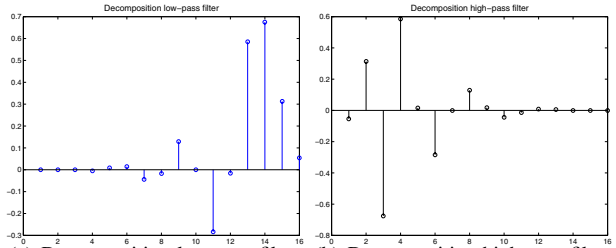
$$d_{t,1} = \sum_{m=0}^{M} h_m y(2t + 1 - m) \Rightarrow d_1 = Hy(t), \quad (2)$$

where $s_{t,1}$'s are elements of the scaling coefficients S_1, and $d_{t,1}$'s are elements of the detail coefficients at the first level d_1. The operators L and H are low-pass and high-pass decomposition filter representations derived from l_m and h_m when these two coefficient vectors are combined using the

Shutao Li and Chen Liao are with the College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China. James T. Kwok is with the Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

0-7803-9490-9/06/$20.00/©2006 IEEE

down-sampling and convolution. The s_1 includes the low-frequency part and d_1 includes the high-frequency part of $y(t)$. The low-frequency part can be decomposed further in the same way. Figure 1 shows the four filters associated with a particular wavelet basis, the Daubechies basis.



(a) Decomposition low-pass filter. (b) Decomposition high-pass filter.
(c) Reconstruction low-pass filter. (d) Reconstruction high-pass filter.

Fig. 1. The four filters associated with the Daubechies wavelet basis. Note that the Daubechies basis is an example of the orthogonal wavelet basis, which has the property that the high-pass and low-pass filters are alternated flip of each other.

The whole process of obtaining the wavelet transform of $y(t)$ using the pyramid algorithm is shown in Figure 2. The data $y(t)$ are defined to be the zeroth scale coefficients, i.e., $S_0 = y(t)$. Then S_0 is decomposed into two subsequences: s_1 and d_1 by (1) and (2), the length of each being $N/2$. The same operations are then repeated to the vector s_1. As shown in Figure 2, s_2 and d_2 are obtained, the length of each is $N/4$. This process is iterated J times to obtain the wavelet detail and scaling coefficients

$$\begin{aligned} \omega &= (s_J, d_J, d_{J-1}, \ldots, d_1) \\ &= (L^J y, HL^J y, HL^{J-1} y, \ldots, Hy). \end{aligned}$$

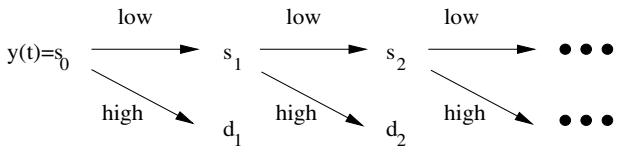


Fig. 2. The 1D wavelet decomposition process.

III. Feature Extraction

In this Section, we describe the proposed DWT-based feature extraction method. We denote the number of samples in the gene expression data set by N. Each sample contains expression values of M genes (features). As is common in gene expression analysis, M is usually much larger than N.

The idea behind the feature extraction process is that the approximation coefficients usually contain the most important information, and hence they will constitute part of the extracted features. However, obviously, there are also some important information in the high-frequency part. Hence, the maximum modulus method is used to select some high-frequency coefficients.

The proposed feature extraction method consists of the following steps:

1) For each sample, the N gene expression data is decomposed using the 1D-DWT. Let the number of decomposition levels be J. The decomposed coefficients can be expressed as $\omega = (s_J, d_J, d_{J-1}, \ldots, d_1)$. An example of the approximation coefficients and high-frequency coefficients are shown in Figures 3 and 4.
2) The approximation coefficients in s_J are included as part of the final feature set.
3) We now use the maximum modulus method to select some high-frequency coefficients.
   a) First, the high frequency coefficients are thresholded.
   $$\text{thd}_{i,j} = \begin{cases} 0 & |d_{i,j}| < \text{TH}, \\ 1 & \text{otherwise}, \end{cases}$$
   for $i = 1, \ldots, N, j = 1, \ldots, L$, where N is the number of samples, L is the length of the high-frequency part. An example of the thresholding operation is shown in Figure 5.
   b) Then the weights of each coefficient are obtained:
   $$\text{weight}_j = \sum_{i=1}^{N} \text{thd}_{i,j}.$$
   c) Finally, the weight_j values are ranked, and locations corresponding to the top n maximum values are found. The coefficients of these locations are used as the high frequency features.
4) The approximation part obtained in step 2 and the high-frequency part obtained in step 3 are combined together to form a new gene subset, which thus has a much lower dimensionality than the original one.

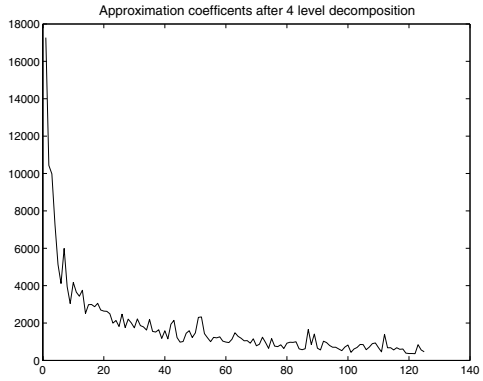


Fig. 3. An example of the approximation coefficients.

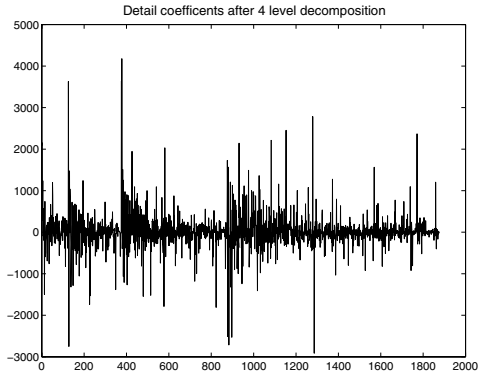


Fig. 4. An example of the high-frequency coefficients.

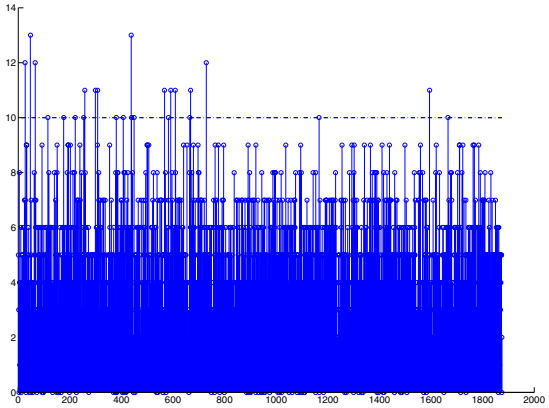


Fig. 5. An example of the thresholding of the high-frequency coefficients.

IV. Experiments

A. Setup

In this Section, experiments are performed on the two benchmark data sets:

1) ALL/AML Leukemia data set: It contains a total of 72 samples of two types of leukemia: 47 of acute myeloid leukemia (AML) and 25 of acute lymphoblastic leukemia (AML). Each sample contains expression values of 7,129 genes (features). This data set is first studied in [12].
2) Colon cancer data set: It consists of a total of 62 samples, 22 of them are from normal colon tissues while the remaining 40 are from tumor tissues. Each sample contains expression values of 2,000 genes. This data set is first studied in [13].

Both data sets can be downloaded from http://www.tsi.enst.fr/∼gfort/GLM/Programs.html.

After feature extraction, we use the support vector machines (SVM) [14] for classification. The SVM has been successfully used in various applications, such as computer vision [15] and bioinformatics [16]. There are a number of kernel functions that can be used with the SVM. In most of our experiments, the default kernel used is the Gaussian kernel:

$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right), \tag{3}$$

where σ^2 is a user-defined variance parameter. The effect of using other kernel functions will be studied in Section IV-D. The SVM package, downloaded from http://svm.sourceforge.net/docs/3.00/api/, is used here.

To be in line with the reported works (e.g., [17]) on these two benchmark data sets, the classification accuracies will be computed using the leave-one-out cross-validation procedure.

B. Use of Different Wavelet Basis and Decomposition Levels

We first compare the performance of different wavelet bases. In general, there are two main types of bases: orthogonal and biorthogonal. Common orthogonal bases include the Daubechies (db), Coiflets (coif), Symlets (sym), and discrete Meyer (dmey). The Daubechies wavelets are the most popular non-redundant orthogonal wavelet basis. The Symlets orthogonal wavelet basis is a quasi-symmetric extension of the Daubechies wavelets. The Coiflets orthogonal wavelet basis is another extension, with vanishing moment conditions for both the wavelets and scaling functions. It is also more symmetrical than the classical Daubechies. The biorthogonal (bior) wavelets are orthogonal in a more general sense. They are sometimes more desirable than orthogonal ones because they can preserve linear phase, have finite impulse response, and the mother wavelets can have arbitrarily high regularity [18]. Implementations based on the Matlab Wavelet Toolbox will be used here.

These wavelet filters also come with different lengths. In the sequel, the length of the wavelet filter is indicated after the name of the basis. For example, "db1" stands for the Daubechies wavelet filter with length 1. Note also that the low pass and high pass filters for biorthogonal wavelets do not have the same length, and their lengths are separated by a dot. For example, "bior2.6" stands for the biorthogonal wavelet filter where the low pass filter has length 2, while the high pass filter has length 6.

In this experiment, we compare the performance of different filters with different lengths, using 3 or 4 levels of decomposition. Preliminary studies suggest that the use of more decomposition levels does not improve classification results. Because of the large number of possible combinations, we only show the results for (TH=70, $n = 100$) on the ALL/AML data; and (TH=10, $n = 250$) on the Colon data. The effect of using different (TH,n) combinations will be studied in Section IV-C. The σ parameter in the Gaussian kernel is fixed at $\sigma = 2^{-8}$. Results are shown in Table I.

C. Use of Different Parameters

In this experiment, we study the effect of the parameters TH and n. Again, because of the large number of possible combinations, we only show the results for the use of the biorthogonal wavelet filter (bior2.6) with 3 levels of decomposition on the ALL/AML data; and the Daubechies wavelet filter (db8) with 4 levels of decomposition on the

TABLE I

RESULTS ON USING DIFFERENT WAVELET BASIS AND DECOMPOSITION LEVELS.

date set	wavelet filter		#levels	accuracy (%)
ALL/AML	Daubechies	db1	3	93.06
			4	93.06
		db8	3	95.83
			4	94.44
	Coiflets	coif1	3	95.83
			4	95.83
		coif3	3	95.83
			4	94.44
	Symlets	sym2	3	91.67
			4	91.67
		sym15	3	94.44
			4	93.06
	discrete Meyer	dmey	3	95.83
			4	95.83
	biorthogonal	bior1.1	3	93.06
			4	93.06
		bior2.6	3	**100.00**
			4	97.22
Colon tumor	Daubechies	db1	3	85.48
			4	79.03
		db8	3	82.26
			4	**93.55**
	Coiflets	coif1	3	85.48
			4	83.87
		coif3	3	83.87
			4	85.48
	Symlets	sym2	3	85.48
			4	79.03
		sym15	3	85.48
			4	87.10
	discrete Meyer	dmey	3	83.87
			4	80.65
	biorthogonal	bior1.1	3	85.48
			4	79.03
		bior2.6	3	87.10
			4	80.65

Colon data. The σ parameter in the Gaussian kernel is fixed at $\sigma = 2^{-8}$.

Results are shown in Table II. As can be seen, the use of a larger threshold (TH) seems to be beneficial. However, the extraction of a large number of features (n) leads to inferior performance in most cases. This agree with our intuition that feature selection is mandatory for these small training sets.

D. Use of Different Kernels in the SVM

In this experiment, we also show the performance with two other popularly-used kernel functions:

1) Linear kernel: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$;
2) Polynomial kernel: $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y})^d$, where d is the polynomial degree.

Results are shown in Tables III and IV respectively. As a summary, the highest accuracies attained by the three kernel functions are compared in Table V. As can be seen, all three kernels attain the best possible accuracy of 100% on the Leukemia data, and they again attain the same performance of 93.55% on the Colon data set. This insensitivity to the choice of kernels shows that the features extracted by the proposed method contain all the information useful for classification.

TABLE II

RESULTS ON USING DIFFERENT TH AND n PARAMETERS.

data set	TH	n	accuracy (%)
ALL/AML	36	100	97.72
		200	95.83
	45	100	97.72
		200	95.83
	70	100	**100.00**
		200	98.61
Colon tumor	5	200	83.87
		250	83.87
	8	200	87.10
		250	83.87
	10	200	91.94
		250	**93.55**

E. Comparison with the Other Methods

Table VI shows the performance of the other methods on the two data sets as reported in the literature. All these methods use leave-one-out cross-validation and so their classification accuracies can be directly compared. As can be seen, the proposed method, together with JCFO (*Joint Classifier and Feature Optimization*) [17] with the linear kernel, attain the best classification accuracy (of 100%) on the ALL/AML data set. On the colon data set, the proposed method also outperforms all other methods except for the JCFO with linear kernel. However, it should be noted that the proposed method is based on wavelet transforms and so the computation is very fast. On the other hand, JCFO relies on the Expectation-Maximization (EM) algorithm [19] and is much slower.

V. CONCLUSIONS

In this paper, we proposed a wavelet-based feature extraction method for microarray data. The approximation coefficients obtained from the discrete wavelet transform (DWT), together with useful features from the high frequency coefficients as selected by the maximum modulus method, are used as input features to a SVM classifier. Experimental results on standard benchmark data sets show that the proposed method outperforms the other methods in terms of classification accuracy.

In general, wavelet transforms have been increasingly popular in the field of bioinformatics because of its capability in multi-resolution analysis and spatial-frequency localization. In the future, we will investigate the combination of wavelet decompositions with other feature extraction methods. A systematic method for the setting of the parameters will be investigated in the future. Moreover, experiments on some other bioinformatics data sets will also be performed.

ACKNOWLEDGMENT

This paper is supported by the Research Grants Council of the Hong Kong Special Administrative Region under grant 615005, the National Nature Science Foundation of China (No. 6040204) and the Program for New Century Excellent Talents in University.

TABLE VI

CLASSIFICATION ACCURACIES (%) OBTAINED BY THE VARIOUS METHODS AS REPORTED IN THE LITERATURE.

methods	ALL/AML	Colon tumor
Adaboost (decision stumps) [20]	95.8	72.6
SVM (quadratic kernel) [20]	94.4	74.2
SVM (linear kernel) [20]	93.0	77.4
RVM (linear kernel) [17]	94.4	80.6
RVM (no kernel: on feature space) [17]	97.2	88.7
logistic regression (no kernel: on feature space) [17]	97.2	71.0
sparse probit regression (quadratic kernel) [17]	95.8	84.6
sparse probit regression (linear kernel) [17]	97.2	91.9
sparse probit regression (no kernel: on feature space) [17]	97.2	85.5
JCFO (quadratic kernel) [17]	98.6	88.7
JCFO (linear kernel) [17]	**100.0**	**96.8**
proposed method	**100.0**	93.6

REFERENCES

[1] Y. Tang, Y.-Q. Zhang, and Z. Huang, "FCM-SVM-RFE gene feature selection algorithm for leukemia classification from microarray gene expression data," in *Proceedings of IEEE International Conference on Fuzzy Systems*, 2005, pp. 97–101.

[2] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Informatics*, vol. 13, pp. 51–60, 2002.

[3] K. Duan and J. Rajapakse, "A variant of SVM-RFE for gene selection in cancer classification with expression data," in *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2004, pp. 49–55.

[4] N. Kawagashira, Y. Ohtomo, K. Murakami, K. Matsubara, J. Kawai, P. Carninci, Y. Hayashizaki, and S. Kikuchi, "Wavelet profiles: Their application in Oryza sativa DNA sequence analysis," in *Proceedings of IEEE Computer Society Bioinformatics Conference*, Stanford, CA, USA, Aug. 2002, pp. 345–346.

[5] J. Ning, C. Moore, and J. Nelson, "Preliminary wavelet analysis of genomic sequences," in *Proceedings of IEEE Computer Society Bioinformatics Conference*, Stanford, CA, USA, Aug. 2003, pp. 509–510.

[6] J. Song, T. Ware, S.-L. Liu, and M. Surette, "Comparative genomics via wavelet analysis for closely related bacteria," *EURASIP Journal on Applied Signal Processing*, vol. 1, pp. 5–12, 2004.

[7] X.-Y. Zhang, F. Chen, Y.-T. Zhang, S. Agner, M. Akay, Z.-H. Lu, M. Waye, and S.-W. Tsui, "Signal processing techniques in genomic engineering," *Proceedings of the IEEE*, vol. 90, no. 12, pp. 1822–1833, 2002.

[8] X. Zhou, X. Wang, and E. Doughert, "Nonlinear-probit gene classification using mutual-information and wavelet based feature selection," *Biological Systems*, vol. 12, no. 3, pp. 371–386, 2004.

[9] T. Tokuyasu, D. Albertson, D. Pinkel, and A. Jain, "Wavelet transforms for the analysis of microarray experiments," in *Proceedings of IEEE Computer Society Bioinformatics Conference*, 2003, pp. 429–430.

[10] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, July 1989.

[11] ——, *A Wavelet Tour of Signal Processing*, 2nd ed. San Diego: Academic Press, 1999.

[12] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[13] 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," in *Proceedings of the National Academy of Science*, vol. 96, June 1999, pp. 6745–6750.

[14] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[15] S. Li, J. Kwok, H. Zhu, and Y. Wang, "Texture classification using support vector machines," *Pattern Recognition*, vol. 36, no. 12, pp. 2883–2893, Dec. 2003.

[16] J.-P. Vert, K. Tsuda, and B. Schölkopf, "A primer on kernel methods," in *Kernel Methods in Computational Biology*, B. Schölkopf, K. Tsuda, and J.-P. Vert, Eds. MIT, 2004, pp. 35–70.

[17] B. Krishnapuram, L. Carin, and A. Hartemink, "Gene expression analysis: Joint feature selection and classifier design," in *Kernel Methods in Computational Biology*, B. Schölkopf, K. Tsuda, and J.-P. Vert, Eds. MIT, 2004, pp. 299–318.

[18] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Transactions on Image Processing*, vol. 1, no. 2, pp. 205–220, 1992.

[19] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[20] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," in *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 2000, pp. 54–64.

TABLE III

PERFORMANCE OF THE LINEAR KERNEL.

data set	#levels	wavelet filter		accuracy (%)
ALL/AML (TH=70, $n = 100$)	3	Daubechies	db1	97.22
			db8	94.44
		Coiflets	coif1	95.83
			coif3	97.22
		Symlets	sym2	91.67
			sym15	94.44
		discrete Meyer	dmey	94.44
		biorthogonal	bior1.1	93.06
			bior2.6	**100.00**
	4	Daubechies	db1	93.06
			db8	94.44
		Coiflets	coif1	95.83
			coif3	94.44
		Symlets	sym2	91.67
			sym15	93.06
		discrete Meyer	dmey	95.83
		biorthogonal	bior1.1	93.06
			bior2.6	97.22
Colon tumor (TH=10, n=250)	3	Daubechies	db1	85.48
			db8	80.65
		Coiflets	coif1	83.87
			coif3	82.26
		Symlets	sym2	82.26
			sym15	85.48
		discrete Meyer	dmey	83.87
		biorthogonal	bior1.1	85.48
			bior2.6	88.71
	4	Daubechies	db1	85.48
			db8	**93.55**
		Coiflets	coif1	87.10
			coif3	85.48
		Symlets	sym2	85.48
			sym15	88.71
		discrete Meyer	dmey	85.48
		biorthogonal	bior1.1	85.48
			bior2.6	83.87

TABLE IV

PERFORMANCE OF THE POLYNOMIAL KERNEL (DEGREE=2).

data set	#levels	wavelet filter		accuracy (%)
ALL/AML (TH=70, $n = 100$)	3	Daubechies	db1	93.06
			db8	95.83
		Coiflets	coif1	97.22
			coif3	95.83
		Symlets	sym2	90.28
			sym15	95.83
		discrete Meyer	dmey	95.83
		biorthogonal	bior1.1	93.06
			bior2.6	**100.00**
	4	Daubechies	db1	93.06
			db8	95.83
		Coiflets	coif1	95.83
			coif3	94.44
		Symlets	sym2	91.67
			sym15	93.06
		discrete Meyer	dmey	95.83
		biorthogonal	bior1.1	93.06
			bior2.6	95.83
Colon tumor (TH=10, $n = 250$)	3	Daubechies	db1	80.65
			db8	82.26
		Coiflets	coif1	85.48
			coif3	85.48
		Symlets	sym2	82.26
			sym15	83.87
		discrete Meyer	dmey	85.48
		biorthogonal	bior1.1	80.65
			bior2.6	87.01
	4	Daubechies	db1	77.42
			db8	**93.55**
		Coiflets	coif1	80.65
			coif3	80.65
		Symlets	sym2	79.03
			sym15	88.71
		discrete Meyer	dmey	80.65
		biorthogonal	bior1.1	77.42
			bior2.6	79.03

TABLE V

HIGHEST ACCURACIES ATTAINED BY THE DIFFERENT KERNELS.

data set	kernel	highest accuracy (%)
ALL/AML	linear	100.00
	polynomial	100.00
	Gaussian	100.00
Colon tumor	linear	93.55
	polynomial	93.55
	Gaussian	93.55