Osong Public Health Res Perspect 2014 5(6), 324–332
http://dx.doi.org/10.1016/j.phrp.2014.09.002
pISSN 2210-9099 eISSN 2233-6052

ORIGINAL ARTICLE

Predicting 5-Year Survival Status of Patients with Breast Cancer based on Supervised Wavelet Method

Maryam Farhadian [a], Hossein Mahjub [b,*], Jalal Poorolajal [c], Abbas Moghimbeigi [d], Muharram Mansoorizadeh [e]

[a]*Department of Epidemiology and Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran.*
[b]*Research Center for Health Sciences and Department of Epidemiology and Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran.*
[c]*Modeling of Noncommunicable Diseases Research Center, Department of Epidemiology and Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran.*
[d]*Modeling of Noncommunicable Disease Research Center, Department of Biostatistics and Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran.*
[e]*Department of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamadan, Iran.*

Received: August 4, 2014
Revised: September 15, 2014
Accepted: September 22, 2014

KEYWORDS:
breast cancer,
microarray data,
supervised wavelet,
support vector machine

Abstract
Objectives: Classification of breast cancer patients into different risk classes is very important in clinical applications. It is estimated that the advent of high-dimensional gene expression data could improve patient classification. In this study, a new method for transforming the high-dimensional gene expression data in a low-dimensional space based on wavelet transform (WT) is presented.
Methods: The proposed method was applied to three publicly available microarray data sets. After dimensionality reduction using supervised wavelet, a predictive support vector machine (SVM) model was built upon the reduced dimensional space. In addition, the proposed method was compared with the supervised principal component analysis (PCA).
Results: The performance of supervised wavelet and supervised PCA based on selected genes were better than the signature genes identified in the other studies. Furthermore, the supervised wavelet method generally performed better than the supervised PCA for predicting the 5-year survival status of patients with breast cancer based on microarray data. In addition, the proposed method had a relatively acceptable performance compared with the other studies.
Conclusion: The results suggest the possibility of developing a new tool using wavelets for the dimension reduction of microarray data sets in the classification framework.

*Corresponding author.
E-mail: mahjub@umsha.ac.ir

1. Introduction

Metastatic breast cancer is a stage of breast cancer where the disease has spread to distant organs or tissues. Treatments against metastasis exist, but usually further treatments after surgery can have serious side effects and involve high medical costs [1]. An important task to optimize the adjuvant chemotherapy of metastasis related to breast cancer is to diagnose the risk of metastasis accurately [2–4].

Classification of cancer patients into different risk classes is very important in clinical applications. Traditional methods for patient classification were mainly based on a series of clinical and histological features [3]. It is estimated that the advent of high-dimensional gene expression data could improve patient classification [5]. Gene expression profiles of breast tumor samples could be used to predict relapse and metastatic patterns in breast cancer patients that could be potential candidate targets for new treatments [4]. It is reasonable to assume that any difference between the two tumors should be represented by some difference in gene expression. However, in microarray studies, the number of samples is relatively small compared to the number of genes per sample. Furthermore, from the biological aspect, only a small portion of genes have predicted the power for phenotypes. If all or most of the genes are considered in the predictive model, they can induce substantial noise and thereby lead to poor predictive performance [6]. Thus, in order to obtain good classification accuracy, a crucial step towards the application of microarray data is the dimensional reduction from the gene expression profiles. In recent years, both feature selection and feature extraction methods have been widely used for classifying gene expression data [7]. Bair and Tibshirani [8] and Bair et al. [9] explored the use of supervised principal component analysis (PCA), which is similar to conventional PCA except that it uses a subset of the predictors selected based on their association with the outcome. Wavelet-based methods have also been used to solve the dimension reduction problem. The primary intuition for applying wavelets in the case of gene expression is that genes are often coexpressed in groups. Therefore, it would be useful to treat the group as a single variable, akin to the motivation behind methods such as PCA [10]. One-dimensional discrete wavelet transform (DWT) is frequently used for feature extraction in the analysis of high-dimensional biomedical data [11]. Studies showed that this method has an acceptable performance in the field of feature extraction in the classification framework [11–15].

The current study aimed to introduce a dimension reduction strategy for transforming the high-dimensional gene expression data in a low-dimensional space based on wavelet transform (WT) in order to predict metastasis of breast cancer. Accordingly, a predictive support vector machine (SVM) model was built upon the reduced dimensional space. Then, the proposed novel supervised wavelet method of feature extraction was compared with the supervised PCA.

2. Materials and methods

The proposed method was applied to three publicly available microarray data sets related to breast cancer.

2.1. Data

2.1.1. Breast cancer data from van't Veer (NKI_97)

The first data set is reported by van't Veer et al [2] and referred to as NKI_97. The original van't Veer data consists of gene expression profiles and clinical information for 97 samples of primary breast cancer tumors, and each case is described by the expression levels of 24,481 genes. Fifty-one patients remained free from metastasis for at least 5 years and were metastasis-negative, and 46 cancer patients developed metastasis within 5 years and were metastasis-positive. All patients were <55 years old and were lymph node-negative. They had no tumor cells in local lymph nodes [2]. The data used in this study is a filtered version of the van't Veer data including gene expression values of 4948 genes in 97 tumor samples [2]. The data are publicly available at the "cancer data" R package (http://www.bioconductor.org/packages/release/data/experiment/html/cancerdata.html).

2.1.2. Breast cancer data from van de Vijver (NKI_295)

The second data set is reported by van de Vijver et al [4] and referred to as NKI_295. The data set provides the gene information for 295 primary breast cancer patients, of which 234 patients were new and the remaining 61 patients were involved in the first data set. Of the total 295 patients, 194 patients were metastasis-negative and 101 patients were metastasis-positive. Of the 234 new patients, 164 patients were metastasis-negative and 70 patients were metastasis-positive. Of the 61 patients involved in the first data set, 30 were metastasis-negative and 31 patients were metastasis-positive. The data is a filtered version of the van de Vijver data including gene expression values of 4948 genes in 295 tumor samples [4]. The data are publicly available at the "cancer data" R package.

2.1.3. Breast cancer data from the Wang study (VDX_286)

The last data set, reported by Wang et al [16] and referred to as VDX_286, contains 286 lymph node-negative breast cancer patients who had not received any adjuvant systemic treatment [16]. Among them, 106

patients had distant metastasis within 5 years of follow up and were considered as metastatic patients, while the rest were considered as nonmetastatic patients. A set of 22,283 genes is available for this data set. The data are publicly available at the "breast cancer VDX" R package.

2.2. Wavelet Transform

A wavelet is a "small wave", which has its energy concentrated in time. In signal processing, a transformation technique is used to transfer data in another domain where hidden information can be extracted. Wavelets have a nice feature of local description and separation of signal characteristics, and provide a tool for the analysis of transient or time-varying signals [11].

A wavelet is a set of orthonormal basis functions generated from dilation and translation of a single scaling function or father wavelet (φ) and a mother wavelet (ψ).

WTs are classified into two different categories: the continuous WT and the DWT. The DWT is a linear operation that operates on a data vector, transforming it into a wavelet coefficient. The idea underlying DWT is to express any function $f(t) \in L^2(R)$ in terms of ö (t) and ø (t) as follows:

$$f(t) = \sum_k c_0(k)\varphi(t-k) + \sum_k \sum_{\mathrm{j}=1} d_{\mathrm{j}}(k) 2^{\frac{-\mathrm{j}}{2}} \psi\left(2^{-\mathrm{j}}t - k\right)$$

$$= \sum_k c_{\mathrm{j}_0}(k) 2^{\frac{-\mathrm{j}0}{2}} \varphi\left(2^{-\mathrm{j}0}t - k\right) + \sum_k \sum_{\mathrm{j}=\mathrm{j}_0} d_{\mathrm{j}}(k) 2^{\frac{-\mathrm{j}}{2}} \psi\left(2^{-\mathrm{j}}t - k\right) \quad (1)$$

where φ(t), ψ(t), c_0, and d_{j} represent the scaling function, mother wavelet function, scaling coefficients (approximation coefficients) at scale zero, and detail coefficients at scale j, respectively. The variable k is the translation coefficient for the localization of gene expression data. The scales denote the different (low to high) scale bands. The variable symbol j_0 is the scale (level) number selected [10].

One-dimensional DWT decomposes a signal as a sum of wavelets at different time shifts and scales (frequencies) using DWT. For this purpose, the signal is passed through a series of high-pass and low-pass filters in order to analyze low as well as high frequencies in the signal as follows:

$$c_{j+1} = \sum_m h(m-2k) c_j \left(m\right) \quad (2)$$

$$d_{j+1} = \sum_m h_1(m-2k) c_j \left(m\right) \quad (3)$$

where $h(m-2k)$ and $h_1(m-2k)$ are the low-pass filters and high-pass filters, respectively.

At each level, the high-pass filter produces detail coefficients (wavelet coefficients) d_1, while the low-pass filter associated with the scaling function produces approximation coefficients (scaling coefficients) c_1. Subsequently, the approximation coefficients c_1 are split into two parts by using the same algorithm and are replaced by c_2 and d_2, and so on. This decomposition process is repeated until the required level is reached. The coefficient vectors are produced by down sampling and are only half the length of the signal or the coefficient vector at the previous level [12].

The main advantage of the WT is that each basis function is localized jointly in both the time and frequency domains. From a viewpoint of time-frequency, the approximation coefficients correspond to the larger-scale low-frequency components, and the detail coefficients correspond to the small-scale high-frequency components. Generally, the former can be used to approximate the original signal, and the latter represents some local details of the original signal [14,15].

There are different families of wavelets: symlets, coiflets, Daubechies, and biorthogonal wavelets. They vary in the various basic properties of wavelets, such as compactness. Haar wavelets, belonging to Daubechies wavelet family, are the most commonly used wavelets in database literature because they are easy to comprehend and fast to be computed.

2.3. Q-value

It is usual to simultaneously test many hundreds or thousands of genes in microarray studies to determine which are differentially expressed. Each of these tests will produce a p value. One main challenge in those studies is to find suitable multiple testing procedures that provide an accurate control of the error rates. Whereas the p value is a measure of significance in terms of the false positive rate, the q value is an approach used to measure statistical significance based on the concept of the false discovery rate. Similar to the p value, the q value gives each feature its own individual measure of significance [17].

2.4. Supervised WT

Firstly, any patients who remain free from metastasis for at least 5 years are placed into Class 1, otherwise into Class 2. The proposed DWT-based feature selection method consists of the following steps: (1) A t test is taken as the measure to identify differently expressed genes and a list of q values is derived. All the genes are ranked according to their corresponding q value and the required numbers of genes are selected from the list; and (2) in each step the top number of genes based on the q value are picked out. Then, this reduced set of genes is modeled by the one-dimensional DWT using Haar mother wavelet and finally, the wavelet approximation coefficients in the first and second levels of decomposition are used in the SVM model, respectively.

2.5. Supervised PCA

Bair and Tibshirani [8] and Bair et al [9] proposed supervised principal components regression. This procedure first picks out a subset of the gene expressions that correlates with response by using univariate selection, and then applies PCA to this subset. In our analysis, we pick out the top number of genes based on q values. We then apply PCA to this subset of genes, and in each step include the top numbers of principal components into a SVM model. The top numbers of principal components that will be comprised of at least 75% of the total variance are included in the SVM model.

2.6. SVM

The SVM model proposed by Vapnik [18] is a supervised learning method that is widely used in microarray data classification. Unlike many modeling techniques which aim to minimize the objective function (such as mean square error) for all instances, SVM attempts to find the hyperplanes that produce the largest separation between the decision function values for the instances located on the borderline between the two classes. The optimally identified hyperplane in the feature space corresponds to a nonlinear decision boundary in the input space. The SVM takes a set of input data with corresponding class labels and predicts a new input which belongs to the classes.

In the binary classification mode, given a training set of instance-label pairs (x_i, y_i) $i=1,2,\ldots,N$, where $x_i \in R^p$ and $y \in \{-1,+1\}$ SVM can be regarded as the solution of the following quadratic optimization problem:

$$\min_{\mathrm{W,b},\gamma} \frac{1}{2} W^T W + C \sum_{i=1}^{N} \gamma_i \tag{4}$$

$$subject\ to\ y_i\left(W^T \varphi(x_i) + b\right) \geq 1 - \gamma_i, \gamma_i \geq 0$$

where the training data are mapped to a higher dimensional space by the function φ and C is a user-defined penalty parameter on the training error that controls the trade-off between classification errors and the complexity of the model. By solving the optimization problem (1) by finding the parameters w and b for a given training set, a decision hyperplane over an n-dimensional input space that produces the maximal margin in the space is designed. Thus, the decision function can be formulated as follows:

$$f(x) = sign\left(W^T \varphi(x) + b\right) \tag{5}$$

SVM can derive the optimal hyperplane for nonlinearly separated data by mapping the impute data into the n-dimensional space using kernel function $[K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)]$. There are four basic kernels: linear, polynomial, radial basic function, and sigmoid [18,19].

In this study, the goal of SVM modeling was to classify patients who had a high risk of breast cancer recurrence. The predictive performance of the SVM-classifier was reported based on sensitivity, specificity, accuracy, and the area under the receiver operating characteristic curve (AUC). These criteria are defined as follows: (TP = true positive; TN = true negative; FN = false negative; and FP = false positive):

Accuracy: ACC $= \frac{TP+TN}{TP+FP+TN+FN}$

Sensitivity: SN $= \frac{TP}{TP+FN}$

Specificity: SP $= \frac{FP}{FP+TN}$

The method is implemented using MATLAB r2012a software (MATLAB Release 2012a, the MathWorks, Inc., Natick, Massachusetts, United States) and R statistical package (e1071, q value).

2.7. Cross data set comparison

To avoid over fitting and to provide a realistic evaluation, the cross data method was used. In this method, features obtained from one data set were used to construct classifiers for the other data set. In this regard, common patients in the NKI_295 and NKI_97 data were removed and the remaining data (NKI_234) were used as a test data set. This method was implemented using genes selected from NKI_234 breast cancer data as input in the supervised wavelet method in the NKI_61 data.

3. Results

The t test statistics were used to identify discriminative genes in each data set. After selecting the top ranked genes based on q values, one-dimensional WT in the first and second levels was applied to these preselected genes. SVMs with three types of kernels—linear, sigmoid, and radial, were used based on wavelet approximation coefficients in the first and the second levels of decomposition. For further assessment of the reported subsets of 70 genes selected by van't Veer et al [2] (for NKI_97 and NKI_295) and 76 signature genes selected by Wang et al [16] (for VDX_286), the supervised wavelet method and supervised PCA were applied. The predictive performance of SVM models was tested by cross-validation, consisting of 10 times 10-folding experiments. The results of supervised wavelet and supervised PCA for the three data sets are shown in Tables 1–3, respectively.

In the NKI_97 data set, the results showed that the SVM with radial kernels based on wavelet approximation coefficients in the first level extracted from 58 preselected genes had the best performance in terms of the evaluation criteria with regard to accuracy (83.11) as well as AUC (83.45). In addition, the SVM with radial kernel based on the first supervised PCA computed based on 84 preselected genes had the best performance

in terms of accuracy (79.22) as well as specificity (83.25), sensitivity (75.22), and AUC (79.24). In both methods (supervised wavelet and supervised PCA), the classifier performance based on the 70 genes selected by q values was better than the 70 gene signature from the van't Veer study (Table 1).

In the NKI_295 data set (Table 2), the results showed that the SVM with radial kernels based on wavelet approximation coefficients in the first level extracted from 91 preselected genes had the best performance in terms of the evaluation criteria, with the highest accuracy (75.37) as well as AUC (70.03). In addition, the SVM with linear kernel based on the first supervised PCA computed based on 91 preselected genes had the best performance in terms of accuracy (73.03) as well as AUC (66.63). In both methods (supervised wavelet and supervised PCA), the classifier performance based on the 70 genes selected by q values was better than the 70 gene signature from the van't Veer study.

In the VDX_286 data set (Table 3), the results showed that the SVM with linear kernels based on wavelet approximation coefficients in the second level extracted from 67 preselected genes had the best performance with the highest accuracy (79.21) as well as AUC (76.04). In addition, the SVM with linear kernel based on the first supervised PCA computed based on 67 preselected genes had the best performance in terms of accuracy (76.00) as well as AUC (74.71). In both methods (supervised wavelet and supervised PCA), the classifier performance based on the selected 76 genes using t statistics was better than the 76 gene signature identified in the Wang study.

To evaluate the reproducibility of the proposed method, a cross data-set comparison was also performed. As shown in Table 4, the results confirmed that

Table 1. Results for supervised wavelet and supervised principal component analysis (PCA): NKI_97, 10 times 10-fold cross-validation.

Method	No. of preselected genes.	Method	Accuracy	Sensitivity	Specificity	AUC
SVM (linear)	70 genes (van't Veer)	Wavelet (Db1.1)	77.11	78.30	76.15	77.22
		Wavelet (Db1.2)	69.11	64.47	73.00	68.74
		Supervised PCA	73.77	75.72	71.84	73.78
SVM (radial)	70 genes (van't Veer)	Wavelet (Db1.1)	77.55	82.28	73.24	77.76
		Wavelet (Db1.2)	75.66	82.20	69.76	75.98
		Supervised PCA	71.77	71.25	72.21	71.73
SVM (sigmoid)	70 genes (van't Veer)	Wavelet (Db1.1)	78.88	78.57	79.18	78.87
		Wavelet (Db1.2)	71.88	74.82	69.26	72.04
		Supervised PCA	68.77	67.58	69.73	68.66
SVM (linear)	70 genes	Wavelet (Db1.1)	72.33	67.55	76.38	71.97
		Wavelet (Db1.2)	76.44	75.53	77.24	76.38
		Supervised PCA	74.00	72.51	75.31	73.91
SVM (radial)	70 genes	Wavelet (Db1.1)	82.77	90.14	74.46	82.30
		Wavelet (Db1.2)	82.00	88.47	76.21	82.34
		Supervised PCA	75.88	75.22	76.52	75.87
SVM (sigmoid)	70 genes	Wavelet (Db1.1)	77.44	86.74	68.93	77.84
		Wavelet (Db1.2)	77.00	82.86	71.72	77.29
		Supervised PCA	78.22	76.83	79.45	78.14
SVM (linear)	$q < 0.02$ (84 genes)	Wavelet (Db1.1)	71.00	68.40	73.09	70.75
		Wavelet (Db1.2)	72.88	72.09	73.67	72.88
		Supervised PCA	78.00	78.01	77.98	78.00
SVM (radial)	$q < 0.02$ (84 genes)	Wavelet (Db1.1)	82.55	87.55	78.21	82.88
		Wavelet (Db1.2)	81.66	84.47	79.00	81.73
		Supervised PCA	79.22	83.25	75.22	79.24
SVM (sigmoid)	$q < 0.02$ (84 genes)	Wavelet (Db1.1)	79.88	88.17	72.53	80.35
		Wavelet (Db1.2)	78.88	86.62	70.94	78.78
		Supervised PCA	75.55	80.00	71.48	75.74
SVM (linear)	$q < 0.01$ (58 genes)	Wavelet (Db1.1)	73.77	76.62	71.34	73.98
		Wavelet (Db1.2)	70.88	67.78	73.95	70.86
		Supervised PCA	76.66	79.36	74.07	76.71
SVM (radial)	$q < 0.01$ (58 genes)	Wavelet (Db1.1)	83.11	88.27	78.63	83.45
		Wavelet (Db1.2)	82.33	85.11	79.55	82.33
		Supervised PCA	77.33	82.43	72.72	77.58
SVM (sigmoid)	$q < 0.01$ (58 genes)	Wavelet (Db1.1)	80.66	89.69	72.51	81.10
		Wavelet (Db1.2)	80.77	85.77	76.07	80.92
		Supervised PCA	76.00	80.87	71.86	76.37

AUC = area under the receiver operating characteristic curve; SVM = support vector machine.

Table 2. Results for supervised wavelet and supervised principal component analysis (PCA): NKI_295, 10 times 10-fold cross-validation.

Method	No. of preselected genes	Method	Accuracy	Sensitivity	Specificity	AUC
SVM (linear)	70 genes (van't Veer)	Wavelet (Db1.1)	65.10	38.32	77.82	58.07
		Wavelet (Db1.2)	66.13	29.71	84.33	57.02
		Supervised PCA	67.00	28.55	87.38	57.97
SVM (radial)	70 genes (van't Veer)	Wavelet (Db1.1)	70.96	32.82	90.37	61.59
		Wavelet (Db1.2)	67.96	26.37	88.64	57.50
		Supervised PCA	65.72	18.36	91.14	54.75
SVM (sigmoid)	70 genes (van't Veer)	Wavelet (Db1.1)	63.17	24.70	81.82	53.26
		Wavelet (Db1.2)	64.55	19.25	88.10	53.67
		Supervised PCA	66.27	23.73	89.04	56.39
SVM (linear)	70 genes	Wavelet (Db1.1)	70.20	48.68	81.29	64.98
		Wavelet (Db1.2)	72.65	53.08	82.52	67.80
		Supervised PCA	69.37	45.83	81.71	63.77
SVM (radial)	70 genes	Wavelet (Db1.1)	71.13	36.98	88.76	62.87
		Wavelet (Db1.2)	70.06	39.92	86.22	63.07
		Supervised PCA	70.10	34.41	89.37	61.89
SVM (sigmoid)	70 genes	Wavelet (Db1.1)	65.79	43.03	77.08	60.06
		Wavelet (Db1.2)	63.44	44.50	73.72	59.11
		Supervised PCA	68.86	33.92	87.55	60.74
SVM (linear)	$q < 0.001$ (56 genes)	Wavelet (Db1.1)	69.68	48.65	80.87	64.76
		Wavelet (Db1.2)	67.20	41.12	80.87	60.99
		Supervised PCA	71.68	46.81	84.56	65.68
SVM (radial)	$q < 0.001$ (56 genes)	Wavelet (Db1.1)	70.37	33.90	89.40	61.65
		Wavelet (Db1.2)	65.72	28.30	86.48	57.39
		Supervised PCA	70.82	40.54	86.62	63.58
SVM (sigmoid)	$q < 0.001$ (56 genes)	Wavelet (Db1.1)	65.79	44.68	76.53	60.60
		Wavelet (Db1.2)	66.37	41.38	79.49	60.43
		Supervised PCA	71.10	45.46	84.21	64.83
SVM (linear)	$q < 0.002$ (91 genes)	Wavelet (Db1.1)	72.37	46.50	86.00	66.25
		Wavelet (Db1.2)	70.43	80.97	57.00	67.24
		Supervised PCA	73.03	46.51	86.76	66.63
SVM (radial)	$q < 0.002$ (91 genes)	Wavelet (Db1.1)	75.37	52.85	87.21	70.03
		Wavelet (Db1.2)	74.58	49.18	86.48	67.83
		Supervised PCA	71.06	39.56	88.05	63.81
SVM (sigmoid)	$q < 0.002$ (91 genes)	Wavelet (Db1.1)	72.44	42.36	88.01	65.19
		Wavelet (Db1.2)	74.34	47.21	88.38	67.80
		Supervised PCA	69.10	49.47	78.63	64.05

AUC = area under the receiver operating characteristic curve; SVM = support vector machine.

the supervised wavelet method also had an acceptable performance, although the improvements were not as high as in the inner data set comparison. The results of other studies based on the same data sets are shown in Table 5. It can be seen that the proposed method had a higher capability for the prediction of metastasis than the other studies [20–29].

4. Discussion

This study proposed a new method based on WT to develop a novel predictive model for the prediction of breast cancer metastasis. Furthermore, the performance of this method was compared with supervised PCA.

The main purpose of the feature extraction method using WT is that the approximation coefficients usually comprise the majority of the important information [11]. In addition, the powerful capability of the DWT to compress the signal energy makes it a good candidate for feature extraction applications. The DWT compresses most of the energy from the input signal and concentrates it in a few high-magnitude coefficients in the transformed matrix.

The wavelet feature extraction method does not depend on the training data set to obtain the basis of feature space compared to the PCA method. Therefore, the wavelet feature extraction method dramatically reduces the computation load compared to PCA [11,12].

Considering the fact that most genes are irrelevant to patients' metastasis, we analyzed the reduced data set given by selecting genes that were significantly related to metastasis based on the t test statistics. If the WT is performed directly by using all of the genes in a data set,

Table 3. Results for supervised wavelet and supervised principal component analysis (PCA): VDX_286, 10 times 10-fold cross-validation.

Method	No. of preselected genes	Method	Accuracy	Sensitivity	Specificity	AUC
SVM (linear)	76 genes (Wang)	Wavelet (Db1.1)	64.42	44.42	76.25	60.33
		Wavelet (Db1.2)	66.39	44.86	79.13	61.99
		Supervised PCA	68.17	39.13	85.82	62.47
SVM (radial)	76 genes (Wang)	Wavelet (Db1.1)	63.89	35.74	79.77	57.75
		Wavelet (Db1.2)	65.10	28.97	87.45	58.21
		Supervised PCA	67.82	33.97	87.88	60.92
SVM (sigmoid)	76 genes (Wang)	Wavelet (Db1.1)	66.92	45.49	79.66	62.58
		Wavelet (Db1.2)	65.64	43.42	79.11	61.27
		Supervised PCA	67.39	43.54	81.28	62.41
SVM (linear)	76 genes	Wavelet (Db1.1)	75.17	61.97	83.02	72.50
		Wavelet (Db1.2)	76.35	59.94	85.99	72.96
		Supervised PCA	67.96	42.04	83.65	62.85
SVM (radial)	76 genes	Wavelet (Db1.1)	76.07	60.80	84.86	72.83
		Wavelet (Db1.2)	77.25	56.48	89.23	72.86
		Supervised PCA	67.32	37.17	85.37	61.27
SVM (sigmoid)	76 genes	Wavelet (Db1.1)	77.21	62.41	86.10	74.26
		Wavelet (Db1.2)	71.57	61.79	77.34	69.56
		Supervised PCA	68.10	42.85	82.77	62.81
SVM (linear)	$q < 0.04$ (67 genes)	Wavelet (Db1.1)	78.21	67.05	84.60	75.83
		Wavelet (Db1.2)	79.21	64.46	87.61	76.04
		Supervised PCA	76.00	68.76	80.66	74.71
SVM (radial)	$q < 0.04$ (67 genes)	Wavelet (Db1.1)	77.00	58.65	87.56	73.10
		Wavelet (Db1.2)	75.17	54.41	88.33	71.37
		Supervised PCA	75.00	60.97	83.68	72.33
SVM (sigmoid)	$q < 0.04$ (67 genes)	Wavelet (Db1.1)	77.03	65.75	83.54	74.65
		Wavelet (Db1.2)	78.50	66.79	85.59	76.19
		Supervised PCA	75.21	64.96	81.63	73.30
SVM (linear)	$q < 0.05$ (86 genes)	Wavelet (Db1.1)	77.00	67.04	83.02	75.03
		Wavelet (Db1.2)	78.17	65.57	85.62	75.60
		Supervised PCA	75.96	66.14	82.11	74.12
SVM (radial)	$q < 0.05$ (86 genes)	Wavelet (Db1.1)	75.96	55.15	88.20	71.68
		Wavelet (Db1.2)	76.17	53.57	89.45	71.51
		Supervised PCA	75.57	63.50	82.98	73.24
SVM (sigmoid)	$q < 0.05$ (86 genes)	Wavelet (Db1.1)	77.32	66.18	83.91	75.04
		Wavelet (Db1.2)	74.67	59.40	83.36	71.38
		Supervised PCA	74.28	65.61	79.19	72.40

AUC = area under the receiver operating characteristic curve; SVM = support vector machine.

there is no guarantee that the resulting wavelet coefficients will be related to metastasis. Thus, this study introduced a supervised form of WT that can be considered as a supervised wavelet. After extracting supervised wavelet approximation coefficients using discrete Haar WT, these coefficients had higher predictive performances than the first three principal components. Therefore, our results suggested that the wavelet coefficients are the efficient way to characterize the features of high-dimensional microarray data. Because the performance of the proposed supervised wavelet method is likely to be improvable compared to some other studies, we conclude that this method is worth further investigation as a tool for cancer patient classification based on gene expression data. For example, to achieve optimal classification performance, a suitable combination of the classifier and the gene selection method needs to be specifically selected for a given data set.

Some studies reported misclassification rates that were obtained by the application of their classifier to a one splitting of the test and training set. For example, van't Veer et al [2] developed a 70-gene classifier predicting a distant metastasis of breast cancer. In the training set, the classifier predicted the class of 65/78 cases correctly (i.e., with an accuracy of 83.3%, corresponding to a weighted accuracy of 83.6%), whereas in the test set it predicted the class of 17/19 cases correctly (i.e., with an accuracy of 89.5%, corresponding to a weighted accuracy of 88.7%). However, in the present study, in order to avoid the over fitting problem, we followed the 10 times 10-fold cross-validation for

Table 4. External validation for supervised wavelet: NKI_234_61, 10 times 10-fold cross-validation.

Method	No. of preselected genes	Wavelet	Accuracy	Sensitivity	Specificity	AUC
SVM (linear)	70 genes	Db1. Level 1	67.83	75.63	59.15	67.39
		Db1. Level 2	64.33	69.45	58.82	64.13
SVM (radial)	70 genes	Db1. Level 1	64.50	72.47	54.94	63.71
		Db1. Level 2	67.66	67.94	67.36	67.65
SVM (sigmoid)	70 genes	Db1. Level 1	65.66	72.93	58.24	65.59
		Db1. Level 2	62.16	56.06	68.47	62.27
SVM (linear)	$q < 0.00$ (13 genes)	Db1. Level 1	64.00	68.81	59.34	64.07
		Db1. Level 2	61.50	53.96	69.82	61.89
SVM (radial)	$q < 0.003$ (13 genes)	Db1. Level 1	71.83	78.33	65.33	71.83
		Db1. Level 2	69.00	70.16	67.86	69.01
SVM (sigmoid)	$q < 0.003$ (13 genes)	Db1. Level 1	70.66	65.06	76.73	70.90
		Db1. Level 2	68.83	67.89	69.76	68.83

AUC = area under the receiver operating characteristic curve; SVM = support vector machine.

Table 5. Previously published analyses for the breast cancer data.

	No. of samples	Feature selection	Classifier	Measure	Validation method
Current study	97	Supervised wavelet	SVM radial kernel	Accuracy: 83.11	CV
		Supervised PCA	SVM radial kernel	Accuracy: 79.22	
	295	Supervised wavelet	SVM radial kernel	Accuracy: 75.37	
		Supervised PCA	SVM linear kernel	Accuracy: 73.03	
	286	Supervised wavelet	SVM linear kernel	Accuracy: 79.21	
		Supervised PCA	SVM linear kernel	Accuracy: 76.00	
Michiels et al (2005) [20]	97	Correlation	Nearest-centroid	Accuracy: 68.00	CV
Peng (2005) [23]	97	Signal to noise ratio	SVM	Accuracy: 75.00	Leave-one-out CV
		Signal to noise ratio	Bagg & Boost SVM	Accuracy: 77.00	
		Subsampling	Ensemble SVM	Accuracy: 81.00	
Pochet et al (2004) [24]	78+19*	None	LS-SVM linear kernel	Accuracy: 69.00	Leave-one-out CV
		None	SVM RBF kernel	Accuracy: 69.00	
		None	SVM linear kernel	Accuracy: 52.00	
Alexe et al (2006) [22]	78+19	Support set identified by logical analysis of data	SVM linear kernel	Accuracy: 77.00	CV
			Artificial NN	Accuracy: 79.00	
			Logistic regression	Accuracy: 78.00	
			Nearest neighbors	Accuracy: 76.00	
			Decision trees (C4.5)	Accuracy: 67.00	
Jahid et al (2012) [26]	295	Steiner tree based method	SVM	Accuracy: 62.00	CV
	286			Accuracy: 61.00	
Chuang et al (2007) [25]	295	Subnetwork marker	SVM	Accuracy: 72.00	CV
	286			Accuracy: 62.00	
van Vliet et al (2012) [21]	295	Filtering approach (t test)	Nearest mean classifier	AUC: 73.80	CV
Dehnavi et al (2013) [27]	286	Rough-set theory	Neuro-fuzzy System	Accuracy: 78.00	10-fold CV
Lee et al (2011) [28]	286	Modules with condition responsive correlations	Naïve Bayesian classifier	AUC: 0.62	Leave-one-out CV
Jahid et al (2014) [29]	295	Patient−patient co-expression networks	PC-classifier	AUC: 0.78	Leave-one-out CV
			Dagging	AUC: 0.72	
			AdaBoost	AUC: 0.66	
	286		PC-classifier	AUC: 0.68	
			Dagging	AUC: 0.61	
			AdaBoost	AUC: 0.55	

AUC = area under the receiver operating characteristic curve; CV = cross validation; PCA = principal component analysis; RBF = radial basic function; SVM = support vector machine.

evaluating the SVM classifier. The evaluation of the classifier based on one test set is very impressed with the data splitting process.

Future investigations can focus on different ways of preselecting genes in the first stage of the proposed method. For example, rather than ranking genes based on their *t* test scores, one would use a different metric to measure the association between a given gene and metastasis occurrence. By contrast, another mother wavelet and a different level of decomposition can be studied. In this study, gene expression data were employed as predictors. However, prediction performance may be improved by adding other covariates such as age, lymph node status, tumor size, and histological grade. It is likely that the classification performances could be improved with the use of some other classifiers.

This study confirmed that the SVM model based on the supervised wavelet feature extraction method was superior with regards to predictive performance than the supervised PCA and some other studies. Gene expression profiling can help to distinguish between patients at high risk and those at low risk for developing distant metastases, therefore, this technology and other high-throughput techniques are helping to alter our view of breast cancer and provide us with new tools for molecular diagnoses. These results exhibit the possibility of developing a new tool using wavelets for the dimension reduction of microarray data sets in the classification framework and therefore, the use of this method in similar classification problems is recommended.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgments

This study is part of a PhD thesis in Biostatistics (Grant no. 16/35/3500). The authors thank the Vice-Chancellor for Research and Technology of Hamadan University of Medical Sciences, Iran, for approving the project and providing financial support.

References

1. Ahr A, Kam T, Solbach C, et al. Identification of high risk breast cancer patients by gene-expression profiling. Lancet 2002 Jan; 359(9301):131−2.
2. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002 Jan;415(6871):530−6.
3. Lee TB. Comparison of breast cancer screening results in Korean middle-aged women: A hospital-based prospective cohort study. Osong Public Health Res Perspect 2013 Aug;4(4):197−202.
4. van de Vijver MJ, Yudong HE, van't Veer LJ, et al. A gene expression signature as a predictor of survival in breast cancer. N Engl J Med 2002 Dec;347(25):1999−2009.
5. Bovelstad HM, Nygard S, Storvold HL, et al. Predicting survival from microarray data−a comparative study. Bioinformatics 2007 Jun;23(16):2080−7.
6. Li L, Li H. Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics 2004 Dec; 20(18):3406−12.
7. Wessel N, Wieringen V, Kun D, et al. Survival prediction using gene expression data: A review and comparison. Comput Stat Data Anal 2007 Mar;53(5):1590−603.
8. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2002 Apr;2(4): 511−22.
9. Bair E, Hastie T, Paul D, et al. Prediction by supervised principal components. J Am Statist Assoc 2006 Jan;101(473):119−36.
10. Tokuyasu TA, Albertson D, Pinkel D, et al. Wavelet transforms for the analysis of microarray experiments. In: Bioinformatics Conference, CSB 2003. Proceedings of the 2003 IEEE, 11−14 Aug. 2003. p. 429−30.
11. Liu Y. Wavelet feature extraction for high-dimensional microarray data. Neurocomputing 2009 Jan;72(4−6):985−90.
12. Liu Y. Feature extraction and dimensionality reduction for mass spectrometry data. Comput Biol Med 2009 Sep;39(9):818−23.
13. Liu Y. Dimensionality reduction and main component extraction of mass spectrometry cancer data. Knowl-Based Syst 2012 Feb;26: 207−15.
14. Sarhan AM. Wavelet-based feature extraction for DNA microarray classification. Artif Intell Rev 2013 Mar;39(3):237−49.
15. Nanni L, Lumini A. Wavelet selection for disease classification by DNA microarray data. Expert Syst Appl 2011 Jan;38(1):990−5.
16. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005 Feb;365(9460):671−9.
17. Storey JD, Tibshirani R. Statistical significance for genome-wide experiments. PNAS 2003 Aug;100(16):9440−5.
18. Vapnik V. Statistical Learning Theory. 2nd ed. New York: Wiley; 1998.
19. Cortes C, Vapnik V. Support vector networks. Mach Learn 1995 Sep;20(3):273−97.
20. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005 Feb;365(9458):488−92.
21. van Vliet MH, Horlings HM, van de Vijver MJ, et al. Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. PLoS ONE 2012 Jul;7(7):e40358.
22. Alexe G, Alex S, Axelrod DE. Breast cancer prognosis by combinatorial analysis of gene expression data. Breast Cancer Res 2006 Jul;8(4):R41.
23. Peng YH. Robust ensemble learning for cancer diagnosis based on microarray data classification. Adv Data Mining Appl 2005;3584: 564−74.
24. Pochet N, Smet FD, SuykensJ AK, et al. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics 2004 Jul; 20(17):3185−95.
25. Chuang HY, Lee E, Liu YT, Lee D, et al. Network-based classification of breast cancer metastasis. Mol Syst Biol 2007 Oct;3:140.
26. Jahid MJ, Ruan JA. Steiner tree-based method for biomarker discovery and classification in breast cancer metastasis. BMC Genomics 2012;13(Suppl. 6):S8.
27. Dehnavi AM, Sehhati MR, Rabbani H. Hybrid method for prediction of metastasis in breast cancer patients using gene expression signals. J Med Signals Sens 2013 Apr;3(2):79−86.
28. Lee S, Lee E, Lee KH, et al. Predicting disease phenotypes based on the molecular networks with condition-responsive correlation. Int J Data Min Bioinform 2011;5(2):131−42.
29. Jahid MJ, Huang TH, Ruan J. A personalized committee classification approach to improving prediction of breast cancer metastasis. Bioinformatics 2014 Jul;30(13):1858−66.