Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 578–587
Expert Systems with Applications
www.elsevier.com/locate/eswa

Prediction model building and feature selection with support vector machines in breast cancer diagnosis

Cheng-Lung Huang [a], Hung-Chang Liao [b,*], Mu-Chen Chen [c]

[a] Department of Information Management, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan
[b] Department of Health Services Administration, Chung-Shan Medical University, Taichung, Taiwan
[c] Institute of Traffic and Transportation, National Chiao Tung University, Taiwan

Abstract

Breast cancer is a serious problem for the young women of Taiwan. Some medical researches have proved that DNA viruses are one of the high-risk factors closely related to human cancers. Five DNA viruses are studied in this research: specific types of HSV-1 (herpes simplex virus type 1), EBV (Epstein-Barr virus), CMV (cytomegalovirus), HPV (human papillomavirus), and HHV-8 (human herpesvirus-8). The purposes of this study are to obtain the bioinformatics about breast tumor and DNA viruses, and to build an accurate diagnosis model about breast cancer and fibroadenoma. Research efforts have reported with increasing confirmation that the support vector machine (SVM) has a greater accurate diagnosis ability. Therefore, this study constructs a hybrid SVM-based strategy with feature selection to render a diagnosis between the breast cancer and fibroadenoma and to find the important risk factor for breast cancer. The results show that {HSV-1, HHV-8} or {HSV-1, HHV-8, CMV} are the most important features and that the diagnosis model achieved high classification accuracy, at 86% of average overall hit rate. A Linear discriminate analysis (LDA) diagnosis model is also constructed in this study. The LDA model shows that {HSV-1, HHV-8, EBV} or {HSV-1, HHV-8} are significant factors which are similar to that of the SVM-based classifier. However, the classificatory accuracy of the SVM-based classifier is slightly better than that of LDA in the negative hit ratio, positive hit ratio, and overall hit ratio.

Keywords: Breast tumor; Support vector machines; Feature selection

1. Introduction

In Taiwan, breast cancer is the second most occurring cancer and the death rate from breast cancer is increasing each year (Cancer Registry Annual Report, 2004; Overview of Public Health, 1998). Almost 64.1% of women with breast cancer are diagnosed before the age of 50 and 29.3% of women with breast cancer are diagnosed before the age of 40 (Cheng, Tsou, Liu, & Jian, 2000). On the average, the women diagnosed with breast cancer in Taiwan are younger than those in the US. Some recognized factors will increase the risk of breast cancer; however, the causes are still unknown. Hence, it is difficult for medical professionals to treat breast cancer with the appropriate preventive methods (Ziegler et al., 1993). Yu, Rohan, Cook, Howe, and Miller (1992) investigated the risk factors for fibroadenoma in a case-control study involving 117 fibroadenoma cases in Australia. This study shows that fibroadenoma shared some risk factors with breast cancer. It is estimated that DNA viruses, which emerge as major causal factors, contribute 20% to the occurrence of human cancers (Dimmock & Primrose, 1994). DNA viruses, as causes, are closely related to the human cancers as part of the high-risk factors. These DNA viruses include specific types of HSV-1 (herpes simplex virus type 1), EBV (Epstein-Barr virus), CMV (cytomegalovirus), HPV (human papillomavirus), and HHV-8 (human herpesvirus-8) (Dimmock & Primrose, 1994).

* Corresponding author. Tel.: +886 4 2473 0022x17170.
E-mail address: hcliao@csmu.edu.tw (H.-C. Liao).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.09.041

Wu et al. (2001) studied biological purging of breast cancer cells using an attenuated replication-competent HSV-1 in human hematopoietic stem cell transplantation. Hu et al. (2004) developed a second generation of genetically modified HSV-1 with paclitaxel in the treatment of breast cancer in vitro. Wang and Vos (1996) studied a hybrid herpesvirus infectious vector, pH300, based on HSV-1 and EBV for gene transfer to human cells in vitro and in vivo.

In other studies, Liu et al. (1993) indicated that the EBV hybridoma technique offers several advantages over the other hybridoma systems for generating anti-breast cancer human monoclonal antibodies. Katano et al. (1995) established a Breast-M and an EBV-infected B-cell line (Hairy-BM) from breast tumor tissue. Also, Yip, Hawkins, Clark, and Ward (1997) used the EBV-transformed peripheral blood mononuclear cells from individuals with breast cancer for the construction of human immunoglobulin gene libraries. Fina et al. (2001) studied the frequency and genome load of EBV in 509 breast cancers from various geographical areas.

Recently, Grinstein et al. (2002) demonstrated EBV in carcinomas of the breast, lung, and other sites. Xue, Lampert, Haldane, Bridger, and Griffin (2003) also studied the EBV gene in human breast cancer. Then Huang, Chen, Hutt-Fletcher, Ambinder, and Hayward (2003) suggested that sporadic lytic EBV infection might contribute to polymerase chain reaction based (PCR-based) detection of EBV in traditionally non-virally associated epithelial malignancies. In addition, Ribeiro-Silva, Ramalho, Garcia, and Zucoloto (2004) studied whether there is a relationship between latent infection with EBV and p53 and p63 expression in breast carcinomas. Lastly, Baeyens et al. (2004) compared the radiation response in EBV cell lines derived from breast cancer patients with or without a BRCA1 mutation and revealed no significant difference.

In the previous studies of CMV for breast cancer, Strender et al. (1981) studied a group of 17 patients who had undergone modified radical mastectomy for breast cancer. Lee, Reimer, Oh, Campbell, and Schnitzer (1998) found that the caveolin expression was significantly reduced in human breast cancer cells provided that the caveolin cDNA linked to the CMV promoter was transfected into human mammary cancer cells. Still, Hamilton, Vince, Wolfman, and Cowell (1999) indicated that the constitutive expression of the gene under the control of CMV promoter in mouse fibroblasts results in cellular transformation and anchorage-independent growth.

Svane et al. (2002) analyzed the impact of high-dose chemotherapy on antigen-specific T cells responsive to CMV immunity in breast cancer patients. Akbulut, Zhang, Tang, and Deisseroth (2003) studied the cytotoxic effect of replication-competent adenoviral vectors carrying L-plastin promoter regulated E1A and cytosine deaminase genes on cancers of breast, ovary, and colon. A similar vector driven by the CMV promoter has also been constructed as a control. This treatment resulted in decreased tumor size and decreased tumor cell growth rate. Ma et al. (2004) demonstrated that the inhibition of the PKB-dependent survival pathway could promote apoptosis and thermosensitization in malignant breast cancer cells, with relative sparing of their normal counterpart. Zhu et al. (2004) showed that CXCR4 had a low expression of luciferase (0.32%) compared to that of the CMV promoter in mice live in vivo. The CXCR4 was proven to be a good candidate as a tissue-specific promoter for cancer gene therapy for melanoma and breast cancers.

Recent studies have revealed a possible role for HPV in the pathogenesis of breast cancer, although no definitive interaction was observed between types of oral contraceptives or with any recognized risk factor for breast cancer. Oral contraceptives may act as a promoter for HPV-induced carcinogenesis (La Vecchia, Tavani, Franceschi, & Parazzini, 1996). Chang et al. (1999) demonstrated a high frequency of abnormalities of this gene in human breast cancer. They found that there was no genomic deletion or rearrangement in spite of the presence of abnormal transcripts and no definite relationship between the abnormal transcripts and HPV infection. Liu et al. (2001) showed that 6 out of 17 (35%) types of breast cancers were identified as being HPV positive in the PCR/dot blot analysis with both the HPV E6-E7 and L1 primer sets. Widschwendter, Brunhuber, Wiedemair, Mueller-Holzner, and Marth (2004) suggested that HPV DNA might be transported from the original site of infection to the breast tissue by the bloodstream, and that it possibly existed in the carcinogenesis of breast neoplasia in some patients. Finally, the researches for HHV-8 and breast cancer have few citations in the literature. Klein and Klein (2005) studied the surveillance against tumor. HHV-8 is a relevant viral agent in this context. Andres (2005) found that Kaposi's sarcoma had a high incidence in the renal-transplanted population, and that it was related to HHV-8.

From the literature reviews, it can be seen that it is important to evaluate the associations among DNA viruses, HSV-1, EBV, CMV, HPV, and HHV-8 with breast cancer and fibroadenoma. In order to obtain the relationship between DNA viruses and breast tumors, this paper uses the support vector machines (SVM) to find the pertinent bioinformatics. Support vector machines were first suggested by Vapnik (1995) and have recently been used in a range of problems including pattern recognition (Pontil & Verri, 1998), bioinformatics (Yu, Ostrouchov, Geist, & Samatova, 2003), text categorization (Joachims, 1998), and cancer diagnosis (Lee & Lee, 2003; Lee, Mangasarian, & Wolberg, 2000; Liu et al., 2003).

When using SVM, two problems are confronted: how to choose the optimal input feature subset for SVM and how to set the best kernel parameters. These two problems are crucial because the feature subset choice influences the appropriate kernel parameters and vice versa (Fröhlich & Chapelle, 2003). Feature selection is an important issue in building classification systems. It is advantageous to limit the number of input features in a classifier in order to have

a good predictive and less computationally intensive model (Zhang, 2000). With a small feature set, the explanation of rationale for the classification decision can be more readily realized. In addition to the feature selection, proper model parameters setting can improve the SVM classification accuracy. The parameters that should be optimized include penalty parameter C and the kernel function parameters such as the gamma (γ) for the radial basis function (RBF) kernel.

To design a SVM, one must choose a kernel function, set the kernel parameters and determine a soft margin constant C. The grid algorithm is an alternative to finding the best C and gamma when using the RBF kernel function (Hsu & Lin, 2002).

To explore five DNA viruses – HSV-1, EBV, CMV, HPV, and HHV-8 – affecting the breast tumor diagnosed by using support vector machines, this study tried grid search to find the best SVM model parameters and used F-score calculation to select input features.

This paper is organized as follows. Section 2 describes basic SVM concepts. Section 3 describes three SVM-based strategies used in this research. Section 4 presents the experimental results from using the proposed method to diagnose the real world breast cancer data set. Section 5 gives remarks and provides a conclusion.

2. Basic concepts of SVM classifier

In this section, we will briefly describe the basic SVM concepts for typical two-class classification problems. These concepts can also be found in (Kecman, 2001, Schölkopf & Smola, 2000, & Cristianini & Shawe-Taylor, 2000).

Given a training set of instance-label pairs (x_i, y_i), $i = 1, 2, \ldots, m$ where $x_i \in R^n$ and $y_i \in \{+1, -1\}$, SVM finds an optimal separating hyperplane with the maximum margin by solving the following optimization problem:

$$\begin{aligned} \underset{w,b}{\text{Min}} \quad & \frac{1}{2} w^{\mathrm{T}} w \\ \text{subject to:} \quad & y_i(\langle w \cdot x_i \rangle + b) - 1 \geqslant 0 \end{aligned} \tag{1}$$

It is known that to solve this quadratic optimization problem one must find the saddle point of the Lagrange function:

$$L_p(w, b, \alpha) = \frac{1}{2} w^{\mathrm{T}} \cdot w - \sum_{i=1}^{m} (\alpha_i y_i (\langle w \cdot x_i \rangle + b) - 1) \tag{2}$$

where the α_i denotes Lagrange multipliers, hence $\alpha_i \geqslant 0$. The search for an optimal saddle point is necessary because the L_p must be minimized with respect to the primal variables w and b and maximized with respect to the non-negative dual variable α_i. By differentiating with respect to w and b, and introducing the Karush–Kuhn–Tucker (KKT) conditions for the optimum constrained function, then is transformed to the dual Lagrangian $L_D(\alpha)$:

$$\begin{aligned} \underset{\alpha}{\text{Max}} \quad & L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle \\ \text{subject to:} \quad & \alpha_i \geqslant 0, \quad i = 1, \ldots, m \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned} \tag{3}$$

To find the optimal hyperplane, a dual Lagrangian $L_D(\alpha)$ must be maximized with respect to non-negative α_i. The solution α_i for the dual optimization problem determines the parameters w^* and b^* of the optimal hyperplane. Thus, the optimal hyperplane decision function $f(x) = \mathrm{sgn}(\langle w^* \cdot x \rangle + b^*)$ can be written as

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i^* \langle x_i, x \rangle + b^* \right) \tag{4}$$

In a typical classification task, only a small subset of the Lagrange multipliers α_i usually tends to be greater than zero. Geometrically, these vectors are the closest to the optimal hyperplane. The respective training vectors having non-zero α_i are called support vectors, as the optimal decision hyperplane $f(x, \alpha^*, b^*)$ depends on them exclusively.

The above concepts can also be extended to the non-separable case (linear generalized SVM). In terms of these slack variables, the problem of finding the hyperplane that provides the minimum number of training errors (i.e., to keep the constraint violation as small as possible) has the formal expression as follows:

$$\begin{aligned} \underset{w,b,\xi}{\text{Min}} \quad & \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{m} \xi_i \\ \text{subject to:} \quad & y_i(\langle w \cdot x_i \rangle + b) + \xi_i - 1 \geqslant 0 \\ & \xi_i \geqslant 0 \end{aligned} \tag{5}$$

where C is a penalty parameter on the training error, and ξ_i is the non-negative slack variables. SVM finds the hyperplane that provides the minimum number of training errors (i.e., to keep the constraint violation as small as possible).

This optimization model can be solved using the Lagrangian method, which is almost equivalent to the method for solving the optimization problem in the separable case. One must maximize the dual variables Lagrangian:

$$\begin{aligned} \underset{\alpha}{\text{Max}} \quad & L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle \\ \text{Subject to:} \quad & 0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, m \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned} \tag{6}$$

To find the optimal hyperplane, a dual Lagrangian $L_D(\alpha)$ must be maximized with respect to non-negative α_i under the constraints $\sum_{i=1}^{m} \alpha_i y_i = 0$ and $0 \leqslant \alpha_i \leqslant C$. The penalty parameter C, which is now the upper bound on α_i, is determined by the user. Finally, the form of optimal hyperplane decision function is the same as (4).

The nonlinear SVM maps the training samples from the input space into a higher-dimensional feature space via a

mapping function Φ. The kernel function $k(x_i, x_j)$ defines an inner product as $k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$.

In the dual Lagrange (6), the inner products are replaced by the kernel function, and the nonlinear SVM dual Lagrangian $L_D(\alpha)$ (7) is similar with that in the linear generalized case

$$L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i \cdot x_j)$$

$$\text{Subject to}: \quad 0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, m \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{7}$$

Followed by the steps described in the linear generalized case, we obtain the decision function of the following form:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i^* \langle \Phi(x), \Phi(x_i) \rangle + b^* \right) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i^* \langle k(x, x_i) \rangle + b^* \right) \tag{8}$$

The kernel function we explored in our experiments was the radial basis function (RBF) which is defined by (9).

$$k(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \tag{9}$$

3. Experiments and methodologies

3.1. Data collection and data partition

The source of 80 data points (tissue samples), including 52 specimens of non-familial invasive ductal breast cancer from women and 28 mammary fibroadenomas, is the Chung-Shan Medical University Hospital (Tsai et al., 2005). After using PCR and Southern hybridization to screen for the presence of β-globin, it was discovered that the 80 specimens screened were DNA virus positive/negative for the presence of β-globin, the internal control.

To guarantee that the present results are valid and can be generalized for making predictions regarding new data, the data set is further randomly partitioned into training and independent testing sets via a stratified 5-fold cross validation. Each of the 5 subsets acts as an independent hold-out test set for the model trained with the rest of the 4 subsets. The advantages of k-fold cross validation are that the impact of data dependency is minimized and the reliability of the results can be improved (Salzberg, 1997). In addition, the classification models are developed with a huge portion of the accessible data (80% in this case) and all the data is utilized to test the trained models.

A pair of training and testing set is called a "fold" or a "group" in this study. As shown in Table 1, due to the number of cases (positive: 52, negative: 28) that can not be divided by 5; the size of each fold is not the same – the number of cases for fold #5 is 20, and for the others it is 15.

Table 1
The data set is further randomly partitioned into training and independent testing sets via a stratified 5-fold cross validation

	Size of training set	Size of testing set
Fold #1	65	15
Fold #2	65	15
Fold #3	65	15
Fold #4	65	15
Fold #5	60	20

3.2. Feature selection

Feature selection is an important issue in building classification systems. It is advantageous to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model (Zhang, 2000). With a small feature set, the explanation of rationale for the classification decision can be more easily realized. In the area of medical diagnosis, a small feature subset means lower test and diagnosis costs. F-score (Chen & Lin, 2005) is a simple technique that measures the discrimination of two sets of real numbers. Given training vectors x_k, $k = 1, 2, \ldots, m$, if the number of positive and negative instances are $n_+$and n_-, respectively, then the F-score of the ith feature is defined as follows (Chen & Lin, 2005):

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \tag{10}$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, and $\bar{x}_i^{(-)}$ are the averages of the ith feature of the whole, positive, and negative data sets, respectively; $x_{k,i}^{(+)}$ is the ith feature of the kth positive instance, and $x_{k,i}^{(-)}$ is the ith feature of the kth negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative (Chen & Lin, 2005).

3.3. Setting model parameters

In addition to the feature subset selection, proper model parameters setting can improve the SVM classification accuracy. The parameters that should be optimized include penalty parameter C and the kernel function parameters such as the gamma (γ) for the radial basis function (RBF) kernel. To design an SVM, one must choose a kernel function, set the kernel parameters and determine a soft margin constant C. With the RBF kernel, there are two parameters to be determined in the SVM model: C and gamma (γ). The grid search approach (Hsu, Chang, & Lin, 2003) is an alternative to finding the best C and gamma when using the RBF kernel function.

In the grid search approach, pairs of (C, γ) are tried and the one with the best cross-validation accuracy is chosen. After identifying a "better" region on the grid, a finer grid search on that region can be conducted. To get good generalization ability, grid search approach uses a

validation process to decide parameters. That is, for each of the k subsets of the data set D, create a training set $T = D - k$, then run a cross-validation process as follows (Chen & Lin, 2005; Hsu et al., 2003):

Step 1. Consider a grid space of (C,γ) with $\log_2 C \in \{-5,-4,\ldots,12\}$ and $\log_2\gamma \in \{-12,-13,\ldots,5\}$.
Step 2. For each hyperparameter pair (C,γ) in the search space, conduct k-fold cross validation on the training set.
Step 3. Choose the parameter (C,γ) that leads to the lowest CV (cross validation) error classification rate.
Step 4. Use the best parameter to create a model as the predictor.

Overall accuracy is averaged across all k partitions. These k accuracy values also give an estimate of the accuracy variance of the algorithms.

3.4. Setting model parameters using grid search and selecting input features using F-score

To build diagnosis models successfully, this study tried a SVM-based strategy using grid search to optimize model parameters and F-score calculation to select input features (see Fig. 1). The procedure of grid search is the same as that shown in Section 3.3. A cross-validation approach with $k = 5$ was also conducted to avoid overfitting during training process. The overall testing accuracy is averaged across all k partitions. That is, for each of the k subsets of the data set$D$, create a training set $T = D - k$, then run a cross-validation process as follows:

Step 1. Calculate and sort the F-scores.
Step 2. For the possible number of features f, $f \in \{1,2,\ldots,m\}$, where m is the total number of features in a data set, do the following steps:
(2.1) Keep the first f features according to the sorted F-scores.
(2.2) For the training set, calculate the average SVM accuracy using 5-fold cross validation (5-CV).
Step 3. Choose the f with the largest average 5-CV accuracy. Retrain the SVM with the training set, and predict the test accuracy with the test set.

4. Experimental results and discussion

4.1. Experimental results

In this experiment, the importance of each feature is measured by F-score, and the SVM parameters are opti-

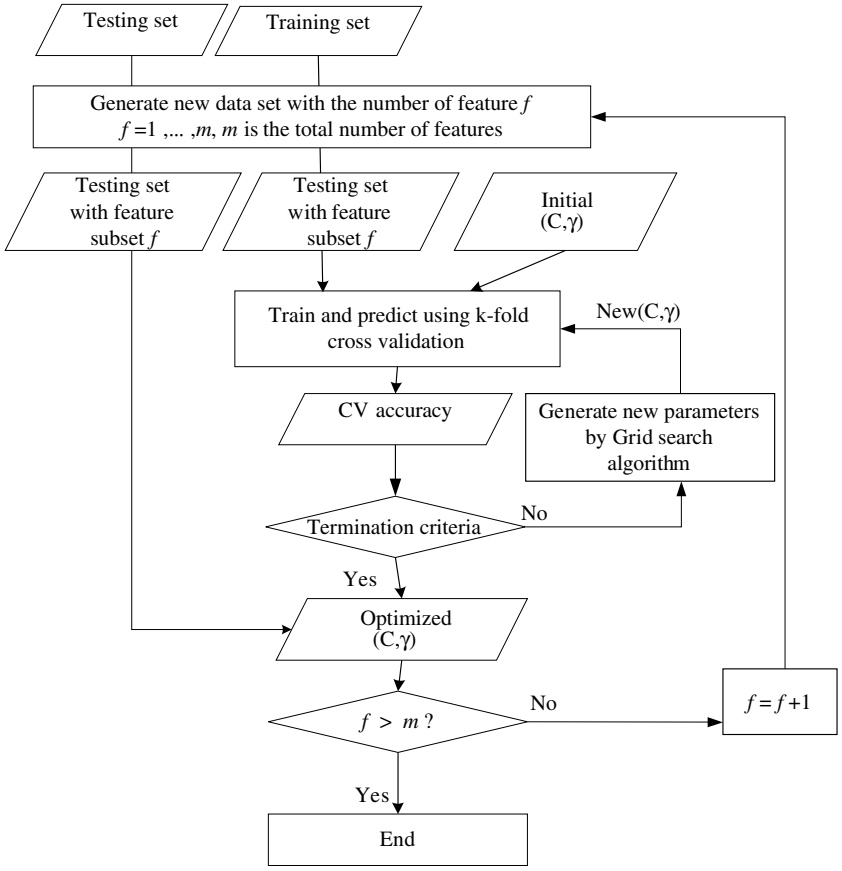

Fig. 1. The SVM-based strategy using grid search to optimize model parameters and F-score calculation to select input features.

Table 2
The relative feature importance with *F*-score

Feature	Fold #1	Fold #2	Fold #3	Fold #4	Fold #5	Average
HSV-1	0.388553	0.646448	0.334284	0.410013	0.615385	0.478937
HHV-8	0.353846	0.292308	0.240615	0.353846	0.272727	0.302668
CMV	0.105321	0.076555	0.033816	0.076555	0.040816	0.066613
EBV	0.002311	0.002311	0.006712	0.000617	0.002268	0.002844
HPV	0.00158	0.000287	0.005803	0.000296	0.001001	0.001793

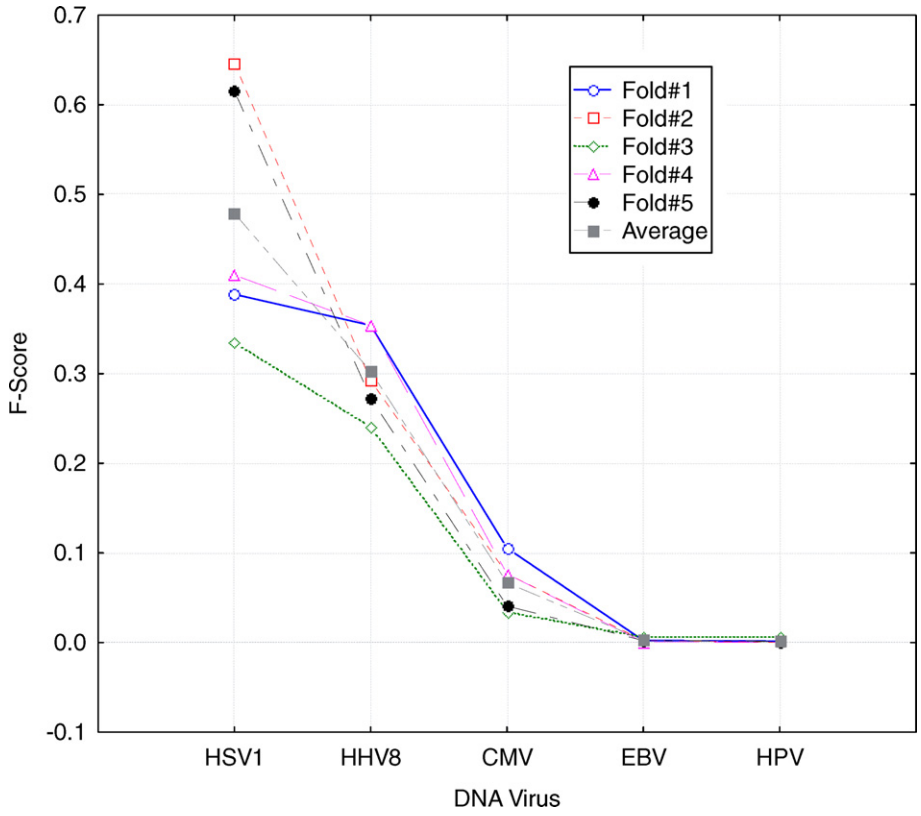

Fig. 2. The relative importance of DNA virus based on the *F*-score.

mized by grid search. Table 2 and Fig. 2 show the relative feature importance with *F*-score for each feature on each fold. The average *F*-score for HSV-1, HHV-8, CMV, EBV, and HPV (from high to low) are 0.478937, 0.302668, 0.066613, 0.002844, and 0.001793, respectively. The degree of breast tumor associated with DNA viruses, from high to low, are HSV-1, HHV-8, CMV, EBV, and HPV. Therefore, five models with a different number of features are constructed further to obtain the SVM classification models. As shown in Table 3, the five models with different feature subsets based on *F*-score are {HSV-1}, {HSV-1 HHV-8}, {HSV-1 HHV-8 CMV}, {HSV-1 HHV-8 CMV EBV}, and {HSV-1 HHV-8 CMV EBV HPV}.

Tables 4 and 5 show the training and testing accuracies for the five models, each model achieved a high average overall accuracy of above 80%. Among the five models, two models with feature subsets, {HSV-1, HHV-8} and {HSV-1, HHV-8, CMV} achieved the highest training and testing accuracy. For these two models, {HSV-1, HHV-8} and {HSV-1, HHV-8, CMV}, Table 6 shows their details for the best SVM parameters (c and gamma), training accuracy, and testing accuracy for each fold. Their average negative, positive, and overall hit rate for model achieved 0.71, 0.946666, and 0.866666, respectively.

Table 3
The five feature subsets based on the *F*-score

Model	Number of selected features	Features
#1	1	HSV-1
#2	2	HSV-1 HHV-8
#3	3	HSV-1 HHV-8 CMV
#4	4	HSV-1 HHV-8 CMV EBV
#5	5	HSV-1 HHV-8 CMV EBV HPV

Only two or three attributes, {HSV-1, HHV-8} or {HSV-1, HHV-8, CMV}, can achieve identical high accuracy. It is not necessary to include all features for the sake of cost saving. For accuracy, the positive hit ratio is higher

Table 4
Overall training accuracy for each feature subset

Number of selected features	SVM + GS + FS: Training accuracy						
	Fold #1	Fold #2	Fold #3	Fold #4	Fold #5	Average	Standard deviation
1	78.4615	83.0769	76.9231	78.4615	83.3333	80.05126	2.636905
2	84.6154	87.6923	84.6154	86.1538	88.3333	86.28204	1.534166
3	84.6154	87.6923	84.6154	86.1538	88.3333	86.28204	1.534166
4	84.6154	87.6923	84.6154	86.1538	88.3333	86.28204	1.534166
5	84.6154	87.6923	86.1538	86.1538	88.3333	86.58972	1.306426

Table 5
Overall testing accuracy for each feature subset

Number of selected features	SVM + GS + FS: Testing accuracy						
	Fold #1	Fold #2	Fold #3	Fold #4	Fold #5	Average	Standard deviation
1	0.866667	0.666667	0.933333	0.866667	0.7	0.806667	0.104137
2	0.933333	0.8	0.933333	0.866667	0.8	0.866667	0.059628
3	0.933333	0.8	0.933333	0.866667	0.8	0.866667	0.059628
4	0.933333	0.666667	0.8	0.866667	0.8	0.813333	0.088443
5	0.933333	0.666667	0.8	0.866667	0.8	0.813333	0.088443

Table 6
Detail testing accuracy for feature subset of size 2 and 3

	Features size = 2			Features size = 3		
	Negative hit ratio	Positive hit ratio	Overall hit ratio	Negative hit ratio	Positive hit ratio	Overall hit ratio
Fold #1	0.8	1	0.93333	0.8	1	0.93333
Fold #2	0.6	0.9	0.8	0.6	0.9	0.8
Fold #3	0.8	1	0.93333	0.8	1	0.93333
Fold #4	0.6	1	0.86667	0.6	1	0.86667
Fold #5	0.75	0.83333	0.8	0.75	0.83333	0.8
Average	0.71	0.946666	0.866666	0.71	0.946666	0.866666

than the negative hit ratio; on average, the overall hit ratio is highly accurate. The results reveal that the SVM model has a good disarmament performance in diagnosing breast cancer according to our data set.

4.2. Comparison with linear discriminate analysis

In this experiment, the importance of each feature and the classificatory accuracy is measured by Linear discriminate analysis (LDA). Table 7 shows the significance of each attribute. The feature subset {HSV-1, HHV-8, EBV} is included in the LDA model except for Fold #2. The attribute EBV is slightly insignificant in the model of Fold #2; therefore, only two attributes {HSV-1, HHV-8} are included. Table 8 shows the details of training and testing accuracy for each fold. Their average negative, positive, and overall hit rate for the model achieved 0.71, 0.893, and 0.833, respectively.

Table 9 shows that the accuracy of SVM is slightly superior to the accuracy of LDA. For the selected features, SVM includes {HSV-1, HHV-8, CMV} or {HSV-1, HHV-8}, while LDA includes {HSV-1, HHV-8, EBV} or {HSV-1, HHV-8}.

Table 7
The P-level of each attribute for LDA

	HSV-1	HHV-8	EBV	CMV	HPV
Fold #1	0.000025	0.000192	0.013389	0.071484	0.615577
Fold #2	1.56E−07	0.000801	0.068939	0.184108	0.309406
Fold #3	2.70E−07	0.000040	0.002272	0.051081	0.096721
Fold #4	0.000002	0.000028	0.027227	0.060013	0.243957
Fold #5	9.12E−08	0.001157	0.020698	0.289498	0.522833

Table 8
Training and testing accuracy for LDA

	Training			Testing		
	Negative hit ratio	Positive hit ratio	Overall hit ratio	Negative hit ratio	Positive hit ratio	Overall hit ratio
Fold #1	69.56522	92.85714	84.61539	0.800	1.000	0.933
Fold #2	73.91304	90.47619	84.61539	0.600	0.900	0.800
Fold #3	86.95652	88.09524	87.69231	0.800	0.900	0.867
Fold #4	71.42857	94.23077	86.25000	0.600	1.000	0.867
Fold #5	70.00000	97.50000	88.33334	0.750	0.667	0.700
Average	74.37267	92.63187	86.30129	0.710	0.893	0.833

Table 9
Comparison summary between SVM and LDA

	Negative hit ratio	Positive hit ratio	Overall hit ratio	Selected features
SVM	0.710	0.947	0.867	{HSV-1, HHV-8, CMV} or {HSV-1, HHV-8}
LDA	0.710	0.893	0.833	{HSV-1, HHV-8, EBV} or {HSV-1, HHV-8}

5. Discussion and conclusion

This paper has explored five DNA viruses – HSV-1, EBV, CMV, HPV, and HHV-8 – affecting a breast tumor diagnosed by using support vector machines. In order to find the correlation DNA viruses with breast tumor, and to achieve a high classificatory accuracy, *F*-score is adapted to find the important features, and the grid search approach is used to search the optimal SVM parameters. The results revealed that the SVM-based model has good performance in diagnosing breast cancer according to our data set.

The present study's results also show that the attributes {HSV-1, HHV-8} or {HSV-1, HHV-8, CMV} can achieve identical high accuracy, at 86% of average overall hit rate. Although these two models have an identical high accuracy, considering the diagnosis cost and accuracy, this study suggests simultaneously considering HSV-1 and HHV-8 is feasible; however, only considering HHV-8 or HSV-1 is less accurate.

From the SVM model and LDA, the authors found that the HSV-1 and HHV-8 are the common important features for breast tumor in distinguishing breast cancer and fibroadenoma. The development of an oncolytic viral therapy for breast cancer with an HSV-1 mutant, HF10 is by Teshigahara et al. (2004); the result also indicated that replication-competent HSV-1 mutants held significant potential as cancer therapeutic agents. Allan et al. (2001) detected two women who developed Kaposi's sarcoma in the lymphedematous arm many years after surgery for breast cancer. Kaposi's sarcoma-associated herpesvirus (KSHV, HHV-8) was suggested to be associated with breast cancer (Newton et al., 2003). Additionally, from the SVM model, CMV has also the important features for breast tumors. Richardson et al. (2004) investigated the association between EBV and CMV immunoglobulin G levels and the risk of breast cancer before age 40 in Australian breast cancer families. Their results are such that CMV is a risk factor for breast cancer. Furthermore, Richardson (1997) suggested that CMV is a risk factor for breast carcinomas. The antibody activity against CMV increased in several seropositive patients. None of these patients, however, developed signs of a CMV infection.

The LDA shows EBV is an important feature in breast tumors beside HSV-1 and HHV-8. The first report of the positive effect of EBV on breast cancer is Labrecque, Barnes, Fentiman, and Grifin (1995). Chu, Chen, and Chang (1998) studied the presence of EBV in breast cancer and suggested that it may not play a significant role in the etiology of breast cancers in Taiwan. Glaser, Ambinder, DiGiuseppe, Horn-Ross, and Hsu (1998) concluded that the EBV EBER-1 transcript is not commonly expressed in breast cancer, based on a broadly representative case series. Bonnet et al. (1999) investigated the presence of EBV in human breast cancers and indicated that EBV might be a cofactor in the development of some breast cancers. McCall et al. (2001) researched nine studies of EBV in breast cancer and found only one of 115 cases that tested positive for EBV.

Murray et al. (2003) concluded that EBV can be regularly detected in whole sections of breast cancers, but its viral copy number is very low. Based on SVM model and LDA, the neglect factor is HPV. This finding is the same as that of Klein and Klein (2005). Klein and Klein also showed that HPV has not been associated with breast cancer. This report concluded that products of the HPV genome induce immortalization of human breast epithelial cells and reduce their growth factor requirement.

The present study shows that the SVM-based classifier for fibroadenoma or breast cancer diagnosis classification model is satisfactory both in classificatory accuracy and in finding the important features to discriminate between fibroadenoma or breast cancer. The practical obstacle of the SVM-based (as well as neural networks) classification model is its black-box nature. A possible solution for this issue is the use of SVM rule extraction techniques or the use of hybrid-SVM model combined with other more interpretable models. These issues remain to be solved in future research.

References

Akbulut, H., Zhang, L., Tang, Y., & Deisseroth, A. (2003). Cytotoxic effect of replication-competent adenoviral vectors carrying L-plastin promoter regulated E1A and cytosine deaminase genes in cancers of the breast, ovary and colon. *Cancer Gene Therapy, 10*(5), 388–395.

Allan, A. E., Shoji, T., Li, N., Buralge, A., Davis, B., & Bhawan, J. (2001). Two cases of Kaposi's sarcoma mimicking Stewart–Treves syndrome found to be human herpesvirus-8. *American Journal of Dermatopathology, 23*(5), 431–436.

Andres, A. (2005). Cancer incidence after immunosuppressive treatment following kidney transplantation. *Critical Reviews in Oncology/Hematology, 56*(1), 71–85.

Baeyens, A., Thierens, H., Claes, K., Poppe, B., de Ridder, L., & Vral, A. (2004). Chromosomal radiosensitivity in BRCA1 and BRCA2 mutation carriers. *International Journal of Radiation Biology, 80*(10), 745–756.

Bonnet, M., Guinebretiere, J.-M., Kremmer, E., Grunewald, V., Benhamou, E., Contesso, G., et al. (1999). Detection of Epstein-Barr virus in invasive breast cancers. *Journal of National Cancer Institute, 91*(16), 1376–1381.

Cancer Registry Annual Report (2004). Executive Yuan (Taipei), Taiwan.

Chang, J.-G., Su, T.-H., Wei, H.-J., Wang, J.-C., Chen, Y.-J., Chang, C.-P., et al. (1999). Analysis of TSG101 tumour susceptibility gene transcripts in cervical and endometrial cancers. *British Journal of Cancer, 79*(3–4), 445–450.

Chen, Y.-W., & Lin, C.-J. (2005). Combining SVMs with various feature selection strategies. Available from http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf.

Cheng, S. H., Tsou, M. H., Liu, M. C., & Jian, J. J. (2000). Unique features of breast cancer in Taiwan. *Breast Cancer Research and Treatment, 63*(3), 213–220.

Chu, J.-S., Chen, C.-C., & Chang, K.-J. (1998). In situ detection of Epstein-Barr virus in breast cancer. *Cancer Letters, 124*(1), 53–57.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge: Cambridge University Press.

Dimmock, N. J., & Primrose, S. B. (1994). *Carcinogenesis and tumor viruses: Introduction to modern virology* (4th ed.). London: Blackwell Science Ltd..

Fina, F., Romain, S., Ouafik, L. H., Palmari, J., Ayed, F. B., Benharkat, S., et al. (2001). Frequency and genome load of Epstein-Barr virus in 509 breast cancers from different geographical areas. *British Journal of Cancer, 84*(6), 783–790.

Fröhlich, H., & Chapelle, O. (2003). Feature selection for support vector machines by means of genetic algorithms. In *Proceedings of the 15th IEEE international conference on tools with artificial intelligence, Sacramento, California, USA*, pp. 142–148.

Glaser, S. L., Ambinder, R. F., DiGiuseppe, J. A., Horn-Ross, P. L., & Hsu, J. L. (1998). Absence of Epstein-Barr virus EBER-1 transcripts in an epidemiologically diverse group of breast cancers. *International Journal of Cancer, 75*(4), 555–558.

Grinstein, S., Preciado, M. V., Gattuso, P., Chabay, P. A., Warren, W. H., De Matteo, E., et al. (2002). Demonstration of Epstein-Barr virus in carcinomas of various sites. *Cancer Research, 62*(17), 4876–4878.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification. Available from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Hsu, C. W., & Lin, C. J. (2002). A simple decomposition method for support vector machine. *Machine Learning, 46*(1–3), 219–314.

Hu, J., Hallden, G., Shorrock, C., Simpson, G., Coffin, R., Kamalati, T., et al. (2004). Combination of a second generation genetically modified herpes simplex virus type 1 (HSV-1) with paclitaxel in the treatment of breast cancer in vitro. *Molecular Therapy, 9*, S105.

Huang, J., Chen, H., Hutt-Fletcher, L., Ambinder, R. F., & Hayward, S. D. (2003). Lytic viral replication as a contributor to the detection of Epstein-Barr virus in breast cancer. *Journal of Virology, 77*(24), 13267–13274.

Joachims, T. (1998). Text categorization with support vector machines. In *Proceedings of European conference on machine learning (ECML), Chemintz, DE*, pp. 137–142.

Katano, M., Kubota, E., Nagumo, F., Matsuo, T., Hisatsugu, T., & Tadano, J. (1995). Inhibition of tumor cell growth by a human B-cell line. *Biotherapy, 8*(1), 1–6.

Kecman, V. (2001). *Learning and soft computing*. Cambridge, MA: The MIT Press.

Klein, G., & Klein, E. (2005). Surveillance against tumors—is it mainly immunological? *Immunology Letters, 100*(1), 29–33.

Labrecque, L. G., Barnes, D. M., Fentiman, I. S., & Grifin, B. E. (1995). Epstein-Barr virus in epithelial cell tumors: a breast cancer study. *Cancer Research, 55*(1), 39–45.

La Vecchia, C., Tavani, A., Franceschi, S., & Parazzini, F. (1996). Oral contraceptives and cancer. A review of the evidence. *Drug Safety, 14*(4), 260–272.

Lee, Y. K., & Lee, C. K. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics, 19*(9), 1132–1139.

Lee, Y.-J., Mangasarian, O. L., Wolberg, & W. H. (2000). In *Breast cancer survival and chemotherapy: a support vector machine analysis. DIMACS series in discrete mathematics and theoretical computer science, Vol. 55*, pp. 1–20.

Lee, S. W., Reimer, C. L., Oh, P., Campbell, D. B., & Schnitzer, J. E. (1998). Tumor cell growth inhibition by caveolin re-expression in human breast cancer cells. *Oncogene, 16*(11), 1391–1397.

Liu, Y., Klimberg, V. S., Andrews, N. R., Hicks, C. R., Peng, H., Chiriva-Internati, M., et al. (2001). Human papillomavirus DNA is present in a subset of unselected breast cancers. *Journal of Human Virology, 4*(6), 329–334.

Liu, H., Xu, Z.-L., Wang, Y., Yang, L., Feng, O., Li, Y., et al. (1993). Production of anti-tumor human monoclonal antibodies using different approaches. *Human Antibodies, 4*(1), 2–8.

Liu, H. X., Zhang, R. S., Luan, F., Yao, X. J., Liu, M. C., Hu, Z. D., et al. (2003). Diagnosing breast cancer based on support vector machines. *Journal of Chemical Information and Computer Sciences, 43*(3), 900–907.

Ma, N., Szmitko, P., Brade, A., Chu, I., Lo, A., Woodgett, J., et al. (2004). Kinase-dead PKB gene therapy combined with hyperthermia for human breast cancer. *Cancer Gene Therapy, 11*(1), 52–60.

McCall, S. A., Lichy, J. H., Bijwaard, K. E., Aguilers, N. S., Chu, W.-S., & Taubenberger, J. K. (2001). Epstein-Barr virus detection in ductal carcinoma of the breast. *Journal of National Cancer Institute, 93*(2), 148–150.

Murray, P. G., Lissauer, D., Junying, J., Davies, G., Moore, S., Bell, A., et al. (2003). Reactivity with a monoclonal antibody to Epstein-Barr virus (EBV) nuclear antigen 1 defines a subset of aggressive breast cancers in the absence of EBV genome. *Cancer Research, 63*(9), 2338–2343.

Newton, R., Ziegler, J., Bourboulia, D., Casabonne, D., Beral, V., Mbidde, E., et al. (2003). The sero-epidemiology of Kaposi's sarcoma-associated herpesvirus (KSHV/HHV-8) in adults with cancer in Uganda. *International Journal of Cancer, 103*, 226–232.

Overview of Public Health (1998). Department of Health, Executive Yuan (Taipei), Taiwan.

Pontil, M., & Verri, A. (1998). Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(6), 637–646.

Ribeiro-Silva, A., Ramalho, L. N. Z., Garcia, S. B., & Zucoloto, S. (2004). Does the correlation between EBN-1 and p63 expression in breast carcinomas provide a clue to tumorigenesis in Epstein-Barr virus-related breast malignancies? *Brazilian Journal of Medical and Biological Research, 37*(1), 89–95.

Richardson, A. (1997). Is breast cancer caused by late exposure to a common virus? *Medical Hypotheses, 48*, 491–497.

Richardson, A. K., Cox, B., McCredie, M. R. E., Dite, G. S., Chang, J.-H., Gertig, D. M., et al. (2004). Cytomegalovirus, Epstein-Barr virus and risk of breast cancer before age 40 years: a case-control study. *British Journal of Cancer, 90*(11), 2149–2152.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery, 1*, 317–327.

Schölkopf, B., & Smola, A. J. (2000). *Statistical learning and kernel methods*. Cambridge, MA: MIT Press.

Still, I. H., Hamilton, M., Vince, P., Wolfman, A., & Cowell, J. K. (1999). Cloning of TACC1, an embryonically expressed, potentially transforming coiled coil containing gene, from the 8p11 breast cancer amplicon. *Oncogene, 18*(27), 4032–4038.

Strender, L.-E., Blomgren, H., Petrini, B., Wasserman, J., Forsgren, M., Norberg, R., et al. (1981). Immunologic monitoring in breast cancer patients receiving postoperative adjuvant chemotherapy. *Cancer, 48*(9), 1996–2002.

Svane, I. M., Nikolajsen, K., Hansen, S. W., Kamby, C., Nielsen, D. L., & Johnsen, H. E. (2002). Impact of high-dose chemotherapy on antigen-specific T cell immunity in breast cancer patients. Application of new flow cytometric method. *Bone Marrow Transplantation, 29*(8), 659–666.

Teshigahara, O., Goshima, F., Takao, K., Kohno, S., Kimata, H., Nakao, A., et al. (2004). Oncolytic viral therapy for breast cancer with herpes simplex virus type 1 mutant HF 10. *Journal of Surgical Oncology, 85*(1), 42–47.

Tsai, J.-H., Tsai, C.-H., Chang, M.-H., Lin, S.-J., Xu, F.-L., & Yang, C.-C. (2005). Association of viral factors with non-familial breast cancer in Taiwan by comparison with non-cancerous, fibroadenoma, and thyroid tumor tissues. *Journal of Medical Virology, 75*, 276–281.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Wang, S., & Vos, J.-M. (1996). A hybrid herpesvirus infectious vector based on Epstein-Barr virus and herpes simplex virus type 1 for gene

transfer to human cells in vitro and in vivo. *Journal of Virology, 70*(12), 8422–8430.

Widschwendter, A., Brunhuber, T., Wiedemair, A., Mueller-Holzner, E., & Marth, C. (2004). Detection of human papillomavirus DNA in breast cancer of patients with cervical cancer history. *Journal of Clinical Virology, 31*(4), 292–297.

Wu, A., Mazumder, A., Martuza, R. L., Liu, X., Thein, M., Meehan, K. R., et al. (2001). Biological purging of breast cancer cells using an attenuated replication-competent herpes simplex virus in human hematopoietic stem cell transplantation. *Cancer Research, 61*(7), 3009–3015.

Xue, S. A., Lampert, I. A., Haldane, J. S., Bridger, J. E., & Griffin, B. E. (2003). Epstein-Barr virus gene expression in human breast cancer: protagonist or passenger? *British Journal of Cancer, 89*(1), 113–119.

Yip, Y. L., Hawkins, N. J., Clark, M. A., & Ward, R. L. (1997). Evaluation of different lymphoid tissue sources for the construction of human immunoglobulin gene libraries. *Immunotechnology, 3*(3), 195–203.

Yu, G. X., Ostrouchov, G., Geist, A., & Samatova, N. F. (2003). An SVM-based algorithm for identification of photosynthesis-specific genome features. In *2nd IEEE computer society bioinformatics conference, CA, USA*, pp. 235–243.

Yu, H., Rohan, T. E., Cook, M. G., Howe, G. R., & Miller, A. B. (1992). Risk factors for fibroadenoma: a case-control study in Australia. *American Journal of Epidemiology, 135*(3), 247–259.

Zhang, G. P. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, 30*(4), 451–462.

Zhu, Z. B., Makhija, S. K., Lu, B., Wang, M., Kaliberova, L., Liu, B., et al. (2004). Transcriptional targeting of adenoviral vector through the CXCR4 tumor-specific promoter. *Gene Therapy, 11*(7), 645–648.

Ziegler, R. G., Hoover, R. N., Pike, M. C., Hildesheim, A., Nomura, A. M., West, D. W., et al. (1993). Migration patterns and breast cancer risk in Asian-American women. *Journal of National Cancer Institute, 85*, 1819–1827.