ELSEVIER

Artificial Intelligence in Medicine 28 (2003) 281–306

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines

C. Lu[a], T. Van Gestel[a], J.A.K. Suykens[a], S. Van Huffel[a,*], I. Vergote[b], D. Timmerman[b]

[a]*Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*

[b]*Department of Obstetrics and Gynecology, University Hospitals KU Leuven, Herestraat 49, B-3000 Leuven, Belgium*

Received 25 March 2002; received in revised form 1 October 2002; accepted 19 February 2003

Abstract

In this work, we develop and evaluate several least squares support vector machine (LS-SVM) classifiers within the Bayesian evidence framework, in order to preoperatively predict malignancy of ovarian tumors. The analysis includes exploratory data analysis, optimal input variable selection, parameter estimation, and performance evaluation via receiver operating characteristic (ROC) curve analysis. LS-SVM models with linear and radial basis function (RBF) kernels, and logistic regression models have been built on 265 training data, and tested on 160 newly collected patient data. The LS-SVM model with nonlinear RBF kernel achieves the best performance, on the test set with the area under the ROC curve (AUC), sensitivity and specificity equal to 0.92, 81.5% and 84.0%, respectively. The best averaged performance over 30 runs of randomized cross-validation is also obtained by an LS-SVM RBF model, with AUC, sensitivity and specificity equal to 0.94, 90.0% and 80.6%, respectively. These results show that the LS-SVM models have the potential to obtain a reliable preoperative distinction between benign and malignant ovarian tumors, and to assist the clinicians for making a correct diagnosis.

Keywords: Ovarian tumor classification; Least squares support vector machines; Bayesian evidence framework; ROC analysis; Ultrasound; CA 125

* Corresponding author. Tel.: +32-16-321703; fax: +32-16-321970.
E-mail addresses: chuan.lu@esat.kuleuven.ac.be (C. Lu), tony.vangestel@esat.kuleuven.ac.be (T. Van Gestel), johan.suykens@esat.kuleuven.ac.be (J.A.K. Suykens), sabine.vanhuffel@esat.kuleuven.ac.be (S. Van Huffel).

0933-3657/$ – see front matter
doi:10.1016/S0933-3657(03)00051-4

1. Introduction

Ovarian masses are a very common problem in gynecology. Detection of ovarian malignancy at an early stage is very important for the survival of the patients. The 5-year survival rate for ovarian cancer when detecting at a late clinical stage is 35% [17]. In contrast, the 5-year survival for patients with stage I ovarian cancer is about 80% [29]. However, nowadays 75% of the cases are only diagnosed at an advanced stage, resulting into the highest mortality rate among gynecologic cancers. The treatment and management of different types of ovarian tumors differ greatly. Conservative management or less invasive surgery suffices for patients with a benign tumor; on the other hand, those with suspected malignancy should be timely referred to a gynecologic oncologist. An accurate diagnosis before operation is critical to obtain the most effective treatment and best advice, and will influence the outcome for the patient and the medical costs. Therefore, a reliable test for preoperative discrimination between benign and malignant ovarian tumors is of considerable help for clinicians in choosing the appropriate treatment for patients.

Several attempts have been made in order to automate the classification process. The risk of malignancy index (RMI) is a widely used score which combines the CA 125 values with the ultrasonographic morphologic findings and the menopausal status of the patient [10]. In a previous study, based on a smaller data set, several types of black-box models such as logistic regression models (LRs) and multi-layer perceptrons (MLPs) have been developed and tested [22,23], using the selected variables via the stepwise logistic regression. Both types of models have been shown to perform better than the RMI. A hybrid approach that integrates the Bayesian belief network (which represents the expert knowledge in the graphical model) into the learning of MLPs, has also been investigated in [2–4]. The integration of the white-box models (e.g. belief networks) with the black-box models (e.g. MLPs) leads to so-called grey-box models. This can be done for example by transformation of the belief network into an informative prior distribution for black-box models by using virtual prior samples. However, finding the structure and learning of the graphical model is not so easy and very time consuming. MLPs also suffer from the problem of multiple local minima. In this paper, we will focus on the development of black-box models, in particular least squares support vector machines (LS-SVMs), to preoperatively predict malignancy of ovarian tumors based on an enlarged data set, and validating the models for clinical purposes.

Support vector machines (SVMs) are extensively used for solving pattern recognition and nonlinear function estimation problems [28,6]. They map the input into a high-dimensional feature space, in which an optimal separating hyperplane can be constructed. The attractive features of these kernel-based algorithms include: good generalization performance, the existence of a unique solution, and strong theoretical background, i.e. statistical learning theory [28], supporting their good empirical results. In this paper, a least squares version of SVMs (LS-SVMs) [19,20] is considered, in which the training is expressed in terms of solving a set of linear equations in the dual space instead of quadratic programming as for the standard SVM case. To achieve a high level of performance with LS-SVM models, some parameters have to be tuned, including the regularization parameter and the kernel parameter corresponding to the kernel type. The Bayesian evidence framework proposed by MacKay provides a unified theoretical treatment of learning in order to cope with similar problems in neural networks [13]. Recently, the Bayesian

method has also been integrated into the LS-SVMs, and a numerical implementation was derived. This approach has been successfully applied to several benchmark problems [26] and to the prediction of financial time series [27]. Within this Bayesian evidence framework, we are able to perform parameter estimation, hyperparameter tuning, model comparison, input selection, and probabilistic interpretation of the output in a unified way.

The paper is organized as follows. In Section 2, the exploratory data analysis is described. In Section 3, the LS-SVMs and the Bayesian evidence framework are briefly reviewed; a design of a LS-SVM classifier within the evidence framework in combination with a sparse approximation process, and a forward input selection procedure are proposed. In Section 4, we demonstrate the application of LS-SVM to the prediction of malignancy of the ovarian tumors, including several practical issues during model development and evaluation; the performance of different models with different kernels are assessed via receiver operating characteristic (ROC) analysis. In Section 5, we will discuss several issues when using these models in clinical practice. Finally, conclusions are drawn and topics for future research are indicated.

2. Data

The data set includes the information of 525 consecutive patients who were referred to a single ultrasonographer at University Hospitals Leuven, Belgium, between 1994 and 1999. These patients have a persistent extrauterine pelvic mass, which was subsequently surgically removed. The study is designed mainly for preoperative differentiation between benign and malignant adnexal masses [22]. Patients without preoperative results of serum CA 125 levels have been excluded from this analysis, the number of which is $N_{\text{miss}} = 100$. Results of histological examination were considered as the gold standard for discrimination of the tumors. Among the available 425 cases, 291 patients (68.5%) had benign tumors, whereas 134 ones (31.5%) had malignant tumors.

The following measurements and observations were acquired before operation: the age and menopausal status of patients; serum CA 125 levels; the ultrasonographic morphologic findings, in particular locularility, papillation, solid areas, echogenic descriptions of the mass, the amount of ascites; color Doppler imaging and blood flow indexing, in particular, the resistance index, and color score (a subjective semi-quantitative assessment of the amount of blood flow). For a detailed explanation, the reader is referred to [22–25].

A rigorous approach to pattern recognition requires a good understanding of the data. Our exploratory data analysis aims to gain insights into the data and consists of the following steps.

2.1. Data preprocessing

The original data set contains 25 features. Feature histograms and boxplots have been used to identify outliers and quantization effects. Some feature values have been transformed prior to further analysis, in particular, CA 125 serum level was rescaled by taking its logarithm; the nominal scaled variable color score with values from 1 to 4 was recoded to three binary variables. Hence, we have in total 27 candidate input variables.

Table 1
Demographic, serum marker, color Doppler imaging and morphologic variables

	Variable (symbol)	Benign	Malignant
Demographic	Age (Age)	45.6 ± 15.2	56.9 ± 14.6
	Post-menopausal (Meno) (%)	31.0	66.0
Serum marker	CA 125 (log) (L_CA125)	3.0 ± 1.2	5.2 ± 1.5
CDI	Weak blood flow (Colsc2) (%)	41.2	14.2
	Normal blood flow (Colsc3) (%)	15.8	35.8
	Strong blood flow (Colsc4) (%)	4.5	20.3
	Pulsatility index (PI)	1.34 ± 0.94	0.96 ± 0.61
	Resistance index (RI)	0.64 ± 0.16	0.55 ± 0.17
	Peak systolic velocity (PSV)	19.8 ± 14.6	27.3 ± 16.6
	(*Time-averaged*) mean velocity (TAMX)	11.4 ± 9.7	17.4 ± 11.5
B-mode ultrasonography	Abdominal fluid (Asc) (%)	32.7	67.3
	Unilocular cyst (Un) (%)	45.8	5.0
	Unilocular solid (Unsol) (%)	6.5	15.6
	Multilocular cyst (Mul) (%)	28.7	5.7
	Multilocular solid (Mulsol) (%)	10.7	36.2
	Solid tumor (Sol) (%)	8.3	37.6
Morphologic	Bilateral mass (Bilat) (%)	13.3	39.1
	Smooth wall (Smooth) (%)	56.8	5.8
	Irregular wall (Irreg) (%)	33.8	73.2
	Papillations (Pap) (%)	13.0	53.2
	Septa > 3 mm (Sept) (%)	13.0	31.2
	Acoustic shadows (Shadows) (%)	12.2	5.7
Echogenicity	Anechoic cystic content (Lucent) (%)	43.2	29.1
	Low level echogenicity (Low_level) (%)	12.0	19.9
	Mixed echogenicity (Mixed) (%)	20.3	13.5
	Ground glass cyst (G_glass) (%)	19.8	8.5
	Hemorrhagic cyst (Haem) (%)	3.9	0.0

Note: For continuous variables, mean $\pm$ S.D. in case of benign and malignant, respectively are reported; for binary variables, the occurrences (%) of the corresponding features are reported.

2.2. Univariate analysis

Table 1 lists the 27 variables that were considered, together with their mean value and standard deviations or the occurrence in case of benign and malignant tumors, respectively.

2.3. Multivariate data analysis

To get a first idea of the important predictors, we performed a factor analysis using the technique of principal components factoring (PCF), which is essentially principal component analysis based on the correlation matrix, with the assumption that estimates of the communalities are one. Fig. 1 shows the biplot in a two-dimensional space generated by the first two principal components called PC1 and PC2. The biplot visualizes

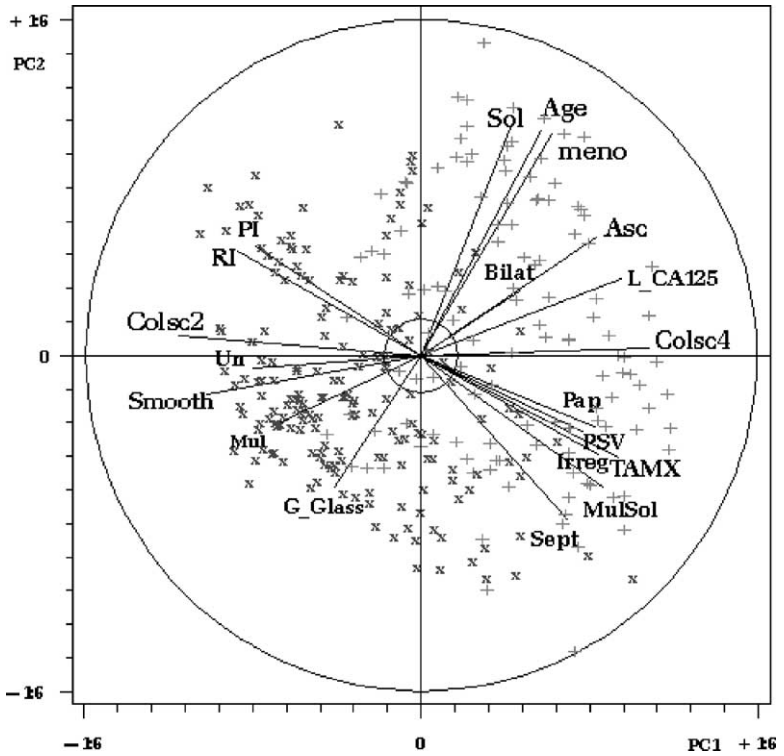

Fig. 1. Biplot of the ovarian tumor data. The observations are plotted as points (×: benign, +: malignant), the variables are plotted as vectors from the origin, i.e. taking the respective factor loadings as the coordinates.

the correlation between the variables, and the relations between the variables and classes. In particular, a small angle between two variables such as (Age, Meno) points out that those variables are highly correlated; the observations of malignant tumors (indicated by '+') have relatively high values for variables Sol, Age, Meno, Asc, L_CA125, Colsc4, Pap, Irreg, etc. but relatively low values for the variables Colsc2, Smooth, Un, Mul, etc. The biplot reveals that many variables are correlated, implying the need of variable selection. On the other hand, quite a lot of overlap between the two classes can also be observed, suggesting that the classical linear techniques might not be enough to capture the underlying structure of the data, and a nonlinear classifier might give better results than a linear classifier.

3. Least squares support vector machines and Bayesian evidence framework

MLPs have become very popular black-box classifiers, however they suffer from several drawbacks like non-convexity of the underlying optimization problem and difficulties in choosing the best number of hidden units. In support vector machines

[28], the learning problem is formulated and represented as a convex quadratic programming (QP) problem. The basic idea of the SVM classifier is the following: map an n-dimensional input vector $x \in \mathbb{R}^n$ into a high n_f-dimensional feature space by the mapping $\varphi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_f} : x \rightarrow \varphi(x)$. A linear classifier is then constructed in this feature space by minimizing an appropriate cost function. Using Mercer's theorem [14], the classifier is obtained by solving a finite-dimensional QP problem in the dual space avoiding explicit knowledge of the high-dimensional mapping and using only the related kernel function. In least squares support vector machines [19], one uses equality constraints instead of inequality constraints and a least squares error term in order to obtain a linear set of equations in the dual space.

However, to achieve a high level of performance, some parameters in the LS-SVM model must be tuned. These adjustable hyperparameters include: a regularization parameter, which determines the tradeoff between minimizing the training errors and minimizing the model complexity; and a kernel parameter such as the width of the RBF kernel. One popular way to choose the hyperparameters is cross-validation. Alternatively, one can utilize an upper bound on the generalization error resulting from Vapnik–Chervonenkis (VC) learning theory [28].

On the other hand, a similar problem of finding good hyperparameters in the training of feedforward neural networks, has been tackled by applying the Bayesian framework [5,15,13]. In comparison with the traditional approaches, the Bayesian methods provide a rigorous framework for the automatic adjustment of the regularization parameters to their near optimal values, without the need to set data aside in a validation set. Moreover, Bayesian techniques also provide assessments of the confidence associated with its prediction, which is essential for any biomedical pattern recognition system. In contrast to the maximum likelihood framework, which finds a set of parameters by minimizing an error function, the Bayesian approach handles uncertainty by integrating over all possible sets of parameters. Particularly the Bayesian evidence method performs integration using an approximate analytic solution.

In [26], the evidence framework has been applied to LS-SVMs for classification. Because of the least squares formulation of LS-SVMs, the derivation of analytic expressions on the different levels of inferences is possible. Relating a probabilistic framework to the LS-SVM formulation on the first level of Bayesian inference, the hyperparameters are inferred on the second level. Model comparison is performed on the third level in order to select the kernel parameters.

In the following subsections, we briefly review the use of LS-SVMs in binary classification problems, and how to apply the Bayesian framework to LS-SVM classifiers. For more mathematical details and other applications the interested reader may consult the book [20] and the papers [19,21,26,27]. Then we introduce an LS-SVM input variable selection scheme and sparse approximation procedures for LS-SVM classifiers within the evidence framework.

3.1. Probabilistic inferences in LS-SVM within the evidence framework

The LS-SVM classifier $y(x) = \text{sign}[w^{\mathrm{T}}\varphi(x) + b]$ is inferred from the data $D = \{(x_i, y_i)\}_{i=1}^{N}$ with binary targets $y_i = \pm 1$ (in this tumor classification problem, $+1$

corresponds to 'malignant' and -1 to 'benign'), by minimizing the following cost function:

$$\min_{w,b,e} \mathcal{J}_1(w,e) = \mu E_W + \zeta E_D = \frac{\mu}{2} w^{\mathrm{T}} w + \frac{\zeta}{2} \sum_{i=1}^{N} e_i^2 \tag{1}$$

subject to the equality constraints

$$e_i = 1 - y_i[w^{\mathrm{T}} \varphi(x_i) + b], \qquad i = 1, \ldots, N. \tag{2}$$

The regularization and sum of squares error term are defined as $E_W = (1/2) w^{\mathrm{T}} w$, and $E_D = (1/2) \sum_{i=1}^{N} e_i^2$, respectively. The tradeoff between the training error and regularization is determined by the ratio $\gamma = \zeta/\mu$.

One defines the Lagrangian

$$\mathcal{L}(w,b,e;\alpha) = \mathcal{J}_1 - \sum_{i=1}^{N} \alpha_i \{ y_i[w^{\mathrm{T}} \varphi(x_i) + b] - 1 + e_i \},$$

where α_i are Lagrange multipliers. The Kuhn–Tucker conditions for optimality $\partial \mathcal{L}/\partial w = 0$, $\partial \mathcal{L}/\partial b = 0$, $\partial \mathcal{L}/\partial e_i = 0$, $\partial \mathcal{L}/\partial \alpha_i = 0$ provide a set of linear equations $w = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i)$, $\sum_{i=1}^{N} \alpha_i y_i = 0$, $\alpha_i = \gamma e_i$, $y_i[w^{\mathrm{T}} \varphi(x_i) + b] - 1 + e_i = 0$, for $i = 1, \ldots, N$, respectively. Elimination of w and e gives

$$\left[\begin{array}{c|c} 0 & Y^{\mathrm{T}} \\ \hline Y & \Omega + \gamma^{-1} I_N \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 1_v \end{array}\right] \tag{3}$$

with $Y = [y_1 \cdots y_N]^{\mathrm{T}}$, $\alpha = [\alpha_1 \cdots \alpha_N]^{\mathrm{T}}$, $e = [e_1 \cdots e_N]^{\mathrm{T}}$, $1_v = [1 \cdots 1]^{\mathrm{T}}$, and I_N the $N \times N$ identity matrix. Mercer's theorem is applied to the matrix Ω with $\Omega_{ij} = y_i y_j \varphi(x_i)^{\mathrm{T}} \varphi(x_j) = y_i y_j K(x_i, x_j)$, where $K(\cdot,\cdot)$ is a chosen positive definite kernel that satisfies Mercer condition [14]. The most common kernels include a linear kernel $K(x_i, x_j) = x_i^{\mathrm{T}} x_j$ and an RBF kernel $K(x_i, x_j) = \exp(- \| x_i - x_j \|_2^2 / \sigma^2)$. The LS-SVM classifier is then constructed in the dual space as:

$$y(x) = \operatorname{sign}\left[\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b \right]. \tag{4}$$

It is interesting to notice here that the least squares formulation is related to kernel Fisher discriminant analysis (FDA). In addition an LS-SVM with linear kernel corresponds to linear Fisher discriminant analysis with regularization term [26].

3.1.1. Inference of model parameters (level 1)

The parameters w and bias term b for given value of μ, ζ are inferred from the data D at the first level. By applying Bayes' rule, a probabilistic interpretation for (1) and (2) is obtained:

$$p(w,b|D, \log\mu, \log\zeta, \mathcal{H}) = \frac{p(D|w,b,\log\mu,\log\zeta,\mathcal{H}) p(w,b|\log\mu,\log\zeta,\mathcal{H})}{p(D|\log\mu,\log\zeta,\mathcal{H})}, \tag{5}$$

where the model $\mathscr{H}$ corresponds to the kernel function K with different kernel parameters such as the width of an RBF kernel σ. The *evidence* $p(D|\log\mu, \log\zeta, \mathscr{H})$ is a normalizing constant and will be needed in the next level of inference.

The LS-SVM learning process can be given the following probabilistic interpretation. The error function is interpreted as the negative log likelihood for a noise model: $p(D|w, b, \log\zeta, \mathscr{H}) \propto \exp(-\zeta E_D)$. Thus the use of the sum of squares error E_D corresponds to an assumption of Gaussian noise on the target variable, and the parameter ζ defines a noise level (variance) $1/\zeta$.

We assume a separable Gaussian prior on the parameters w, with variance $1/\mu$, $p(w|\log\mu, \mathscr{H}) = (\mu/2\pi)^{n_f/2}\exp(-\mu/2w^{\mathrm{T}}w)$, and a Gaussian prior for b with variance $\sigma_b^2 \to \infty$ to approximate a uniform distribution. Thus the regularization term E_W is interpreted in terms of a log prior probability distribution over the parameters w and b : $p(w, b|\log\mu, \log\zeta, \mathscr{H}) = p(w|\log\mu, \mathscr{H})p(b|\log\sigma_b, \mathscr{H}) \propto \exp(-\mu E_W)$.

Hence the expression for the first level of inference becomes

$$p(w, b|D, \log\mu, \log\zeta, \mathscr{H}) \propto \exp(-\mu E_W)\exp(-\zeta E_D) = \exp(-\mathscr{J}_1(w, b)). \tag{6}$$

The maximum a posteriori estimates w_{MP} and b_{MP} are then obtained by minimizing the negative logarithm of (1), i.e. solving the linear set of equations in (3).

3.1.2. Class probabilities for the LS-SVM classifiers (level 1)

Given the posterior probability of the model parameters w and b we will now integrate over all w and b values so as to obtain the posterior probability $p(y|x, D, \log\mu, \log\zeta, \mathscr{H})$.

In the evidence framework, we assume that the posterior distribution of w can be approximated by a single Gaussian at w_{MP}. We define two error variables corresponding to different classes (indicated by subscripts '+' and '−') as $e_\pm = w^{\mathrm{T}}(\varphi(x) - \hat{m}_\pm)$, where $\hat{m}_+$ and $\hat{m}_-$ are the centers of the positive and negative class, respectively. After marginalizing over w the distribution of $e_\pm$ will also be Gaussian, centering around mean $m_{e_\pm}$ with variance $(\zeta_\pm^{-1} + \sigma_{e\pm}^2)$. The expression for the mean is

$$m_{e_\pm} = w_{\mathrm{MP}}^{\mathrm{T}}(\varphi(x) - \hat{m}_\pm) = \sum_{i=1}^{N}\alpha_i y_i K(x, x_i) - \frac{1}{N_\pm}\sum_{i=1}^{N}\alpha_i y_i \sum_{j\in\mathscr{I}_\pm} K(x_i, x_j), \tag{7}$$

where $\mathscr{I}_+$ and $\mathscr{I}_-$ indicate the sets of indices whose corresponding data points have positive and negative labels, respectively. The computation of the variance from the target noise $\zeta_\pm^{-1}$ will be discussed in the next section. While the corresponding expression of the additional variance due to the uncertainty in the parameters w is

$$\sigma_{e_\pm}^2 = [\varphi(x) - \hat{m}_\pm]^{\mathrm{T}} Q_{11}[\varphi(x) - \hat{m}_\pm], \tag{8}$$

where Q_{11} is the upper left $n_f \times n_f$ block of the covariance matric $Q = \mathrm{covar}([w; b], [w; b])$, which is related to the Hessian H of the LS-SVM cost function $\mathscr{J}_1(w, b)$,

$$Q = H^{-1} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \dfrac{\partial^2 J_1}{\partial w^2} & \dfrac{\partial^2 J_1}{\partial w \partial b} \\ \dfrac{\partial^2 J_1}{\partial b \partial w} & \dfrac{\partial^2 J_1}{\partial b^2} \end{bmatrix}^{-1}. \tag{9}$$

And the variance will be finally computed in the dual space. Let $\theta(x) = [K(x,x_1)\cdots K(x,x_N)]^{\mathrm{T}}$, Ψ be the $N \times N$ kernel matrix with element $\Psi_{ij} = K(x_i,x_j)$, and a centering matrix $M = (I_N - (1/N)(1_v 1_v^{\mathrm{T}}))$. Further define $1_+, 1_- \in \mathbb{R}^N$ as the vector with element zero or one, for $i = 1,\ldots,N$, $1_{\pm,i} = 1$ if $y_i = \pm 1$, otherwise $1_{\pm,i} = 0$. By using matrix algebra and applying the Mercer condition, we obtain:

$$\sigma^2_{e\pm} = \frac{1}{\mu}K(x,x) - \frac{2}{\mu N_\pm}\sum_{i\in\mathscr{I}_\pm} K(x,x_i) + \frac{1}{\mu N_\pm^2}\sum_{i,j\in\mathscr{I}_\pm} K(x_i,x_j) - \frac{\zeta}{\mu}(\theta^{\mathrm{T}}(x) - \frac{1}{N_\pm}1_\pm^{\mathrm{T}}\Psi)M(\mu I_N + \zeta M\Psi M)^{-1}M(\theta(x) - \frac{1}{N_\pm}\Psi 1_\pm). \tag{10}$$

Thus the conditional probabilities can be computed as:

$$p(x|y=\pm 1, D, \log\mu, \log\zeta, \log\zeta_\pm, \mathscr{H}) = (2\pi(\zeta_\pm^{-1} + \sigma^2_{e_\pm}))^{-1/2}\exp\left(-\frac{m^2_{e_\pm}}{2(\zeta_\pm^{-1} + \sigma^2_{e_\pm})}\right). \tag{11}$$

By applying Bayes' rule the following posterior class probabilities of the LS-SVM classifier are obtained (for notational simplicity, $\log\mu, \log\zeta, \log\zeta_\pm, \mathscr{H}$ are dropped in this expression):

$$p(y|x,D) = \frac{p(y)p(x|y,D)}{p(y=1)P(x|y=1,D) + P(y=-1)p(x|y=-1,D)}, \tag{12}$$

where $p(y)$ corresponds to the prior class probability. The posterior probability could also be used to make minimum risk decisions in case of different error costs. Let c_-^+ and c_+^- denote the cost of misclassifying a case from class '$-$' and '$+$', respectively. One trick to combine the posterior probability with the different error costs is by replacing $p(y)$ in (12) with the adjusted class prior:

$$P'(y=1) = \frac{P(y=1)c_+^-}{P(y=1)c_+^- + P(y=-1)c_-^+},$$

and

$$P'(y=-1) = \frac{P(y=-1)c_-^+}{P(y=1)c_+^- + P(y=-1)c_-^+}.$$

3.1.3. Inference of hyperparameters (level 2)

The second level of inference via Bayes' rule is the following:

$$p(\log\mu, \log\zeta|D,\mathscr{H}) = \frac{p(D|\log\mu,\log\zeta,\mathscr{H})p(\log\mu,\log\zeta|\mathscr{H})}{p(D|\mathscr{H})} \propto p(D|\log\mu,\log\zeta,\mathscr{H}), \tag{13}$$

where a uniform distribution is assumed in $\log\mu$ and $\log\zeta$ for the prior $p(\log\mu,\log\zeta|\mathscr{H}) = p(\log\mu|\mathscr{H})p(\log\zeta|\mathscr{H})$. The probability $p(D|\log\mu,\log\zeta,\mathscr{H})$ is equal to the evidence of the

previous level. Using Gaussian density at the maximum a posteriori estimates $w_{\mathrm{MP}}, b_{\mathrm{MP}}$, we obtain:

$$p(\log\mu, \log\zeta|D, \mathscr{H}) \propto \frac{\sqrt{\mu^{n_f}\zeta^N}}{\sqrt{\det H}} \exp(-\mathscr{J}_1(w_{\mathrm{MP}}, b_{\mathrm{MP}})), \tag{14}$$

with the Hessian H defined in (9). The expression for $\det H$ is given by $N\mu^{n_f - N_{\mathrm{eff}}}\zeta \prod_{i=1}^{N_{\mathrm{eff}}}(\mu + \zeta\lambda_{G,i})$, where N_{eff} eigenvalues $\lambda_{G,i}$ are the non-zero eigenvalues of the centered kernel matrix in the feature space and are the solution of the eigenvalue problem

$$(M\Psi M)v_{G,i} = \lambda_{G,i}v_{G,i}, \quad i = 1, \ldots, N_{\mathrm{eff}} \leq N - 1. \tag{15}$$

The effective number of parameters [5,12] for LS-SVM, is equal to:

$$\gamma_{\mathrm{eff}} = 1 + \sum_{i=1}^{N_{\mathrm{eff}}} \frac{\zeta_{\mathrm{MP}}\lambda_{G,i}}{\mu_{\mathrm{MP}} + \zeta_{\mathrm{MP}}\lambda_{G,i}} = 1 + \sum_{i=1}^{N_{\mathrm{eff}}} \frac{\gamma_{\mathrm{MP}}\lambda_{G,i}}{1 + \gamma_{\mathrm{MP}}\lambda_{G,i}}, \tag{16}$$

where the first term is due to the fact that no regularization is applied on the bias term b of the LS-SVM model. Since $N_{\mathrm{eff}} \leq N - 1$, the estimated number of effective parameters cannot exceed the number of data points N.

In the optimum of the level 2 cost function, the following relations can be obtained: $2\mu_{\mathrm{MP}}E_W(w_{\mathrm{MP}}) = \gamma_{\mathrm{eff}} - 1$ and $2\zeta_{\mathrm{MP}}E_D(w_{\mathrm{MP}}, b_{\mathrm{MP}}) = N - \gamma_{\mathrm{eff}}$. The last equality can be viewed as the Bayesian estimate of the variance $\zeta_{\mathrm{MP}}^{-1} = \sum_{i=1}^{N} e_i^2/(N - \gamma_{\mathrm{eff}})$ of the noise e_i. However in this paper, when computing the posterior of the class probability, the variances of the noise with different classes may differ, and are approximated in this way:

$$\zeta_{\pm}^{-1} = \frac{\sum_{j\in\mathscr{I}_{\pm}} e_{\pm,j}^2}{N_{\pm} - \gamma_{\mathrm{eff}}(N_{\pm}/N)}. \tag{17}$$

In practice, one can reformulate the optimization problem in μ and ζ into a scalar optimization problem in $\gamma = \zeta/\mu$:

$$\min_{\gamma} \mathscr{J}_2(\gamma) = \sum_{i=1}^{N-1} \log\left[\lambda_{G,i} + \frac{1}{\gamma}\right] + (N-1)\log[E_W(w_{\mathrm{MP}}) + \gamma E_D(w_{\mathrm{MP}}, b_{\mathrm{MP}})], \tag{18}$$

with $\lambda_{G,i} = 0$ for $i > N_{\mathrm{eff}}$. The expressions for E_D and E_W can be given in the dual variables using the relation $\alpha_i = \gamma_i e_i$: $E_D = (1/2\gamma^2)\sum_{i=1}^{N}\alpha_i^2, E_W = (1/2)\sum_{i=1}^{N}\alpha_i(y_i - (\alpha_i/\gamma) - b_{\mathrm{MP}})$.

This optimal hyperparameter γ is then obtained by solving the optimization problem (18) with gradients. Given the optimal γ_{MP}, one can easily compute μ_{MP} and ζ_{MP} using their relations in the optimum.

3.1.4. Bayesian model comparison (level 3)

After determining the hyperparameter μ_{MP} and ζ_{MP} on the second level of inference, we still have to select a suitable model $\mathscr{H}_j$. The prior $p(\mathscr{H}_j)$ over all possible models is assumed to be uniform. Thus the posterior for the model $\mathscr{H}_j$ is in the form

of $p(\mathscr{H}_j|D) \propto p(D|\mathscr{H}_j)p(\mathscr{H}_j) \propto p(D|\mathscr{H}_j)$. At this level, no evidence or normalizing constant is used since it is infeasible to compare all possible models $\mathscr{H}_j$.

A separable Gaussian prior for $p(\log\mu_{\mathrm{MP}}, \log\zeta_{\mathrm{MP}}|\mathscr{H}_j)$ is assumed for all models $\mathscr{H}_j$, with the constant standard deviations $\sigma_{\log\mu}$ and $\sigma_{\log\zeta}$. These prior widths of the hyperparameters are generally assumed to be broad and they cancel out when alternative models are compared. Also we assume that $p(\log\mu, \log\zeta|D, \mathscr{H}_j)$ can be well approximated by using a separable Gaussian with error bars $\sigma_{\log\mu|D}$ and $\sigma_{\log\zeta|D}$. The posterior likelihood $p(D|\mathscr{H}_j)$ corresponds to the evidence at the previous level and can be evaluated by:

$$p(D|\mathscr{H}_j) \propto p(D|\log\mu_{\mathrm{MP}}, \log\zeta_{\mathrm{MP}}, \mathscr{H}_j)\frac{\sigma_{\log\mu|D}\sigma_{\log\zeta|D}}{\sigma_{\log\mu}\sigma_{\log\zeta}}. \tag{19}$$

The models can thus be ranked according to the evidence $p(D|\mathscr{H}_j)$, that is the tradeoff between the goodness of fit from the previous level $p(D|\log\mu_{\mathrm{MP}}, \log\zeta_{\mathrm{MP}}, \mathscr{H}_j)$ and the Occam factor $\sigma_{\log\mu|D}\sigma_{\log\zeta|D}/\sigma_{\log\mu}\sigma_{\log\zeta}$ [12].

The error bars of $p(D|\log\mu_{\mathrm{MP}}, \log\zeta_{\mathrm{MP}}, \mathscr{H}_j)$ can be approximated by $\sigma^2_{\log\mu|D} \simeq (2/(\gamma_{\mathrm{eff}} - 1))$ and $\sigma^2_{\log\zeta|D} \simeq (2/(N - \gamma_{\mathrm{eff}}))$. And the expression for the evidence in the dual space is the following:

$$p(D|\mathscr{H}_j) \propto \sqrt{\frac{\mu_{\mathrm{MP}}^{N_{\mathrm{eff}}}\zeta_{\mathrm{MP}}^{N-1}}{(\gamma_{\mathrm{eff}} - 1)(N - \gamma_{\mathrm{eff}})\prod_{i=1}^{N_{\mathrm{eff}}}(\mu_{\mathrm{MP}} + \zeta_{\mathrm{MP}}\lambda_{G,i})}}. \tag{20}$$

One selects the kernel parameters, e.g. the width of an RBF kernel, with maximal posterior $p(D|\mathscr{H}_j)$.

3.2. Design of the LS-SVM classifier in a Bayesian evidence framework

Before building an LS-SVM classifier, it is better to normalize componentwise the training inputs to zero mean and unit variance [5]. We denote the normalized training data as $D = \{(x_i, y_i)\}_{i=1}^N$, with x_i the normalized inputs and $y_i \in \{-1, 1\}$ the corresponding class label. The new inputs collected in the test set and for evaluating the trained model will also be normalized in the same way as the training data, i.e. using the mean and variance estimates from the training data. Now, we start the design of the LS-SVM classifier in a Bayesian framework. Several procedures including hyperparameter tuning, input variable selection and sparse approximation are to be established.

3.2.1. Hyperparameter tuning

Select the model $\mathscr{H}_j$ by choosing a kernel type K_j with possible kernel parameters, e.g. the width of an RBF kernel σ_j. Infer the optimal γ_{MP}, μ_{MP} and ζ_{MP} on level 2 inference and evaluate the model evidence as follows:

1. Solve the eigenvalue problem (15).
2. Solve the scalar optimization problem (18) in $\gamma = \mu/\zeta$ using, e.g. a quasi-Newton method.
3. Given the optimal γ_{MP}, compute μ_{MP} and ζ_{MP} and γ_{eff}.
4. Calculate $p(D|\mathscr{H}_j)$ from (20) at the third level.

For a kernel K_j with tuning parameters, refine the tuning parameters, such that a higher model evidence $p(D|\mathscr{H}_j)$ is obtained. For example, for an RBF kernel, the parameter σ is inferred on the third level.

3.2.2. Input variable selection

In the Bayesian framework, given the likelihoods of the models $\mathscr{H}_0$ and $\mathscr{H}_1$, two models can be compared by the ratio of posterior probabilities: $p(D|\mathscr{H}_1)p(\mathscr{H}_1)/(p(D|\mathscr{H}_0)p(\mathscr{H}_0)) = p(\mathscr{H}_1)/p(\mathscr{H}_0)B_{10}$, where $B_{10} = p(D|\mathscr{H}_1)/p(D|\mathscr{H}_0)$ is the Bayes factor for model $\mathscr{H}_1$ against $\mathscr{H}_0$ from data D. If equal priors are assigned to the models, the posterior odds ratio then equals the Bayes factor, which can be seen as a measure of the evidence given by the data in favor of a model compared to a competing one. When the Bayes factor is greater than 1, the data favor $\mathscr{H}_1$ over $\mathscr{H}_0$; otherwise, the reverse is true. The rules of thumb for interpreting $2\log B_{10}$ include: the evidence for $\mathscr{H}_1$ is very weak if $0 \leq 2\log B_{10} \leq 2.2$, and the evidence for $\mathscr{H}_1$ is decisive if $2\log B_{10} > 10$, etc [9].

In the context of the Bayesian evidence framework, the evidence of the model $p(D|\mathscr{H}_j)$ is computed with (20) on level 3 inference. A higher $p(D|\mathscr{H}_1)$ compared to $p(D|\mathscr{H}_0)$ means the data favor $\mathscr{H}_1$ to $\mathscr{H}_0$. Therefore, given a certain type of kernel for the model, we propose to select the input variables according to the model evidence $p(D|\mathscr{H}_j)$.

The procedure performs a forward selection (greedy search), starting from zero variables, and choosing each time the variable which gives the greatest increase in the current model evidence. The selection is stopped when the addition of any remaining variable no longer increases the model evidence.

3.2.3. Sparse approximation

Due to the choice of the two-norm in the cost function, the sparseness is lost compared with the standard QP type SVMs. However, as has been shown in [21], the sparseness can be imposed to LS-SVMs by a pruning procedure based upon the sorted support value spectrum $|\alpha_i|$. Inspired by the SVM solution whose support vectors are near the decision boundary, we propose here to prune the data points which have negative support values. This is quite intuitive, since in LS-SVMs, $\alpha_i = \gamma e_i$. Negative support value α_i indicate that the data (x_i, y_i) are easy cases. The pruning of easy examples will focus the model more on the harder cases which lie around the decision boundary.

1. $D_{\text{cur}} = D = \{(x_i, y_i)\}_{i=1}^{N}$.
2. Based on D_{cur}, select the regularization parameter γ and possibly a kernel parameter σ within the Bayesian evidence framework. Train the LS-SVM (compute α) on the data D_{cur}, using current γ and σ.
3. If all the support values are positive, then go to 6.
4. Repeat pruning all the data points with non-positive support values, $D_{\text{cur}} \Leftarrow D_{\text{cur}} \setminus \{(x_d, y_d)|\alpha_d \leq 0\}$. Based on the reduced data set D_{cur}, recompute α using the same γ and σ, until all α values on the reduced data set D_{cur} are positive.
5. Go to 2.
6. Stop pruning, return the current α value and set the support values for the pruned data to zero.

3.2.4. Probabilistic interpretation of the output

The designed LS-SVM classifier $\mathscr{H}_j$ can be used to calculate class probabilities in the following steps:

1. Given the parameters $\alpha, b_{\mathrm{MP}}, \mu_{\mathrm{MP}}, \zeta_{\mathrm{MP}}, \gamma_{\mathrm{MP}}, \gamma_{\mathrm{eff}}$ and the eigenvalues and eigenvectors in (15) available from the designed model $\mathscr{H}_j$, calculate $m_{e+}, m_{e-}, \sigma^2_{e+}$ and σ^2_{e-} from (7) and (10), respectively. Compute ζ_+ and ζ_- from (17).
2. Calculate $p(x|y = \pm 1, D, \log\mu, \log\zeta, \log\zeta_\pm, \mathscr{H}_j)$ from (11).
3. Calculate $p(y|x, D, \mathscr{H}_j)$ from (12) by using the prior class probabilities or adjusted priors $P'(y = +1)$ and $P'(y = -1)$, respectively.

4. Application of LS-SVMs to the prediction of malignancy of ovarian tumors

Now we apply the LS-SVMs within the evidence framework to predict malignancy of ovarian tumors. The performance is assessed by receiver operator characteristic curve analysis. The area under the ROC curve (AUC) is computed. Furthermore, by setting various cutoff levels to the output probability, we will derive the sensitivity (true positive rate) and specificity (true negative rate) on the test set. All the experiments are conducted in Matlab.

4.1. Training and test set

First, we try to evaluate the generalization ability of the model, independently of the training data and model fitting process. The data set is split according to the time scale. The data from the first treated 265 patients (collected from 1994 to 1997) are taken as training set. The remaining 160 patient data (collected from 1997 to 1999) are used as test set. The proportion of malignant tumors in the training set and test set are both about $1/3$. Thanks to the Bayesian methods implemented here, no validation set is needed during training; otherwise, the validation during training would make inefficient use of the data set which is already moderately small in the case at hand [13]. The following procedures including input variable selection and model fitting, are independent from the test set.

However, the estimate from such a single hold-out cross-validation, in which the data set is partitioned into just two mutually exclusive subsets, is somehow biased, and depends on the division of the training set and test set. In order to get an estimate with lower bias, and also with potentially better predictive power of our method, we conduct another experiment. The data set is split randomly into two sets, the training set still containing 265 data, and test set 160 data. The sets are stratified, which means that the proportion of the malignant cases in each data set are kept around one third in all the training and test sets. We repeat this hold-out cross-validation 30 times, and the performance of the method is estimated by averaging.

The training and test set splitting issue related to the clinical practice will be further discussed in Section 5.

4.2. Input variable selection

The data set originally contains 27 input variables, some of which are rather relevant, others are only weakly relevant. Selecting the most predictive input variables is critical

to effective model development. A good subset selection of explanatory variables can substantially improve the performance of a classifier. The challenge is finding ways to pick the best subsets of variables.

A variety of techniques have been suggested for variable selection. One of the common approaches is stepwise logistic regression. This approach, with similarities to other correlation-based techniques, encounters problems if the input variables are not independent. Moreover, it is based on linear regression.

Here within this evidence framework, we will adapt the forward selection procedure as introduced in Section 3.2. We select a subset of variables that maximizes the evidence of the LS-SVM classifiers with either linear or RBF kernels. In order to stabilize the selection and computation of the evidence itself, we first compute the evidence of all univariate models each of which contains one single variable, and remove the three input variables which have the smallest evidence. A too small evidence points out that the corresponding variable contributes little to the prediction of malignancy of the ovarian tumors. This has also been verified by their negligible association with the class labels. Then we start the forward selection based on the remaining 24 candidate variables. The 10 selected variables based on an RBF kernel are listed in order of selection: L_CA125, Pap, Sol, Colsc3, Bilat, Meno, Asc, Shadows, Colsc4, Irreg, and will be denoted as MODEL1. The 11 selected variables based on a linear kernel, denoted as MODEL0, are also listed in order of selection: L_CA125, Pap, Sol, Colsc4, Unsol, Colsc3, Bilat, Shadows, Asc, Smooth, Meno. Though the two subsets of variables have nine variables in common, the evidence of the model selected based on the RBF kernel is higher than the one based on the linear kernel. The Bayes factor for MODEL1 (with the RBF kernel) against MODEL0 (with the linear kernel) B_{10} is greater than 1, and $2\log B_{10} = 74$ is greater than 10, indicating a strong evidence against MODEL0 in favor of MODEL1. Therefore, MODEL1 is used here for model building instead of the other.

In previous work [11], stepwise logistic regression was used to select the input variables. Eight variables were selected, which is just a reduced set of MODEL1 by removing variables 'Bilat' and 'Shadows'. However, this smaller subset was chosen based on the whole data set, and therefore validation on the test set might be over optimized. Here, for comparison reasons, we will also show the experimental results using this subset of variables, which is denoted by MODEL2: L_CA125, Asc, Pap, Meno, Colsc3, Colsc4, Sol, Irreg.

4.3. Model fitting and prediction

The model fitting procedure has two stages, the first is the construction of an LS-SVM classifier using the sparse approximation procedure explained in Section 3.2. The output of the LS-SVM classifier at this stage is a continuous number, which could be positive or negative and is located around $+1$ or -1. Remember that our training set ($N_{\text{train}} = 265$) is only moderately sized, thus the main goal of sparse approximation here is not to reduce computation time for training or prediction, but to improve the generalization ability. At the second stage, we will compute the output probability, indicating the posterior probability for a tumor to be malignant. Although some training data might be pruned during the first stage, the class mean and the posterior probabilities for the new data will be computed using all the training data.

In risk minimization decision making, different error costs are considered in order to reduce the expected loss. In this classification problem, misclassification of a malignant tumor is very serious, thus we aim at selecting a model with a high sensitivity while maintaining a high specificity (a low false positive rate). As the classifier will tend to predict the cases to the prevalent class, we need to correct for this tendency in order to increase the sensitivity of the classification by providing a higher adjusted prior for the malignant class. In the following experiments, the adjusted prior for the malignant class is intuitively set to $2/3$ and the benign class to $1/3$. When making a decision, one can take a certain probability cutoff value for the target environment. For example, setting a decision level at 0.5, means that all cases with a probability of malignancy greater than 0.5 are considered to be malignant, otherwise, they are classified as benign.

4.4. Model evaluation

The most commonly used performance measure of a classifier or a model is the classification accuracy, or the rate of correct classification, within the assumptions of equal misclassification costs and constant class distribution in the target environment. Both assumptions are not satisfied in real world problems [18]. Unlike classification accuracy, ROC is independent of class distributions or error costs and has been widely used in the biomedical field. Let us give a brief description about the ROC curves.

Assume a dichotomic classifier $y(x)$, which is the output value of the classifier given input x. Then the ultimate decision is taken by comparing the output $y(x)$ with a certain cutoff value. The *sensitivity* of a classifier is then defined as the proportion of malignant cases that are predicted to be malignant, and *specificity* as the proportion of benign cases that are predicted to be benign. The false positive rate is 1-specificity. When varying the cutoff value, i.e. the decision level, the sensitivity and specificity will change. An ROC curve is constructed by plotting the sensitivity versus 1-specificity, for varying cutoff values. The area under the ROC curve can be statistically interpreted as the probability of the classifier to correctly classify malignant cases and benign cases. The higher the AUC, the better the test. In this study, the AUC is obtained by a nonparametric method based on the Wilcoxon statistic, using the trapezoidal rule, to approximate the area [8]. The method proposed by DeLong et al. [7] will be used to compute the variance and covariance of the nonparametric AUC estimates derived from the same set of cases. AUC can be used for comparing two different ROC curves.

In contrast to the other measures such as the sensitivity and specificity, which require the setting of appropriate cutoff values for classification, the AUC is a one-value measure of the accuracy of a test. Hence here the statistical tests for comparing the performance of different models will be based on the AUC (see Sections 4.5 and 4.6).

4.5. Results from temporal validation

In the first experiment, the data set is split according to the time scale, and the performance of the model will be evaluated in the subsequent patients within the same center. Hence, we call this validation of our models *temporal validation* [1].

Table 2
Comparison of the temporal validation performance on the test set ($N_{train} = 265, N_{test} = 160$)

Model type (N_{SV})	AUC (±S.E.)	Decision level	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
RMI	0.8733 (±0.0298)	100	78.13	74.07	80.19	65.57	85.86
		75	**76.88**	**81.48**	**74.53**	**61.97**	**88.76**
LR1	0.9111 (±0.0246)	0.5	81.25	74.07	84.91	71.43	86.54
		0.4	80.63	75.96	83.02	69.49	87.13
		0.3	80.63	77.78	82.08	68.85	87.88
		0.2	**80.63**	**81.48**	**80.19**	**67.69**	**89.47**
LS-SVM1$_{Lin}$ (118)	0.9141 (±0.0236)	0.5	82.50	77.78	84.91	72.41	88.24
		0.4	81.25	77.78	83.02	70.00	88.00
		0.3	**81.88**	**83.33**	**81.13**	**69.23**	**90.53**
LS-SVM1$_{RBF}$ (97)	0.9184 (±0.0225)	0.5	84.38	77.78	87.74	76.36	88.57
		0.4	83.13	81.48	83.96	72.13	89.90
		0.3	**84.38**	**85.19**	**83.96**	**73.02**	**91.75**
LR2	0.9161 (±0.0218)	0.5	79.37	75.93	81.13	67.21	86.87
		0.4	77.50	75.93	78.30	64.06	86.46
		0.3	**78.75**	**81.48**	**77.36**	**64.71**	**89.13**
		0.2	78.13	85.19	74.53	63.01	90.80
LS-SVM2$_{Lin}$ (115)	0.9195 (±0.0215)	0.5	81.25	77.78	83.02	70.00	88.00
		0.4	80.63	79.63	81.13	68.25	88.66
		0.3	**80.00**	**85.19**	**77.36**	**65.71**	**91.11**
LS-SVM2$_{RBF}$ (99)	0.9223 (±0.0213)	0.5	83.75	81.48	83.96	73.33	90.00
		0.4	82.5	83.33	82.08	70.31	90.63
		0.3	**80.00**	**85.19**	**77.36**	**65.71**	**91.11**

Note: The 'best' results of each model obtained at a certain decision level are indicated in bold; and the highest value among the bold results per column is underlined.

Here we build LS-SVM classifiers with linear and RBF kernels. The input variables used for model building are MODEL1 and MODEL2, respectively. The corresponding LS-SVM models will be denoted as LS-SVM1 and LS-SVM2. Subscripts 'RBF' and 'Lin' indicate the kernel type that is used.

All the training data are normalized in order to have zero mean and unit variance. The same normalization is applied to the test set using the mean and variance estimates from the training set. The model performance measure is estimated based on the output probability of the model. The AUC and its computed standard error (S.E.) [7], which are independent of the cutoff value, are reported in the second column of Table 2. Also listed in Table 2 are the performance measures calculated at different decision levels (for LS-SVMs and LRs, those levels are probability cutoff levels). They include: the accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Predictive value helps in interpreting the test result for an individual. The PPV is the proportion of all positive tests that are true positive; the NPV is the proportion of all negative tests that are true negative. The numbers between the parentheses in the first column indicate the number of support vectors (N_{SV}) in the LS-SVM classifier in the first stage of model building.

The performance of the Risk of Malignancy Index (the RMI is calculated as the product of the CA 125 level, an ultrasound morphologic score, and a score for the patient's menopausal status) and two logistic regression (LR) models LR1 and LR2, using respectively MODEL1 and MODEL2 as inputs, are also reported for comparison.

Note that, the decision levels we used in the experiments are considered 'good' according to our model selection goal. They lead to a high sensitivity and low false positive rate on the training set; those decision levels, which result into a high accuracy but a too low sensitivity or specificity, are considered unacceptable in this context. The 'good' decision levels for LRs here are approximately the same as those for LS-SVMs, since we incorporate the same adjusted class prior, the $2:1$ ratio of the adjusted prior class probability between the malignant and benign cases, into the computation of the final outcome $(0 \sim 1)$. That is, by correcting the bias term b_0 in the LR model as follows: $b = b_0 - \log(N_+/N_-) + \log(2/1)$, where N_+ and N_- denote the number of malignant and benign cases in the training set, respectively. LS-SVM models within the evidence framework also shift the good decision level towards the 0.5 probability level after taking the adjusted class priors into account.

Let us have a look at Table 2. First, we can see that RMI has the worst performance on the test set, all the other models have obviously higher AUCs than RMI, and its accuracy and sensitivity are also lower compared with those of the other models. However the difference in AUC for linear LS-SVMs and LRs versus RMI is not significant according to the comparison measure in [7] (see the P-values in Table 3 obtained from two-tailed z-tests), though the AUCs of LS-SVMs and LRs on the training set are all significantly better than that of RMI (P-values< 0.001). Comparing LS-SVM2_{RBF} with RMI, a significant P-value of 0.048 is obtained, while the difference between LS-SVM1_{RBF} and RMI is close to significant, having a P-value of 0.066. Note that the comparison measure is considered to be conservative (AUC underestimated and the variance overestimated). Moreover, the variance of the estimated AUC will further decreases as more patients are included in the data set.

Now move to the comparison between linear LS-SVMs and LRs. The LS-SVMs with linear kernels have similar performance as LRs. However, the sensitivity for LRs is a little bit lower than that of linear LS-SVMs, at the same specificity level. For example, at a decision level of 0.5, both LR1 and linear LS-SVM1 have the same specificity 85%, but the sensitivity of LR1 is 74% which is lower compared with 78% for the linear LS-SVM1.

We can also easily observe that LS-SVMs with RBF kernels have slightly better performance than both linear LS-SVM and LR models; LS-SVM$_{\mathrm{RBF}}$ models achieve

Table 3
Significance level when two AUCs on the test set from the temporal cross-validation are compared (P-value from pairwise two-tailed z-test)

Model	LR1	LR2	LS-SVM1_{Lin}	LS-SVM2_{Lin}	LS-SVM1_{RBF}	LS-SVM2_{RBF}
RMI	0.183	0.121	0.120	**0.077**	**0.066**	**0.048**
LR1	1.000	0.635	0.553	0.408	0.443	0.324
LR2	0.635	1.000	0.825	0.429	0.809	0.431

Note: P-values that are significant or close to significant are indicated in bold.

consistently the highest AUC, sensitivity and specificity on the test set. The consistently higher positive predictive value and negative predictive value of LS-SVM models compared to those of LRs also point out that LS-SVMs perform better than LRs. Moreover, the LS-SVM1$_{\mathrm{RBF}}$ achieves also higher performance on the training set, with AUC 0.990 versus 0.976 for LR1. Hence based on this result, we conclude that LS-SVM models with RBF kernels are recommended in this case.

As to the effect of using different input variables, by pairwise comparison of the models based on MODEL1 and MODEL2, we find that the models generated by MODEL2 (less variables) have marginally higher AUCs on the test set (though the performance on the training set is the opposite). However the difference is not statistically significant. On the other hand, the models derived from MODEL1 have higher accuracy than those derived from MODEL2given the same sensitivity at those 'good' decision levels. We thus conclude that the input variables selected within the evidence framework, i.e. MODEL1, based on the training data only, have comparable performance with MODEL2, which were selected based on the whole data set using stepwise logistic regression. Actually, the input variables selected by stepwise logistic regression based on only the training data have poorer performance than both MODEL1 and MODEL2. This provides again evidence for the appropriateness of our input selection procedure.

It is also interesting to see how the class probability reflects the uncertainty of the decision making. The uncertainty is the largest, when the probability of one case to be malignant is 0.5. So we could predefine an uncertainty region of the probability $[0.5 \pm t]$, where t is a small positive value between 0 and 0.5. To make the decision more reliable, the classifier should reject the cases whose outcome falls in this uncertainty region. Clinically, this means that those patients will be referred to further examination.

Now, we take the classifier LS-SVM1$_{\mathrm{RBF}}$ as an example. When t is set to 0.2, the uncertainty region becomes (0.3–0.7). Fixing the decision probability level at 0.5, when we accept all the test cases, the accuracy is 84% with a sensitivity of 78% and a specificity of 88%. When rejecting the 14 (9%) uncertain cases, we obtain a reasonably higher performance based on the reduced test set with an AUC of 0.9325, accuracy 88%, sensitivity 83% and specificity 90%.

4.6. Results from randomized cross-validation

We have already described a temporal validation above, where the splitting of training set and test set is non-random. In this section, we will report the results based on 30 runs of stratified cross-validation. In each run of cross-validation, the 265 training data and 160 test data are randomly selected. The same two subsets of input variables MODEL1 and MODEL2 are used. Same types of models as the previous ones are to be evaluated.

The average of AUC ($\overline{\mathrm{AUC}}$), the corresponding standard deviation (S.D., derived from 30 AUC values), accuracy, sensitivity and specificity are reported in Table 4. The number between the parentheses indicates the mean of the number of support vectors ($\overline{N}_{\mathrm{SV}}$). Boxplots in Figs. 2 and 3 illustrate the distribution of the AUCs over the 30 validations on test and training set, respectively.

From this experiment, an increase of the validation performance is observed, which is mainly due to the randomization of the training set and test set. However, we still obtain

Table 4
Averaged performance on the test set from 30 runs of randomized cross-validation ($N_{\text{train}} = 265, N_{\text{test}} = 160$)

Model type ($\overline{N}_{\text{SV}}$)	$\overline{\text{AUC}}$ ($\pm$S.D.)	Decision level	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
RMI	0.8882 ($\pm$0.0284)	100	82.65	81.73	83.06	68.89	90.96
		80	**81.10**	**83.87**	**79.85**	**65.61**	**91.63**
LR1	0.9397 ($\pm$0.0209)	0.5	83.29	89.33	80.55	67.81	94.43
		0.4	**81.94**	**91.60**	**77.55**	**65.16**	**95.38**
LS-SVM1$_{\text{Lin}}$ (150.2)	0.9405 ($\pm$0.0199)	0.5	84.31	87.40	82.91	70.09	93.62
		0.4	**82.77**	**90.47**	**79.27**	**66.61**	**94.88**
LS-SVM1$_{\text{RBF}}$ (137.1)	0.9424 ($\pm$0.0207)	0.5	84.85	86.53	84.09	71.46	93.31
		0.4	**83.52**	**90.00**	**80.58**	**67.98**	**94.71**
LR2	0.9403 ($\pm$0.0211)	0.5	82.37	88.80	79.45	66.53	94.08
		0.4	**80.42**	**91.60**	**75.33**	**63.03**	**95.27**
LS-SVM2$_{\text{Lin}}$ (145.9)	0.9404 ($\pm$0.0206)	0.5	84.10	87.13	82.73	69.96	93.50
		0.4	**81.71**	**90.07**	**77.91**	**65.20**	**94.60**
LS-SVM2$_{\text{RBF}}$ (132.9)	0.9415 ($\pm$0.0201)	0.5	84.60	85.27	84.30	71.49	92.73
		0.4	**82.65**	**88.67**	**79.91**	**66.97**	**94.01**

Note: The 'best' results of each model obtained at a certain decision level are indicated in bold; and the highest value among the bold results per column is underlined.

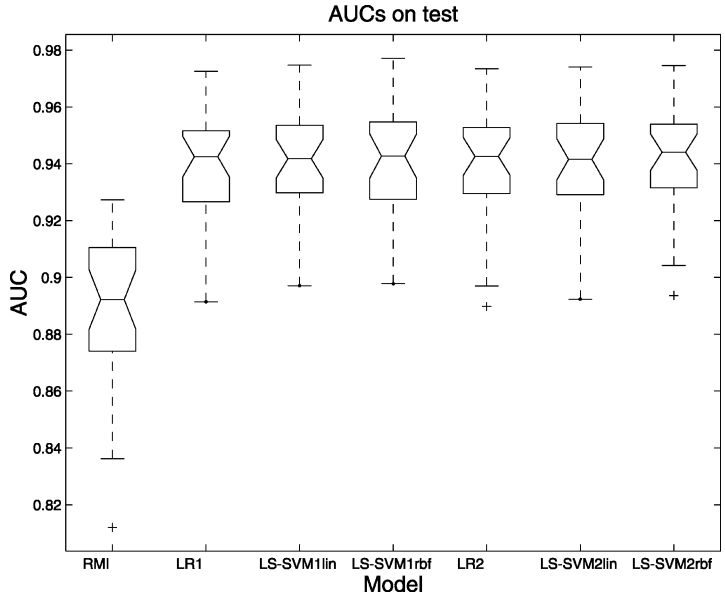

Fig. 2. Boxplot of the AUCs over 30 runs of cross-validation based on the test set (the line in the middle of the notched 'box' is the sample median, the lower and upper lines of the 'box' are the 25th and 75th percentiles of the sample).

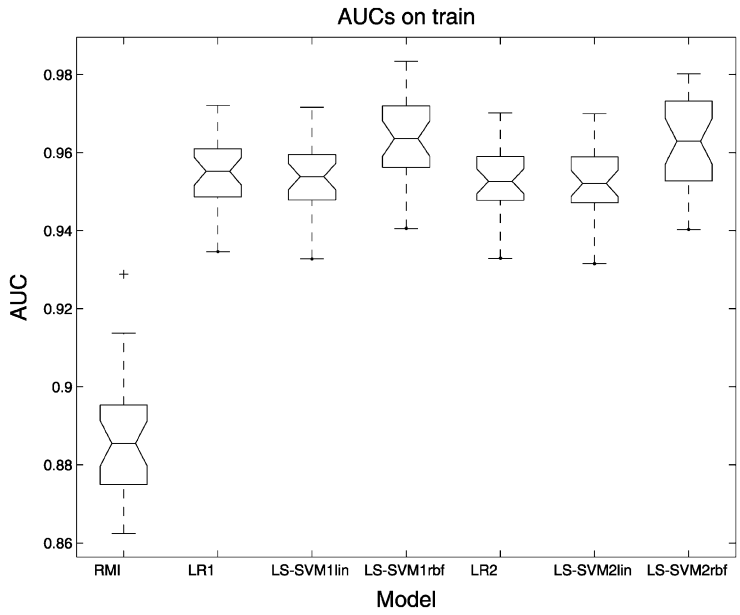

Fig. 3. Boxplot of the AUCs over 30 runs of cross-validation based on the training set (the line in the middle of the notched 'box' is the sample median, the lower and upper lines of the 'box' are the 25th and 75th percentiles of the sample).

Table 5
Rank ordered significant subgroups from multiple comparison on mean AUC from randomized cross-validation

Model	RMI	LR1	LR2	LS-SVM2$_{Lin}$	LS-SVM1$_{Lin}$	LS-SVM2$_{RBF}$	LS-SVM1$_{RBF}$
$\overline{AUC}$	0.8882	0.9397	0.9403	0.9404	0.9405	0.9415	0.9424
S.D.	0.0284	0.0209	0.0211	0.0206	0.0199	0.0201	0.0207

Note: Only the mean AUC of RMI is significantly different from the others.

quite consistent results with the previous single hold-out cross-validation. Among the seven models, RMI has the worst performance. LS-SVM1$_{RBF}$ obtained the best averaged performance, with mean AUC = 0.9424. This can be seen more clearly from Table 5 in which the different models are ordered by the mean of AUCs.

To make a simultaneous comparison of all the models in mean of AUCs, we conduct a one-way ANOVA followed by Tukey multiple comparison [16]. Results are reported in Table 5. The subsets of adjacent means that are not significantly different at 95% confidence level are shown, and are indicated by drawing a line under the subsets. From this comparison, we observe that both LRs and LS-SVMs have significantly better performance than RMI, though the differences among the LR models and LS-SVM models with either linear or RBF kernel are not significant.

4.7. Comparison of the diagnostic performance with the human expert

One might be curious to know: can the computer model beat the expert? Trying to answer this question, we would like to compare the diagnostic results of our models with those of human investigators examining the same patients. The investigators were given all the available information and measurements of the patients before operation [24]. Table 6 shows the diagnostic performance of both the LS-SVM1$_{RBF}$ and the three human investigators. Assessor 1 (DT) is a very experienced expert, who had examined

Table 6
Comparison in diagnostic performance of the model and the human assessors

	Accuracy	Sensitivity	Specificity	PPV	NPV
(a) On the 265 cases collected between 1994 and 1997					
LS-SVM1$_{RBF}$	0.8981	0.9750	0.8649	0.7573	0.9877
Assessor 1	0.9132	0.9750	0.8865	0.7879	0.9880
Assessor 2	0.8189	0.9000	0.7838	0.6429	0.9477
Assessor 3	0.8113	0.8750	0.7838	0.6364	0.9355
(b) On the 160 cases collected between 1997 and 1999					
LS-SVM1$_{RBF}$	0.8313	0.8148	0.8396	0.7213	0.8990
Assessor 1	0.8938	0.8148	0.9340	0.8627	0.9083
(c) Averaged on the test sets from the randomized cross-validation					
LS-SVM1$_{RBF}$	0.8352	0.9000	0.8058	0.6798	0.9471
Assessor 1	0.9058	0.9113	0.9033	0.8123	0.9574

Note: The results of LS-SVM1$_{RBF}$ are obtained at the decision level of 0.4 in (a–c), and in (a) are the performances of recalling on the training set.

ultrasonographically more than 5000 patients. Assessor 2 and 3 are less experienced, who had performed about 200 and 300 ultrasonographical examinations, respectively. Unfortunately, the assessment of the two less experienced assessors are not available on the cases collected after 1997, hence we will mainly focus on the comparison with the expert.

The expert's diagnosis on the 160 newest patients (test set) results in a sensitivity of 81.48%, specificity 93.40% and PPV 86.27%. The LS-SVM1$_{\mathrm{RBF}}$ model gives a diagnostic performance at 0.4 decision level with a sensitivity of 81.48% (same as for the expert), however, a lower specificity of 83.96%, and PPV 72.13%. When looking at the averaged performance on the same randomized cross-validation, similar conclusions can be drawn. The human expert has a sensitivity 91.13%, specificity 90.33% and PPV 81.23%, while the LS-SVM1$_{\mathrm{RBF}}$ has an averaged sensitivity of 90.00%, specificity 80.58% and PPV 67.98%. In summary, the LS-SVM model can achieve the same sensitivity as the expert, however at the cost of a higher false positive rate.

The comparison points out that the models we have till now have not yet been able to beat the experienced human expert. However, from Table 6(a), we observe that the model performs significantly better than the other less experienced assessors 2 and 3 on the old patient group. If the model is assessed by the average performance from the randomized cross-validation, it can also be inferred that the model can better discriminate preoperatively between benign and malignant tumors than the less experienced assessors.

5. Discussion

Next, we would like to discuss several issues related to the application of our diagnostic model in clinical practice.

We first indicate some possible reasons why the expert is still outperforming the models obtained from given amount of data in the positive predictive value. The most important reason is that the expert here is very experienced. The mathematical models would need to reach very high levels of test performance to be comparable in performance to such kind of international top-experts. Comparing the performance of our model to that of less experienced assessors, we can still see the potential value of the mathematical models in helping those investigators with less experience to predict preoperatively the correct outcome.

Another reason might be the absence of prior knowledge in the models, which is abundantly owned by the experts. The quality of a purely data driven model also depends on the quality and quantity of the training data. The representativeness of the training data is critical for the learning and generalization performance. The incorporation of expert knowledge into black-box models is a good idea to compensate for the shortcomings of black-box models. A hybrid approach, which exploits the expert knowledge (represented in a belief network) in the learning of MLPs, has been applied to this ovarian tumor classification problem and has shown its potential to improve the performance of basic MLPs [4]. However, further validation of the approach based on more data is still needed. Future work includes applying a similar hybrid methodology to the LS-SVM models.

A third reason is probably due to the fact that the expert makes his diagnosis based on more information of the patients than available in our black-box model design. Indeed,

some clinical features, e.g. some medical history, family history, genetic factors, and the whole image of transvaginal sonography, etc. are not accessed by the mathematical models.

In addition, the application of the evidence framework here might also be partially responsible for a degradation in performance whenever the given assumptions are not satisfied, though it has several advantages as mentioned before. These Gaussian assumptions still need to be verified. The more training data, the better the assumption will be satisfied.

Another important issue is how to split the data for validating the diagnostic models. The splitting of the data set in a training and test set according to the time scale is more natural in clinical practice. There is a danger for changes in the patient population over time. The more experienced the expert, the more difficult cases are being referred to him for diagnosis, implying that the test set includes a higher number of harder cases (e.g. with borderline malignancy) to diagnose.

Moreover, a homogeneity analysis of the group difference reveals that significant differences (at significance level of 0.05) exist in age between the old patient group (data from 1994 to 1997) and the new patient data set (data from 1997 to 1999). The mean age of the 160 new patients is 48.6 (16–78), which is lower than the mean age of the 265 old patients given by 52.4 (21–93). The proportion of post-menopausal patients in the new data set (41.9%) is also lower than the one in the old data set (48.3%). Moreover, it is well known that the level of tumor marker CA 125 can better predict the presence of cancer in post-menopausal patients, compared to that in pre-menopausal patients. This implies that it is harder to predict correctly the malignancy of the tumors in the new patient group compared to that in the old patient group.

One can observe this trend in time scale from the performance of our model. The performance of the model decreases, from an AUC of 0.99, sensitivity 97.5%, specificity 86.50% when applied to the old patient group (training set), to an AUC of 0.92, sensitivity 81.48% and specificity 83.13% when applied to the new patient group (both obtained by taking 0.4 as the probability decision level). Even for the expert, preoperative detection of cancer in the new patient group is more difficult than in the old patient group, which can be seen from the drop of the sensitivity from 97.5% (specificity 88.65%) in the old patient group, to 81.48% (specificity 93.40%) in the new patient group.

A random splitting of test and training set leads to a more equilibrated distribution of the patient data over both sets other than for a random variation, and is thus a weak procedure and less stringent [1]. This splitting is not representative for the way the models are used in clinical practice, where a prospective evaluation is normally needed.

The temporal validation performance of our LS-SVM model is quite encouraging though not perfect. It has a consistent cancer detection rate comparable to that of the expert, while maintaining an acceptable false positive rate. Furthermore, the output probability of the LS-SVM model enables it to assist the clinicians in making rational management decisions about their patients and to counsel them appropriately.

On the other hand, we must realize the gap between the modeling and the real world. One can expect that this gap will become smaller given a larger amount of training examples; this is also one motivation for the International Ovarian Tumor Analysis (IOTA) project. IOTA is a multi-center study on the preoperative characterization of ovarian tumors based on artificial intelligence models [25]. More than 1000 patient data from more than ten

centers located in different countries, including Belgium, UK, Sweden, France and Italy have been collected. Based on this sufficiently large data set, mathematical models can be developed for preoperative classification of benign and malignant ovarian tumors, and further subclassify the tumors (e.g. borderline malignant, endometrioma). The variation between centers in outcomes of histology and the performance of the models will also be assessed. Then another 1000 patient data will be collected for future prospective validations.

6. Conclusions

In this paper, we apply the LS-SVM models within the Bayesian evidence framework in order to discriminate between benign and malignant ovarian tumors. Advantages of this approach include the ones inherited from the SVM, e.g. a unique solution, and support of statistical learning theory. Moreover, after integration with a Bayesian approach, the determination of the model, regularization and kernel parameters, can be done in a unifying way, without the need of selecting an additional validation set.

A forward selection procedure which aims to maximize the model evidence has been proved to be able to identify the important variables for model building. A sparse approximation procedure applied to the LS-SVM classifier also further improves the generalization performance of the LS-SVM models.

The posterior class probability for malignancy of ovarian tumor for each individual patient can be computed through Bayes' rule, incorporating the prior class probability and misclassification cost. This output probability enables the possible application of our mathematical model in clinical practice.

Two types of LS-SVM models with linear and RBF kernels, and logistic regression models have been built based on 265 training data, and evaluated on 160 newly collected patient data from the same center. They all have much better performance than RMI. The LS-SVM classifier with an RBF kernel achieves the best performance compared with the others, evidenced by consistently achieving the highest rank in AUC, sensitivity, and positive predicting value. Our randomized cross-validation does also confirm the good generalization performance of LS-SVM models. Though the discrepancy between the performance of the linear and nonlinear models is not statistically significant, this can only be verified by using a larger amount of cases for training and testing.

We conclude that LS-SVM models have the potential to reliably predict malignancy of the ovarian tumors, though the models by now have not yet been able to beat the very experienced human expert. Furthermore, a hybrid approach, which combines the learning ability of black-box models and the expert knowledge of white-box models (e.g. Bayesian network) might further improve the model performance. This will be the subject of the future research.

Acknowledgements

We would like to thank our reviewers for their constructive comments. The research work is supported by the Belgian Programme on Interuniversity Poles of Attraction

(IUAP V-22), initiated by the Belgian State, Prime Minister's Office, Federal Office for Scientific, Technical and Cultural Affairs, of the Concerted Research Action (GOA) projects of the Flemish Government MEFISTO-666, of the IDO/99/03 and IDO/02/09 projects (K.U. Leuven), 'Predictive computer models for medical classification problems using patient data and expert knowledge' and of the FWO (Fund for Scientific Research Flanders) projects G.0407.02, G.0269.02, G.0413.03, G.0115.02, G.0388.03, G.0229.03. TVG and JS are postdoctoral researcher with the National Fund for Scientific Research FWO—Flanders. CL is supported by a K.U. Leuven doctoral fellowship.

References

[1] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73.
[2] Antal P, Fannes G, Verrelst H, De Moor B, Vandewalle J. Incorporation of prior knowledge in black-box models: comparison of transformation methods from Bayesian network to multilayer perceptrons. In: Workshop Notes: Workshop on Fusion of Domain Knowledge with Data for Decision Support in Conjunction with the 16th Uncertainty in Artificial Intelligence Conference. Standford, CA, 2000, p. 42–8.
[3] Antal P, Verrelst H, Timmerman D, Moreau Y, Van Huffel S, De Moor B, et al. Bayesian networks in ovarian cancer diagnosis: potentials and limitations. In: Proceeding of the 13th IEEE Symposium on Computer-Based Medical Systems (CBMS 2000). Houston (TX): IEEE Computer Science Press; 2000. p. 103–9.
[4] Antal P, Fannes G, Timmerman D, De Moor B, Moreau Y. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. Artif Intell Med, in press.
[5] Bishop CM. Neural networks for pattern recognition. Oxford: Oxford University Press; 1995.
[6] Cristianini, N, Shawe-Taylor J. An introduction to support vector machines. Cambridge (UK): Cambridge University Press; 2000.
[7] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
[8] Hanley JA, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
[9] Jeffreys H. Theory of probability. New York: Oxford University Press; 1961.
[10] Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. Br J Obstet Gynaecol 1990;97:922–9.
[11] Lu C, De Brabanter J, Van Huffel S, Vergote I, Timmerman D. Using artificial neural networks to predict malignancy of ovarian tumors. In: Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2001). Istanbul, Turkey, 2001 (CD-ROM).
[12] MacKay DJC. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. Network: Comput Neural Syst 1995;6:469–505.
[13] MacKay DJC. The evidence framework applied to classification networks. Neural Comput 1992;4(5): 698–741.
[14] Mercer J. Function of positive and negative type and their connection with the theory of integral equations. Philos Trans R Soc Lond Ser A: Math Phys Eng Sci 1909;209:415–46.
[15] Neal RM. Bayesian learning for neural networks: lecture notes in statistics, vol. 118. New York: Springer; 1996.
[16] Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied linear statistical models. 4th ed. Chicago (Ill): McGraw-Hill/Irwin; 1996.
[17] Ozols RF, Rubin SC, Thomas GM, Robboy SJ. Epithelial ovarian cancer. In: Hoskins WJ, Perez CA, Young RC, editors. Principles and Practice of Gynecologic Oncology. Philadelphia: Lippincott Williams and Wilkins; 2000. p. 981–1058.
[18] Provost F, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. In: Shavlik J, editor. Proceedings of the 15th International Conference on Machine Learning (IMLC-98). Morgan Kaufmann, San Francisco, CA, 1998, p. 445–53.

[19] Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9(3):293–300.
[20] Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least squares support vector machines. Singapore: World Scientific; 2002.
[21] Suykens JAK, De Brabanter J, Lukas L, Vandewalle J. Weighted least squares support vector machine: robustness and sparse approximations. Neurocomputing 2002;48(1–4):85–105 (Special issue on fundamental and information processing aspects of neurocomputing).
[22] Timmerman D, Bourne TH, Tailor A, Collins WP, Verrelst H, Vandenberghe K, et al. A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: the development of a new logistic regression model. Am J Obstet Gynecol 1999;181:57–65.
[23] Timmerman D, Verrelst H, Bourne TH, Moor B D, Collins WP, et al. Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. Ultrasound Obstet Gynecol 1999;13:17–25.
[24] Timmerman D, Schwärzler P, Collins WP, Claerhout F, Coenen M, Amant F, et al. Subjective assesment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. Ultrasound Obstet Gynecol 1999;13:11–6.
[25] Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group. Ultrasound Obstet Gynecol 2000;16:500–5.
[26] Van Gestel T, Suykens JAK, Lanckriet G, Lambrechts A, De Moor B, Vandewalle J. A Bayesian framework for least squares support vector machine classifiers. Neural Comput 2002;15(5):1115–48.
[27] Van Gestel T, Suykens JAK, Baestaens D-E, Lamrechts A, Lanckriet G, Vandaele B, et al. Financial time series prediction using least squares support vector machines within the evidence framework. IEEE Trans Neural Network 2001;12(4):809–21 (Special issue on financial engineering).
[28] Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag; 1995.
[29] Vergote I, De Brabanter J, Fyles A, Bertelsen K, Einhorn N, Sevelda P, et al. Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. Lancet 2001;357(9251):176–82.