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 Leuven, Herestraat 49, B-3000 Leuven, Belgium*

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.

Key words: 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
Email addresses: Johan.Suykens@esat.kuleuven.ac.be (J. A. K. Suykens), Sabine.VanHuffel@esat.kuleuven.ac.be (S. Van Huffel).

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. However, nowadays two thirds 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 also 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 an oncological surgeon. 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 ultrosonographic morphologic findings and the menopausal status of the patient [1]. 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 [2,3], 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 which combines the Bayesian belief network, which represents the expert knowledge in the graphical model, into the learning of MLPs, has also been investigated in [6–8]. 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. 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 [10]. 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, supporting their good empirical results. In this paper, a least squares version of SVMs (LS-SVMs) [11] 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 [15]. 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 benchmarking problems [17] and to the prediction of financial time series [18]. 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 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 the University Hospitals Leuven, Belgium, between 1994 and 1999. These patients have a persistent extrauterine pelvic mass, and the study is designed mainly for preoperative differentiation between benign and malignant adnexal masses [2]. Patients without preoperative results of serum CA 125 levels (n=100) have been excluded from this analysis. 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 [2–5].

The exploratory data analysis that has been first conducted includes the following steps.

Data preprocessing: The original data set contains 25 features. 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.

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.

Multivariate data analysis: To get a first idea of the important predictors, we performed a factor analysis using the principal components as factors. Fig. 1 shows the biplot in a 2-dimensional space generated by the first two principal components called FACTOR1 and FACTOR2. The biplot visualizes 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 '$\times$') 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. On the other hand, quite a lot of overlap between the two classes can also be observed.

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 (SVMs) [10], 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, 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 (LS-SVMs) [11], 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 VC theory [10].

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 evidence framework [13–15]. 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 [17], the evidence framework has been applied to LS-SVMs for classification. Thanks to 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 please consult the papers [11,12,17,18]. Then we introduce a LS-SVM input 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 $y(x) = \text{sign}\left[w^T\varphi(x) + b\right]$ is inferred from the data $D = \{(x_i, y_i)\}_{i=1}^N$ with binary targets $y_i = \pm 1$, by minimizing the following cost function:

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

subject to the equality constraints

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

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

One defines the Lagrangian

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

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

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

with $Y = [y_1 \cdots y_N]^T$, $\alpha = [\alpha_1 \cdots \alpha_N]^T$, $e = [e_1 \cdots e_N]^T$, $1_v = [1 \cdots 1]^T$, and I_N the $N \times N$ identity matrix. Mercer's theorem is applied to the matrix Ω with $\Omega_{i,j} = y_i y_j \varphi(x_i)^T\varphi(x_j) = y_i y_j K(x_i, x_j)$, where $K(\cdot, \cdot)$ is a chosen positive definite kernel that satisfies Mercer condition. The most common kernels include a linear kernel $K(x_1, x_2) = x_1^T x_2$ and an RBF kernel $K(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2/\sigma^2)$. The LS-SVM classifier is then constructed in the dual space as:

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

Here also remarkable is that, the least squares formulation is related to kernel Fisher Discriminant Analysis. And an LS-SVM with linear kernel corresponds to linear Fisher Discriminant Analysis (FDA) with regularization term [17].

3.1.1 Inference of Model Parameters (Level 1)

The parameter 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 $\mathcal{H}$ corresponds to the kernel function K with different kernel parameters such as the width of an RBF kernel σ. The *evidence* $p(D|\log\mu, \log\zeta, \mathcal{H})$ 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 minus the log likelihood for a noise model: $p(D|w, b, \log\zeta, \mathcal{H}) \propto \exp(-\zeta E_D)$. 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 for w, with variance $1/\mu$, $p(w|\log\mu, \mathcal{H}) = (\frac{\mu}{2\pi})^{n_f/2}\exp(-\frac{\mu}{2}w^Tw)$, 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 parameter w and b: $p(w, b|\log\mu, \log\zeta, \mathcal{H}) = p(w|\log\mu, \mathcal{H})p(b|\log\sigma_b, \mathcal{H}) \propto \exp(-\mu E_W)$.

Hence the expression for the first level of inference becomes

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

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 $p(y|x, D, \log\mu, \log\zeta, \mathcal{H})$.

In the evidence framework, we assume that the posterior distribution of w can be approximated by a single Gaussian at w_{MP}. We define an error term corresponding to different classes (indicated by subscripts '+' and '−') as $e_\pm = w^T(\varphi(x) - \hat{m}_\pm)$, 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

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

where $\mathcal{I}_+$ and $\mathcal{I}_-$ 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

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

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

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

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

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

Thus the conditional probabilities can be computed as:

$$\begin{aligned} & p(x | y = \pm 1, D, \log \mu, \log \zeta, \log \zeta_{\pm}, \mathcal{H}) \\ & \quad = (2\pi(\zeta_{\pm}^{-1} + \sigma_{e_{\pm}}^2))^{-1/2} \exp(-\frac{m_{e_{\pm}}^2}{2(\zeta_{\pm}^{-1} + \sigma_{e_{\pm}}^2)}). \end{aligned} \tag{11}$$

By applying Bayes' rule the following posterior class probabilities of the LS-SVM classifier are obtained (for notation simplicity, $\log \mu, \log \zeta, \log \zeta_{\pm}, \mathcal{H}$ 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. These posterior class probabilities could also be used to make minimum risk decisions in the case of different misclassification costs.

3.1.3 Inference of Hyperparameters (Level 2)

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

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

where a uniform distribution is assumed in $\log\mu$ and $\log\zeta$ for the prior $p(\log\mu, \log\zeta|\mathcal{H}) = p(\log\mu|\mathcal{H})p(\log\zeta|\mathcal{H})$. The probability $p(D|\log\mu, \log\zeta, \mathcal{H})$ is equal to the evidence of the previous level. Using a local Gaussian approximation at the maximum a posteriori estimates $w_{\mathrm{MP}}, b_{\mathrm{MP}}$, we obtain:

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

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 Gram matrix in the feature space and are the solution of the eigenvalue problem

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

The effective number of parameters [13,16] 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 be approximated in this way:

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

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

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

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

This optimal hyperparameter γ is then obtained by solving the optimization problem (16) 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 $\mathcal{H}_j$. The prior $p(\mathcal{H}_j)$ over all possible models is assumed to be uniform. Thus the posterior for the model $\mathcal{H}_j$ is in the form of $p(\mathcal{H}_j|D) \propto p(D|\mathcal{H}_j)p(\mathcal{H}_j) \propto p(D|\mathcal{H}_j)$. At this level, no evidence or normalizing constant is used since it is infeasible to compare all possible models $\mathcal{H}_j$.

A separable Gaussian prior for $p(\log\mu_{\text{MP}}, \log\zeta_{\text{MP}}|\mathcal{H}_j)$ is assumed for all models $\mathcal{H}_j$, with the constant standard deviation $\sigma_{\log\mu}$ and $\sigma_{\log\zeta}$. And we assume that $p(\log\mu, \log\zeta|D, \mathcal{H}_j)$ can be well approximated by using a separable Gaussian with error bars $\sigma_{\log\mu|D}$ and $\sigma_{\log\zeta|D}$. The posterior likelihood $p(D|\mathcal{H}_j)$ corresponds to the evidence at the previous level and can be evaluated by:

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

The models can thus be ranked according to the evidence $p(D|\mathcal{H}_j)$, that is the tradeoff between the goodness of fit from the previous level $p(D|\log\mu_{\text{MP}}, \log\zeta_{\text{MP}}, \mathcal{H}_j)$ and the Occam factor $\frac{\sigma_{\log\mu|D}\sigma_{\log\zeta|D}}{\sigma_{\log\mu}\sigma_{\log\zeta}}$ [16].

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

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

One selects the kernel parameters with maximal posterior $p(D|\mathcal{H}_j)$.

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 [13]. We denote the nor-

malized 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. Now, we start the design of the LS-SVM classifier in a Bayesian framework. Several procedures including hyperparameter tuning, input selection and sparse approximation are to be established.

Hyperparameter Tuning

Select the model $\mathcal{H}_j$ 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 $p(D|\mathcal{H}_j)$ from (20) at the third level.

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

Input Selection

Within the Bayesian evidence framework, given a certain type of kernel, we can conduct input selection.

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 can no longer increase the model evidence.

Sparse Approximation

Due to the choice of the 2-norm in cost function, the sparseness is lost compared with the standard QP type SVMs. However, as has been shown in [12], 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.

Probabilistic Interpretation of the Output

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

1. Given the parameters $\alpha, b_{\text{MP}}, \mu_{\text{MP}}, \zeta_{\text{MP}}, \gamma_{\text{MP}}, \gamma_{\text{eff}}$ and the eigenvalues and eigenvectors in (14) available from the designed model $\mathcal{H}_j$, calculate $m_{e+}, m_{e-}, \sigma_{e+}^2$ and σ_{e-}^2 from (7) and (10) respectively. Compute ζ_+ and ζ_- from (17).
2. Calculate $p(x|y = \pm 1, D, \log \mu, \log \zeta, \log \zeta_{\pm}, \mathcal{H}_j)$ from (11).
3. Calculate $p(y|x, D, \mathcal{H}_j)$ from (12) by using the prior class probabilities $P(y = +1)$ 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 (ROC) 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 perform a prospective evaluation, independently of the training data and model fitting process. The data set is split into two parts 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 the inefficient use of the data set which is already moderately small in this case [15]. The following procedures including input selection and model fitting, are independent from the test set.

However, a single hold-out cross-validation estimate 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 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 which 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 the univariate models which each contain 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 little 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: *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 are: *L_CA125, Pap, Sol, Colsc4, Unsol, Colsc3, Bilat, Shadows, Asc, Smooth, Meno*. Though the two subsets of variables have 9 variables in common, the evidence of the model selected based on the RBF kernel is higher than the one based on the linear kernel, and therefore MODEL1 is used here for model building instead of the other.

In previous work [9], 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 this classification problem, we aimed 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 prior for the malignant class. In the following experiments, the prior for the malignant class is 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. 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 AUC is a one-value measure of the accuracy of a test. It 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 [19]. The method proposed by DeLong et al. [20] will be used to compute the variance and covariance of the nonparametric AUC estimates derived from the same set of cases. This can be used for comparing two different ROC curves.

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* [22].

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. The model performance measure is estimated based on the output probability of the model. The AUC and its computed standard error (SE) [20], 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). 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 (RMI) 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 class prior, the 2 : 1 ratio of the 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 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 [20] (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 different with that of RMI (p-values< 0.001). Comparing LS-SVM2$_{\mathrm{RBF}}$ with RMI, a significant p-value of 0.048 is obtained, while the difference between LS-SVM1$_{\mathrm{RBF}}$ 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 become lower when derived from more cases.

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$_{\text{RBF}}$ models achieve 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$_{\text{RBF}}$ 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.

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 MODEL2 given the same sensitivity at those 'good' decision levels. Hence, we 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$_{\text{RBF}}$ as an example. When t is set to 0.2, the uncertainty region becomes $(0.3 \sim 0.7)$. Fixing the decision probability level at 0.5, when we accept all the test cases, the accuracy is 84% with sensitivity 78% and specificity 88%. When rejecting the 14 (9%) uncertain cases, we obtain a reasonably higher performance based on the reduced test set: AUC - 0.9325, accuracy - 88%, sensitivity - 83%, 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 (SD, 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 Fig. 2 and Fig. 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 quite consistent results with the previous single hold out cross-validation. Among the 7 models, RMI has the worst performance. LS-SVM1RBF obtained the best averaged performance, with mean AUC=0.9424. This can be seen clearer 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 [21]. 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.

5 Discussion

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

One might be curious to know: can the computer model beat the expert? Try to answer this question, we would like to compare the diagnostic results of our models with those of experts examining the same patients. The expert (DT) was given all the available information and measurements of the pa-

tients before operation. The expert's diagnosis on the later 160 patients (test set) are: accuracy - 89.38%, sensitivity - 81.48%, specificity - 93.40%, positive predictive value - 86.27% and negative predictive value - 90.83%; while the best diagnostic performance from the LS-SVM1RBF model has at 0.4 decision level: accuracy - 83.13%, sensitivity - 81.48% (same as for the expert) while the specificity is rather low, 83.96% only, positive predictive value - 72.13% and negative predictive value - 89.90%. When looking at the averaged performance on the same randomized cross-validation, similar conclusions can be drawn: the human expert has an overall accuracy 90.58%, sensitivity 91.13% and specificity 90.33%, positive predictive value 81.23%, and negative predictive value 95.74%; while the LS-SVM1RBF has an averaged sensitivity 90.00% and specificity 80.58%, positive predictive value 67.98% and negative predictive value 94.71%. In summary, the LS-SVM model can have 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 (DT). There are some possible reasons for the lesser performance of our models compared to the expert in the positive predicting value. The most important reason is that the expert here is very experienced, and has examined ultrasonographically more than 5000 - 10000 patients. Comparing the performance of our model to that of less experienced experts (having diagnosed 1000 patients or less) [4,8] shows that our models beat this class of less experienced experts. 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 [8]. 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 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 assumptions still need to be verified, as well as the Gaussianity of the approximation. The more training data, the better the latter assumption will be satisfied.

Another important issue is how to split the data for validating the prognostic 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. One can observe this trend in time scale from the performance of our model. The performance of the model decreases on the new cases, from as AUC 0.99, sensitivity 97.5%, specificity 86.50% when applied to the old 265 training cases, to AUC 0.92, sensitivity 81.48% and specificity 83.13% when applied to the new 160 cases (both obtained by taking 0.4 as the probability decision level). Even for the expert, the preoperative detection of cancer in the new cases (data from 1997 to 1999) is more difficult than in the old ones, which can be seen from the drop of the sensitivity from 97.5% (specificity 88.65%) in the old cases, to 81.48% (specificity 93.40%) in the new ones.

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. However, this splitting is not representative for the way the models are used in clinical practice.

The temporal validation performance of our LS-SVM model is quite encouraging though not perfect. It has a comparable and consistent cancer detection rate as 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 modelling and the real world. We are hoping this gap will become smaller given a larger amount of training examples; this is also one motivation for the International Ovarian Tumour Analysis (IOTA) project. IOTA is a multi-center study on the preoperative characterization of ovarian tumours based on artificial intelligence models [5]. More than 1000 patient data from more than ten centers located in different countries, including Belgium, USA, UK, Finland, Sweden, France, Italy and Austria, are 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 different 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. 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

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction (IUAP V-10-29), 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 project (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) project G.0407.02. JS is a postdoctoral researcher with the National Fund for Scientific Research FWO - Flanders.

References

[1] I. Jacobs, D. Oram, J. Fairbanks, J. Turner, C. Frost, J.G. Grudzinskas, 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) 97:922-929.

[2] D. Timmerman, T.H. Bourne, A. Tailor, W.P. Collins, H. Verrelst, K. Vandenberghe, I. Vergote, 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.

[3] D. Timmerman, H. Verrelst, T.H. Bourne, B. De Moor, W.P. Collins, I. Vergote and J.Vandewalle, Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses, *Ultrasound Obstet Gynecol* (1999) 13:17-25.

[4] D. Timmerman, P. Schwärzler, W.P. Collins, F. Claerhout, M. Coenen, F. Amant, I. Vergote, T.H. Bourne, Subjective assesment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience, *Ultrasound Obstet Gynecol* (1999) 13:11-16.

[5] D. Timmerman, L. Valentin, T.H. Bourne, W.P. Collins, H. Verrelst, I. Vergote, 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) 16:500-505.

[6] P. Antal, G. Fannes, H. Verrelst, B. De Moor, J. Vandewalle, Incorporation of prior knowledge in black-box models : comparison of transformation methods from Bayesian network to multilayer perceptrons, in *workshop on Fusion of Domain Knowledge with Data for Decision Support, 16th Uncertainty in Artificial Intelligence Conference* (2000) 42-48.

[7] P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, I. Vergote, Bayesian networks in ovarian cancer diagnosis: potentials and limitations. in *Proceedings 13th IEEE Symposium on Computer-Based Medical Systems* (CBMS 2000) 103-109.

[8] P. Antal, G. Fannes, D. Timmerman, B. De Moor, Y. Moreau, Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor

classification with rejection, 2002, submitted to the *Journal of Artificial Intelligence in Medicine*.

[9] C. Lu, J. De Brabanter, S. Van Huffel, I. Vergote, D. Timmerman, Using artificial neural networks to predict malignancy of ovarian tumors, in *Proc. of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Istanbul, Turkey, EMBC 2001) CD-ROM.

[10] V. Vapnik, *The Nature of Statistical Learning Theory* (Springer-Verlag, 1995).

[11] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, *Neural Processing Letters* (1999) 9(3):293-300.

[12] J.A.K. Suykens, J. De Brabanter, L. Lukas, J. Vandewalle, Weighted least squares support vector machine : robustness and sparse approximation, 2000, *Neurocomputing*, in press.

[13] C.M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1995).

[14] R.M. Neal, Bayesian Learning for Neural Networks, *Lecture Notes in Statistics* (Springer, New York, 1996) vol. 118.

[15] D.J.C. MacKay, The evidence framework applied to classification networks, *Neural Computation* (1992) 4(5): 698-741.

[16] D.J.C. MacKay, Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* (1995) 6:469-505.

[17] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, J. Vandewalle, Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis, 2001, *Neural Computation*, in press.

[18] T. Van Gestel, J.A.K. Suykens, D.-E. Baestaens, A. Lamrechts, G. Lanckriet, B. Vandaele, B. De Moor, J. Vandewalle, Financial time series prediction using least squares support vector machines within the evidence framework, *IEEE Transactions on Neural Networks* (Special Issue on Financial Engineering, 2001) 12(4):809-821.

[19] J.A. Hanley, B. McNeil, The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve, *Radiology* (1982) 143:29-36.

[20] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* (1988) 44:837-845.

[21] J. Neter, M.H. Kutner, C.J. Nachtsheim, W. Wasserman, *Applied Linear Statistical Models* (fourth edition, WCB/McGraw-Hill 1996).

[22] D.G. Altman, P. Royston, What do we mean by validating a prognostic model? *Statistics in Medicine* (2000) 19:453-473.

Figures

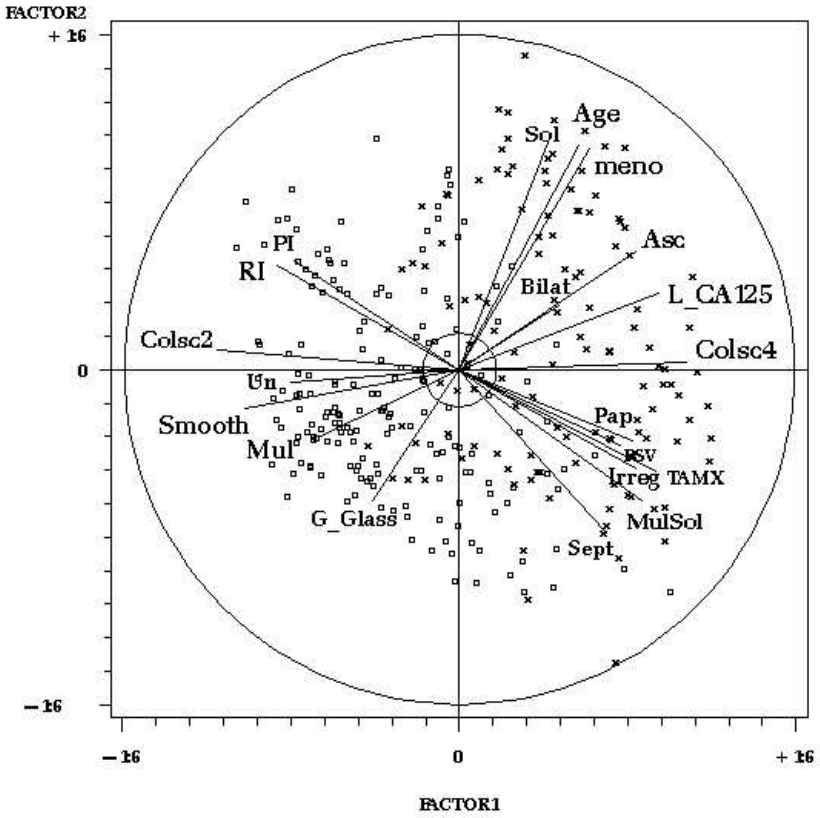

Fig. 1. Biplot of Ovarian Tumor data. The observations are plotted as points ('○'- benign, '×'- malignant), the variables are plotted as vectors from the origin, i.e. taking the respective factor loadings as the coordinates.

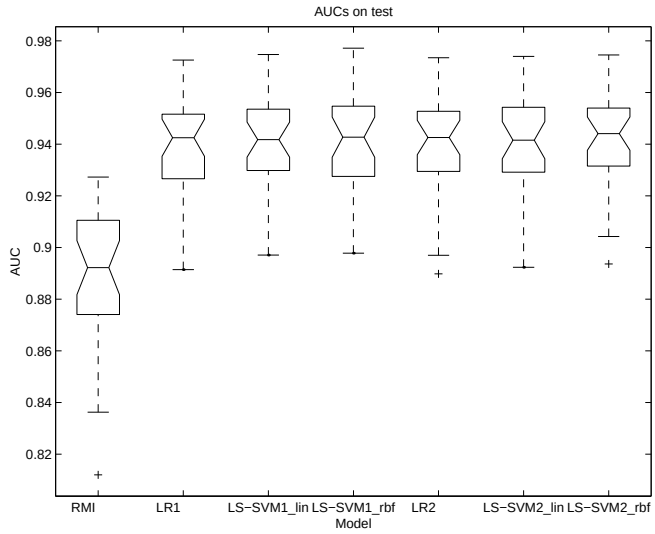

Fig. 2. Boxplot of the AUCs over 30 runs of cross-validation based on the test set (the line in the middle of the notched "box" is the sample median, the lower and upper lines of the "box" are the 25th and 75th percentiles of the sample.)

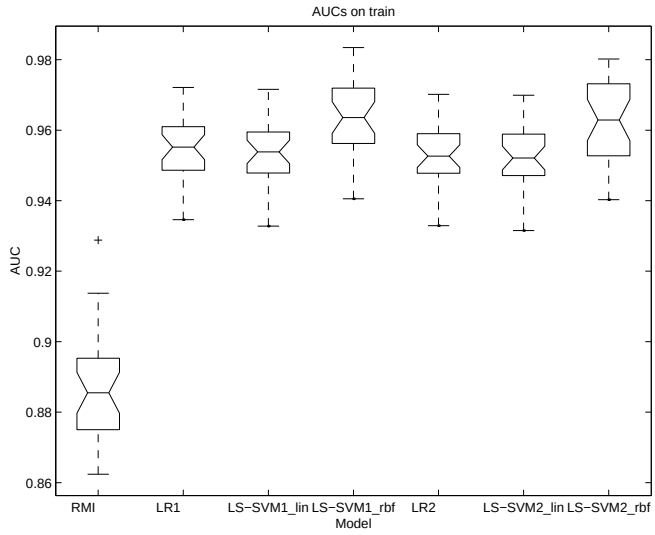

Fig. 3. Boxplot of the AUCs over 30 runs of cross-validation based on training set (the line in the middle of the notched "box" is the sample median, the lower and upper lines of the "box" are the 25th and 75th percentiles of the sample.)

Table 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
	Postmenopausal (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 doppler frequency (PSV)	19.8±14.6	27.3±16.6
	Mean doppler frequency (TAMX)	11.4± 9.7	17.4±11.5
B-Mode	Abdominal fluid (Asc)	32.7 %	67.3 %
Ultrasono	Unilocular cyst (Un)	45.8 %	5.0 %
Graphy	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> 3mm (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±SD in case of benign and malignant respectively are reported; for binary variables, the occurrences (%) of the corresponding features are reported.

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

Model Type (N_{SV})	AUC $\pm SE$	Decision Level	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
RMI	0.8733	100	78.13	74.07	80.19	65.57	85.86
	±0.0298	**75**	**76.88**	**81.48**	**74.53**	**61.97**	**88.76**
LR1	0.9111	0.5	81.25	74.07	84.91	71.43	86.54
	±0.0246	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$_{\text{Lin}}$	0.9141	0.5	82.50	77.78	84.91	72.41	88.24
(118)	±0.0236	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$_{\text{RBF}}$	0.9184	0.5	84.38	77.78	87.74	76.36	88.57
(97)	±0.0225	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	0.5	79.37	75.93	81.13	67.21	86.87
	±0.0218	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$_{\text{Lin}}$	0.9195	0.5	81.25	77.78	83.02	70.00	88.00
(115)	±0.0215	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$_{\text{RBF}}$	0.9223	0.5	83.75	81.48	83.96	73.33	90.00
(99)	±0.0213	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.

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-SVM1Lin	LS-SVM2Lin	LS-SVM1RBF	LS-SVM2RBF
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.

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	$\overline{\text{AUC}}$	Decision	Accuracy	Sensitivity	Specificity	PPV	NPV
($\overline{N_{\text{SV}}}$)	$\pm SD$	Level	(%)	(%)	(%)	(%)	(%)
RMI	0.8882	100	82.65	81.73	83.06	68.89	90.96
	±0.0284	**80**	**81.10**	**83.87**	**79.85**	**65.61**	**91.63**
LR1	0.9397	0.5	83.29	89.33	80.55	67.81	94.43
	±0.0209	**0.4**	**81.94**	**<u>91.60</u>**	**77.55**	**65.16**	**<u>95.38</u>**
LS-SVM1Lin	0.9405	0.5	84.31	87.40	82.91	70.09	93.62
(150.2)	±0.0199	**0.4**	**82.77**	**90.47**	**79.27**	**66.61**	**94.88**
LS-SVM1RBF	0.9424	0.5	84.85	86.53	84.09	71.46	93.31
(137.1)	±0.0207	**0.4**	**<u>83.52</u>**	**90.00**	**<u>80.58</u>**	**<u>67.98</u>**	**94.71**
LR2	0.9403	0.5	82.37	88.80	79.45	66.53	94.08
	±0.0211	**0.4**	**80.42**	**<u>91.60</u>**	**75.33**	**63.03**	**95.27**
LS-SVM2Lin	0.9404	0.5	84.10	87.13	82.73	69.96	93.50
(145.9)	±0.0206	**0.4**	**81.71**	**90.07**	**77.91**	**65.20**	**94.60**
LS-SVM2RBF	0.9415	0.5	84.60	85.27	84.30	71.49	92.73
(132.9)	±0.0201	**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.

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
$\overline{\text{AUC}}$	0.8882	0.9397	0.9403	0.9404	0.9405	0.9415	0.9424
SD	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.