A case study on the System for Paleographic Inspections (SPI): challenges and new developments

Fabio AIOLLI [a,1], *Arianna CIULA* [b]

[a] Dept. of Pure and Applied Mathematics, University of Padova, Italy
[b] Centre for Computing in the Humanities, King's College London, England

Abstract. *Of all the disciplines that study the past through its written heritage, paleography*[2] *is the one that focuses mainly on the analysis of the script borne by manuscript books. The goal of the paleographical method is to date and to localize books produced by hand through the analysis of their ancient scripts. This chapter describes the System for Paleographic Inspections (SPI) software suite developed at the University of Pisa, how the system has been used by paleographers in their attempts to classify and identify scripts, and how it can be improved further to meet the research needs of paleographers.*

Keywords. *Paleographic Analysis, Segmentation, Tangent Distance, SPI*

Introduction

What allow paleographers to locate handwritten material, being it entire books or tiny snippets of handwriting, within a particular chronological interval and within a specific cultural environment, are many different aspects of the book (e.g. various crafts involved in its manufacture, notes made by scribes or illuminators, subject specific indications, history of provenance and textual tradition, decoration). However, a crucial role is played by the graphic comparison of the manuscripts under study with similar witnesses, which are dated or datable, and localized. It deals with a method based on formal and stylistic comparison between what has reached us with a defined paternity (or date) and what can supposedly be brought back to the same paternity (or date).[3] *It is this comparison and the particular observation of graphic occurrences in context that enables the paleographer to discover pertinent similarities. It is the same comparative method that reveals unconsciously individual graphic styles, which are indispensable clues to establish the identity or non-identity of hands.*

---

[1] *Corresponding Author: Fabio Aiolli, Dept. of Pure and Applied Mathematics, Via Trieste 63, 35121 Padova, Italy; E-mail: aioli@math.unipd.it*

[2] *For an introduction to paleography see [1].*

[3] *Free translation from the Italian "confronto formale e stilistico fra quanto ci è giunto con paternità (o data) vicina e quanto si presume possa essere ricondotto alla stessa paternità (o data)" in [2], pag. 18.*

Consequentially, while carrying on an analysis of scripts, the relationship between writing styles is crucial for paleographers to infer spatio-temporal locations of documents especially when there is no other clear evidence about the origin of a manuscript. Often, in these cases, experts disagree on which morphological features should be considered relevant and different experts may have different types of expertise and experience.

The System for Paleographic Inspections (SPI) is a software that was developed by Aiolli [3,4] during its Master Degree thesis under the supervision of Prof. Sperduti and Prof. Starita at the Computer Science department of the University of Pisa in the 1999, in collaboration with Dr. Zaccagnini of the department of Medieval Studies at the University of Pisa. The software was the result of a preliminary study of several years.[4]

The aim behind the development of SPI was to provide an adaptive support system, built on rigorous statistical techniques, that, on the basis of old manuscript images, generates models of different writing styles. These models are in turn used to relate 'new' unseen documents with the ones already processed by the system. The idea is to support the paleographical research by facilitating a rigorous analytical approach, where quantitative comparisons can be used to confirm or discard hypotheses concerned with dating and localization.

Recently, SPI has been successfully tested on a recent corpus of manuscripts held at the public library in Siena (Italy). The software is currently under revision and redevelopment so as to become multiplatform and open source, to provide image processing filters for ancient and damaged manuscript images, to include a markup model to wrap annotations that can be then searched semantically, and to enrich the visualization capability of its internal comparative tools.

In Section 1, a brief survey of the SPI system is given, its three main modules, namely, the segmentation, the learning, and the analysis modules are described in some detail. In Section 2, the application of the system on a real paleographical study performed at the University of Siena is described together with its methodological implications. Finally, in Section 3, we report on the current development of the system and suggest new research directions in order to provide paleographers with an even more sophisticated tool.

1. SPI (System for Paleographic Inspections)

SPI consists basically of three modules or tools:

- a Segmentation Tool, exploiting image processing techniques that is able to extract letters from a digitalized manuscript;
- a Learning Tool, where letter-based prototypes based on tangent distance are generated (these models provides an approximate representation of the handwriting style identifiable in a given manuscript image);
- an Analysis Tool, where a paleographer can classify and study the relationships between handwriting styles and sets of manuscript images.

[4] Previous work was devoted to study the applicability of pattern recognition techniques to palaeographic analysis (e.g., [5], [6], [7]).

using various algorithms for the classification of ovarian tumors as benign or malignant. The variables in the data set are related to the personal and family history of ovarian and breast cancer, clinical and demographical data, grey scale and color Doppler results (i.e. around 40 morphologic and blood flow characteristics describing the tumor), and also included presence or absence of pain during the ultrasound examination. A more detailed account of data collection and inclusion criteria is given in [10]. Descriptive statistics of some important variables are shown in Table 1.

The basic model is a logistic regression (LR) model containing 12 inputs [10]: personal history of ovarian cancer (1; binary, coded as 1 versus 0), current use of hormonal therapy (2; binary), pain during examination (3; binary), presence of ascites (4; binary), presence of blood flow within papillary projections (5; binary), presence of a purely solid tumor (6; binary), irregular internal cyst walls (7; binary), presence of acoustic shadows (8; binary), age in years (9), maximum diameter of the lesion in mm (10), maximum diameter of the largest solid component in mm (11; bounded at 50), and the color score of intratumoral blood flow (12; ordinal with values 1 to 4: no, minimal, moderately strong, or very strong flow). The model predicts the probability of malignancy as $\frac{1}{1+e^{-z}}$, with

$$\begin{aligned} z = & -6.7468 + 1.5985(1) - 0.9983(2) - 0.8577(3) + 1.5513(4) \\ & +1.1737(5) + 0.9281(6) + 1.1421(7) - 2.3550(8) \\ & +0.4916(9) + 0.0326(10) + 0.00841(11) + 0.0496(12). \end{aligned} \tag{1}$$

Several more advanced algorithms have been applied since, many of which with a Bayesian foundation. Techniques used include multi-layer perceptrons (MLPs) [11], least squares support vector machines (LS-SVMs) [12], and Bayesian networks [13].

The aim of this chapter is to present the Bayesian classification models for ovarian tumor diagnosis. The outline of this chapter is as follows. First, the Bayesian approach to statistical analysis is shortly introduced in very general terms. Then, Section 2 presents the Bayesian MLP and LS-SVM models for the diagnosis of ovarian tumors. Section 3 presents the application of a recently developed rule extraction system to a Bayesian MLP model. Section 4 presents a Bayesian network using a genetic algorithm to incorporate variable cost in the input selection analysis. Section 5 investigates whether the widely used tumor marker CA-125 is an indispensable input for ovarian tumor diagnostic models. Section 6 concludes the chapter by discussing the major findings.

1. Short introduction to Bayesian analysis

Bayesian analysis is based on Bayes' theorem, in which prior information relating to a model's parameters (the prior distribution) is combined with information from collected data to yield the posterior distribution. A short introduction can be found in [14]. Generally, a Bayesian analysis proceeds as follows. After we have chosen a suitable model M for the problem at hand, a prior probability distribution on the model parameters θ is defined, $p(\theta|M)$. This prior reflects our prior knowledge and/or beliefs concerning likely values for the parameters. Next, we collect data (D) and compute the likelihood of the observed data assuming different values for the model parameters. This is the likelihood function $p(D|\theta, M)$, which is computed using the chosen model with its specific

Table 1. Descriptive statistics for a selection of IOTA variables, with indication of the classification models in which the variables were used as inputs.

Variable	Benign (n=800)	Malignant (n=266)	LR	BMLP 11-2a	BMLP 11-2b	BLS-SVMs
Continuous	*Median*	*Median*				
Age, years	42	56	×	×	×	×
Max. diameter ovary, mm	61	100				×
Max. diameter lesion, mm	63	100.5	×	×	×	
Max. diam. solid component, mm†	0	46.5	×	×	×	×
CA-125, U/ml‡	17	167				
Ordinal	*Mean*	*Mean*				
Number of papillations, 0-4	0.38	1.38		×		
Color score of tumoral blood flow, 1-4	2.12	3.15	×	×		×§
Binary (0 vs 1)	*%*	*%*				
Unilocular tumor	38.9	0.8			×	
Multilocular-solid tumor	21.0	43.6		×		×
Purely solid tumor	6.5	31.6	×	×	×	
Tumor of suspected ovarian origin	81.9	83.5				×
Personal history of ovarian cancer	0.8	3.0	×		×	×
Ascites	2.9	42.1	×	×	×	×
Blood flow within papillations†	6.5	38.3	×	×	×	×
Irregular internal cyst walls	32.8	81.6	×	×	×	×
Acoustic shadows	13.0	1.5	×		×	×
Current use of hormonal therapy	23.5	17.7	×	×		×
Pelvic pain during examination	28.8	19.6	×		×	

†If not solid component/papillation was observed, value 0 was given

‡CA-125 was available for only 567 and 242 women with a benign and malignant tumor, respectively

§In these models, color score is used as a binary variable (levels 1-3 vs level 4)

underlying assumptions. The posterior distribution is the result of multiplying the prior distribution with the likelihood function, divided by a normalization factor to ensure that the posterior is a proper probability distribution:

$$p(\theta|D,M) = \frac{p(D|\theta,M)p(\theta|M)}{p(D|M)}. \tag{2}$$

One advantage of a Bayesian approach is that uncertainty regarding the true value of model parameters is incorporated: rather than looking for a point estimate, a probability distribution is generated. The fundamental difference between Bayesian and traditional statistics is the view of probability. In the Bayesian framework, probability is seen as a degree of belief, whereas it has a frequentist interpretation within the context of traditional statistics. Other advantages are that hyperparameters can be incorporated in the model such that they do not need to be optimized using cross-validation or similar techniques, and that regularization of the model can be incorporated automatically. Further in the chapter, we do not have the possibility to elaborate at length on the spe-

cific Bayesian aspects of different algorithms, but we refer to more detailed descriptions where necessary.

2. Multi-layer Perceptrons and Least Squares Support Vector Machines

As an extension to logistic regression, diagnostic models using MLPs and LS-SVMs were developed to investigate whether these flexible algorithms would lead to better predictive performance.

2.1. Bayesian Multi-layer Perceptrons

A two-layer feed-forward MLP links an input vector $\mathbf{x}$ of size q to an output Y through connections with the k neurons in the hidden layer (i.e. the hidden neurons). Here, the output has value 0 for a benign tumor and 1 for a malignant tumor. The activation of a hidden neuron, h_k, is a linear combination of the inputs sent through a transfer function:

$$h_k = f\Big(\mathbf{w}_k^T \mathbf{x} + b_k\Big). \tag{3}$$

For the ovarian tumor models, the widely used tanh transfer function is used for f. The output unit activation y is computed in a similar fashion, but it is based on a linear combination of the h_k's:

$$y = g\big(\mathbf{w}^T \mathbf{h} + b\big), \tag{4}$$

with $\mathbf{h}$ the vector of hidden neuron activations. For binary classification problems the logistic sigmoidal function is typically used for g. This ensures that the output activation lies in the $[0, 1]$ interval such that it can be interpreted as the estimated probability of malignancy given $\mathbf{x}$, $P(Y = 1|\mathbf{x})$. An estimate for the parameter vector θ, which consists of $\mathbf{w}_k$, b_k, $\mathbf{w}$, and b, is obtained by optimizing the cross-entropy error function (the negative log-likelihood of the data when g is the logistic output function, which corresponds to a Bernoulli output distribution) augmented with the regularization term $\frac{\alpha}{2}\theta^T\theta$ to keep the parameter estimates small in order to avoid overfitting. The amount of regularization is controlled by the regularization parameter α.

In a Bayesian approach, the posterior distribution $p(\theta|D)$ is sought (we omit the conditioning on M for convenience). The estimated probability of malignancy, then, is obtained by averaging over (or integrating out) the posterior distribution:

$$p(Y|\mathbf{x}, D) = \int p(Y|\mathbf{x}, \theta) p(\theta|D) d\theta. \tag{5}$$

This procedure, however, requires solving complex integrals that often do not have a closed form solution. For the ovarian tumor diagnostic models, the evidence procedure was used [15], which is a Bayesian method that approximates the posterior distribution by a Gaussian. It also optimizes hyperparameters (such as α) rather than integrating them out. Therefore, it is not a fully Bayesian technique such as Markov chain Monte Carlo

or variational methods. The prior distribution is taken to be Gaussian with mean zero and variance α^{-1}. The hyperparameter α is a regularization parameter because larger values make the prior more sharply peaked around zero, thus favoring smaller values for the model parameters. In this work, consistent priors were used, meaning that different regularization parameters were used for $\mathbf{w}_k$, b_k, $\mathbf{w}$, and b. In the evidence procedure, the most probable model parameter values θ_{MP} are found by maximizing the (Gaussian) posterior or, similarly, maximizing the product of the likelihood function and the prior. After taking the negative logarithm, this boils down to minimizing the regularized cross-entropy function mentioned above. Because hyperparameters are optimized in the evidence procedure, the formula to obtain the posterior for θ,

$$p(\theta|D) = \int\int p(\theta|\alpha, D)p(\alpha|D)d\alpha. \tag{6}$$

reduces to

$$p(\theta|D) \approx p(\theta|\alpha_{MP}, D). \tag{7}$$

Model selection can be incorporated in the Bayesian framework by specifying a separate hyperparameter α_i for each input i's weights (i.e. the connections between the input and the hidden neurons). When the most probable value for an α_i is small, large weights are allowed for that input indicating that the input may be important to predict the outcome. In this way, the inputs can be ranked from most to least important. To obtain a sensible ranking, continuous inputs were rescaled into the $[-1,+1]$ interval and binary inputs were coded as -1 versus $+1$. The ARD model was fitted ten times using different initial values for the model hyperparameters, and the input with the worst median ranking was dropped. This process was repeated until three inputs remained. This procedure resulted in a final ranking of all inputs, which was then used to determine how many of the most important inputs were to be used in the final model (cf. infra).

The evidence procedure has been criticized and mainly appears to give satisfactory results for medium- to large-sized data sets [16]. Notwithstanding this, good results have been reported [17].

2.2. Bayesian Least Squares SVMs

Whereas the training of standard support vector machines (SVMs) represents a quadratic programming problem, the least squares variant (LS-SVMs) is trained by solving a linear system [18]. For (LS-)SVMs, the input space is mapped into a high-dimensional feature space using mapping $\varphi : \mathbb{R}^q \rightarrow \mathbb{R}^r$. In the feature space, a linear separation between both classes is created by finding a balance between maximization of the margin between both classes (this corresponds to regularization) and minimization of the number of misclassifications (Figure 2). Coding the output as -1 versus $+1$, the classifier $y(\mathbf{x}) = \text{sign}\big[\mathbf{w}^T\varphi(\mathbf{x}) + b\big]$ is obtained by minimizing the cost function

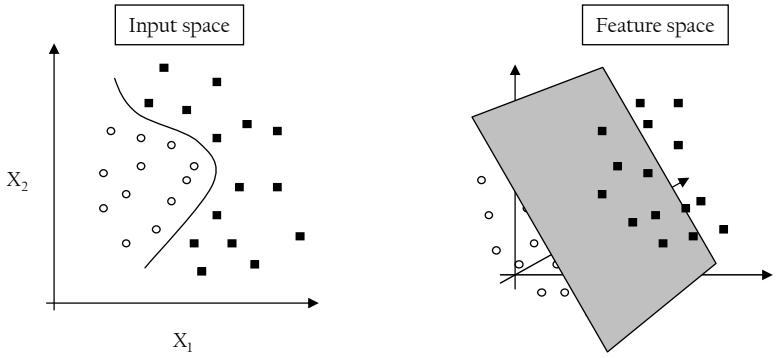


Figure 2. Graphical representation of the underlying rationale for support vector machines.

$$\min_{\mathbf{w},b,\mathbf{e}} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}\gamma\sum_{n=1}^{N} e_n^2 \tag{8}$$

$$\text{such that} \quad y_n\Big[\mathbf{w}^T\varphi(\mathbf{x}_n) + b\Big] = 1 - e_n,\ n = 1, ..., N$$

with y_n the outcome of case n, $\mathbf{w}$ the weight vector of length r, b the bias term, e_n the error variable, and γ the regularization hyperparameter. After taking the Lagrangian of the cost function, the LS-SVM classifier can be reformulated as $y(\mathbf{x}) = \text{sign}\big[\sum_{n=1}^{N}\alpha_n y_n K\big(\mathbf{x},\mathbf{x}_n\big) + b\big]$, with $\alpha_1, ..., \alpha_N$ the support values for the N training cases and $K(\cdot,\cdot)$ a kernel function. This reformulation allows us to work in the feature space without explicitly constructing it by using a positive definite kernel $K\big(\mathbf{x},\mathbf{x}_n\big) = \varphi\big(\mathbf{x}\big)^T\varphi\big(\mathbf{x}_n\big)$. The choice of kernel affects how the linear separation in the feature space relates to the input space. Here, the linear kernel $\mathbf{x}^T\mathbf{x}_n$ was used for constructing a linear classifier in the input space and the radial basis function (RBF) kernel $K\big(\mathbf{x},\mathbf{x}_n\big) = \exp\big(-\|\mathbf{x}-\mathbf{x}_n\|^2\big)/\sigma^2$ for a nonlinear classifier. The support values and the bias term are found by solving a linear Karush-Kuhn-Tucker system.

Standard LS-SVMs do not have the sparseness property of SVMs, where many support values turn out to be zero such that these cases are not used in the classifier. In LS-SVM models, typically no support value will be zero due to the 2-norm in the second term of the cost function (Eq. (8)). For easy cases, support values can be negative [20]. Therefore, a post hoc sparseness procedure was applied in this study by repeatedly pruning cases with negative support values [20].

A disadvantage of (LS-)SVM classifiers is that they do not provide class probabilities. Applying a Bayesian framework to LS-SVMs can overcome this drawback. The Bayesian approach to LS-SVMs in [19] also uses the evidence procedure: the posterior distribution is approximated by a Gaussian around the mode that represents the most probable values $\mathbf{w}_{MP}$. Hyperparameters such as γ and σ (in case of an RBF kernel) are also optimized rather than integrated out. For the Bayesian LS-SVM, we work with a slightly modified cost function:

$$\min_{\mathbf{w},b,\mathbf{e}} = \frac{\mu}{2}\mathbf{w}^T\mathbf{w} + \frac{\zeta}{2}\sum_{n=1}^{N} e_n^2, \tag{9}$$

such that $\gamma = \zeta/\mu$. We can look at the Bayesian approach as a hierarchical method with three levels. On the lowest level, **w** and b are of interest. The prior distribution is taken to be multivariate normal, where the prior for **w** is Gaussian with mean 0 and variance μ^{-1}, and the prior for b is Gaussian with mean 0 and variance σ_b^2 with $\sigma_b^2 \to \infty$ to approximate a uniform distribution. Applying Bayes' theorem results in $\mathbf{w}_{MP}$ and b_{MP}. Similar to the Bayesian MLP, obtaining the most probable values boils down to solving an LS-SVM model. On the second level, the most probable values for μ and ζ are obtained using a uniform prior on log(μ) and log(ζ). The third level deals with model selection. When the RBF kernel is used, this level involves the update of σ using the model evidence $p(D|M)$, which is proportional to $p(M|D)$ since the prior $p(M)$ is taken to be uniform. The final output of a Bayesian LS-SVM is a class probability obtained by integrating over the posterior distribution for **w** and b using the prior class probabilities. These prior probabilities are often taken to be the proportion of cases from each class in the training data set.

Input selection for the Bayesian LS-SVMs to predict ovarian tumor malignancy was performed using a forward selection strategy based on the model evidence [12]. Because the forward selection method is known to be greedy, some inputs were dropped based on knowledge of the subjectivity and accuracy of variables, knowledge of associations between variables, and by checking which variables least decrease the model evidence.

2.3. Experimental Setup

The IOTA data set was split up in a training data set containing 754 tumors (71%) and a test data set containing 312 tumors. This split was stratified for outcome and center. Model development was done using the training set. The test set was only used to independently evaluate model performance.

For the Bayesian MLP, the ARD input ranking analyses used ten hidden neurons to allow for possible nonlinearity (further referred to as ARD10). Using five-fold cross-validation (5CV), the number of hidden neurons and the number of most important inputs to be used in the final model were tuned (i.e. the network architecture). The criteria of interest were the average validation area under the receiver operating characteristic (ROC) curve (AUC) and the average validation cross-entropy error.

Concerning the Bayesian LS-SVMs, two models were built: one with the linear and one with the RBF kernel. Input selection was performed with both types of kernel, but with the aim of selecting one set of inputs for use in both models.

Model evaluation on the test set was based on the AUC and the true positive rate at a true negative rate of 0.75 (Sens75). The AUC can be interpreted as the probability that the model correctly identifies the malignant tumor when confronted with one randomly chosen benign and one randomly chosen malignant tumor. If we define N_b and N_m as the number of benign and malignant tumors in the test set, the AUC is computed as $\frac{1}{N_b N_m}\sum_{n_b=1}^{N_b}\sum_{n_m=1}^{N_m} c_{n_b n_m}$, where $c_{n_b n_m}$ has value 1 if the probability of malignancy is largest for the malignant tumor, 0 if it is largest for the benign tumor, and 0.5 if there is no difference.

2.4. Results and Conclusions

2.4.1. Input Selection, Hidden Neurons, Support Vectors

Regardless of the number of inputs, a model using only two hidden neurons appeared optimal. Focusing on this hidden layer size, 20 runs of 5CV (R5CV) were used to select the number of inputs. This analysis suggested that the eleven most important inputs were to be used (Table 1). The final model, called BMLP11-2a, was obtained by fitting a Bayesian MLP with the selected architecture to the entire training data set.

Because two hidden neurons was suggested to be the optimal choice, the ARD analyses were repeated using two instead of ten hidden neurons (ARD2). Again, the number of inputs was selected using R5CV. The eleven most important inputs according to ARD2 were selected (see Table 1). Fitting a Bayesian MLP with this architecture to the entire training set resulted in model BMLP11-2b.

For the LS-SVM models, the model evidence favored the input selection results based on the linear kernel, so we will not elaborate on the RBF-based input selection results. Eighteen inputs were selected, of which four were again removed based on input variable knowledge as mentioned above. Next, a backward elimination procedure pointed at two variables whose elimination did not decrease model evidence, such that twelve inputs were selected for the Bayesian LS-SVMs (see Table 1).

The Bayesian LS-SVM models were obtained by training the model, using the 12 selected inputs, on the entire training data set. The models are labeled BLS-SVMlin and BLS-SVMrbf. The former has 405 support vectors (54%) whereas the latter has 356 (47%).

2.4.2. Test Set Performance and Conclusions

Table 2 presents the performance of the models on the test set. The ROC curves are shown in Figure 3. All models have an AUC between 0.93 and 0.95, suggesting very good discrimination between benign and malignant tumors. The 95% confidence intervals suggest that the Bayesian models' AUCs do not differ much from the AUC of the basic logistic regression model. At 75% specificity, all models achieve between 92 and 96% sensitivity.

It is clear that the mathematical models can diagnose ovarian masses very well, with AUCs up to 0.95. The basic logistic regression model had very good performance. The Bayesian LS-SVM models performed slightly better, but the differences were very small. A disadvantage of the latter models is the set of inputs that is used. Whether the tumor is thought to be of ovarian origin or not (i.e. para-ovarian or tubal) is very subjective. Also, the maximum diameter of the ovary is clinically not a logical input. Its inclusion can be understood, however, by noting that it is highly related to the maximum diameter of the lesion (i.e. the mass). The CA-125 tumor marker was not considered as a possible input for the models. This is an important clinical remark because the use of CA-125 is controversial. We come back to this later.

The observation that the logistic regression model and SVMlin performed very well together with the fact that only two hidden neurons were selected for the Bayesian MLP models suggest that ovarian tumor diagnosis is a classification task with a low degree of nonlinearity.

Table 2. Test set results of the diagnostic models on based logistic regression, Bayesian MLPs and Bayesian LS-SVMs.

Model	AUC	AUC diff. with LR (with 95% CI†)	Sens75	Sens75 diff. with LR (with 95% CI†)
LR	0.942		0.933	
BMLP11-2a	0.942	$0.000\ (-0.013; 0.013)$	0.920	$-0.013\ (-0.071; 0.039)$
BMLP11-2b	0.933	$-0.009\ (-0.027; 0.004)$	0.920	$-0.013\ (-0.071; -0.039)$
BLS-SVMlin	0.946	$0.004\ (-0.008; 0.014)$	0.960	$0.027\ (-0.029; 0.094)$
BLS-SVMrbf	0.945	$0.003\ (-0.009; 0.013)$	0.947	$0.014\ (-0.033; 0.072)$

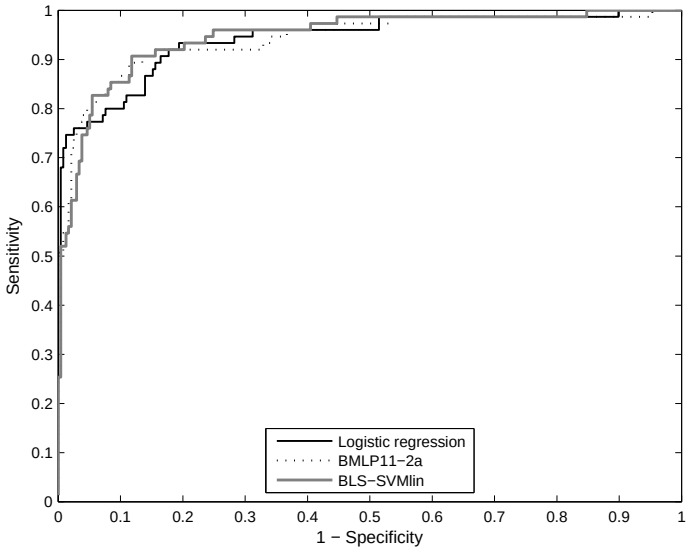


Figure 3. Test set ROC curves.

3. Orthogonal Search-based Rule Extraction

For models with numerical output such as probabilities, a threshold value can be chosen in order to arrive at crisp predictions of tumors as either benign or malignant. If one uses q inputs to generate the probabilities, the use of a threshold value means that a decision boundary is defined in q-dimensional space to separate regions with a different crisp prediction. However, in advanced models such as LS-SVMs or MLPs, the decision boundary can be highly nonlinear such that it is often not clear how the inputs are used to yield class probabilities (i.e. black-box models). Rule extraction methods have been developed in order to gain insight in the operation of a specific model and its decision boundary. One such method is called orthogonal search-based rule extraction (OSRE) [21], and this method was applied to BMLP11-2a [22].

OSRE automatically extracts low order rules using the decision boundary and the training data that were used to derive the model. For each training data point that is predicted to be malignant, the algorithm searches in orthogonal directions for hypercubes spanning the part of the q-dimensional data space for which the model makes the pre-

Table 3. OSRE rules for BMLP11-2a.

Rule	Conjunctive conditions	Rule	Conjunctive Conditions
1	Irregular cyst walls	4	Irregular cyst walls
	Color score 4		Purely solid tumor
	$51.6 \leq$ Age ≤ 93.5		$49.7 \leq$ Age ≤ 92.3
2	No hormonal therapy	5	No hormonal therapy
	Irregular cyst walls		At least 2 papillations
	Color score 4		Color score 3 or 4
	$108 \leq$ Max. diameter lesion ≤ 403		$28.7 \leq$ Max. diameter solid comp. ≤ 224.2
3	At least 4 papillations		
	Blood flow within papillations		
	Color score 3 or 4		
	$59 \leq$ Max. diameter lesion ≤ 401		
	$19.9 \leq$ Max. diameter solid comp. ≤ 227		

diction of malignancy. The limits or size of the hypercube is determined by the decision boundary or by the extremes of the data space. The hypercube represents a rule, consisting of a set of conjunctive conditions expressed by the boundary values for each input. There are as many rules as there are cases that are predicted to be malignant. The full (disjunctive) set of rules is pruned using some criterion. In the application of OSRE to BMLP11-2a (using a threshold probability of 0.15), rule pruning was performed by maximizing the positive predictive value of the set of rules. This lead to rule sets with very high specificity but low to moderate sensitivity. The final set of rules is listed in Table 3. The set of five rules for BMLP11-2a had a sensitivity of 52.6% and a specificity of 99.6%. The positive and negative predictive values were 97.9% and 86.3%. This set of rules is careful in predicting malignancy, but if malignancy is predicted, this is nearly always correct.

It is clear that the extracted rules are not simple or polished. Therefore, their main use is the clarification of the operation of a model. Clinicians may also be interested in a short list of easy-to-use rules that they can apply directly when performing an ultrasound examination of an adnexal mass. Such rules have been derived on the IOTA phase 1 data [23]. The procedure used to extract the rules from Table 3 has lead to a set of disjunctive rules with near perfect specificity and PPV. Thus, these rules can help to detect cases for which one can be highly confident that they are malignant.

4. Controlling Input Cost Using a Genetic Algorithm for Bayesian Networks

A very interesting issue that can have a clear positive impact on clinical practice is whether we can develop well-performing models for which the cost to measure the inputs is low. The cost of an input reflects its subjectivity, measurement accuracy, financial cost, and time cost. A typical example of a low cost input is the age of the patient: it is objective, accurate, and requires time nor money. Inputs derived from the Doppler flow velocity waveforms to measure intratumoral blood flow, on the contrary, have a much

higher cost. Models with low input cost are cheaper and easier to implement, and may be more robust. Gevaert et al. [13] examined whether input selection favoring a low total input cost would result in models that perform similar to models based on unconstrained input selection. To this end, Bayesian networks were used with input selection based on a genetic algorithm.

4.1. Bayesian Networks

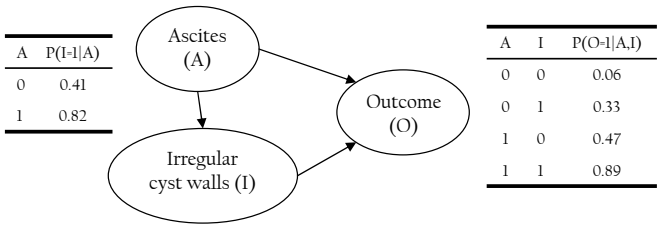
A Bayesian network consists of a network structure and of local probability models [24,25]. The network structure is a directed acyclic graph where the nodes in the graph represent the inputs and the edges between nodes represent dependencies between inputs. The set of parents of input x_i is denoted as $\mathbf{a}_i$. The local probability models specify how $\mathbf{a}_i$ influences x_i. Different kinds of local probability models exist, depending on the nature of the inputs (e.g. discrete or continuous). In this work, the focus was on discrete-valued Bayesian networks since many IOTA variables are discrete (see Figure 4 for an example). The local probability models were represented by conditional probability tables (CPTs), which specify the probability that an input takes a certain value given the value of its parents. Finally, note that a Bayesian network structure implicitly constrains the ordering of the inputs since a directed edge from x_i to x_j is only allowed if x_i precedes x_j in the input order. Using the chain rule of probability, a Bayesian network can thus be represented as:

$$p(\mathbf{x}) = \prod_{i=1}^{q} p\big(x_i|\mathbf{a}_i\big). \tag{10}$$

Building a Bayesian network requires learning the structure and the parameters of the CPTs. Here, structure learning was based on the K2 search algorithm [24]. K2 generates new structures that are evaluated by the so-called Bayesian Dirichlet score metric [24]. The K2 search algorithm uses a pre-specified input order, because this constrains the number of networks and hence the search space. The input order for the K2 algorithm was generated by the genetic algorithm which also dealt with input selection. After structure learning, the parameters of the CPTs have to be learned. For each input x_i and each instantiation of its parents, there is a CPT whose parameter vector θ contains the probabilities for each value of x_i. The prior distribution for θ was the uniform Dirichlet distribution, and the likelihood function assumed a multinomial distribution for the input. This results in a Dirichlet posterior distribution for θ. Finally, the technique of 'probability propagation in trees of clusters' [26] was used to obtain the estimated probability of malignancy based on the learned Bayesian network.

4.2. Genetic Algorithm for Input Selection and Input Order

Genetic algorithms (GAs) mimic evolutionary processes in biology to tackle optimization problems. A GA starts with a set of solutions that is called the population. Here, solutions consisted of a set of inputs with a particular order. Based on a fitness measure, the best solutions are selected to create new solutions (offspring) using GA operators. The new solutions replace bad solutions in the population such that a new population is created (the next generation) that is hoped to be better than the previous one. Thus,



A	P(I=1|A)
0	0.41
1	0.82

A	I	P(O=1|A,I)
0	0	0.06
0	1	0.33
1	0	0.47
1	1	0.89

Figure 4. Simple fictitious example of a Bayesian network.

the GA's goal is to naturally evolve to more fit solutions. An application of GAs in the context of head and neck squamous cell carcinoma can be found in [4].

For each solution in the population, a Bayesian network was developed and its fitness evaluated by the AUC penalized by the total input cost of the model. The fittest models were selected in pairs, and new solutions were created by implementing the cross-over and mutation operators on each pair. The cross-over operator randomly selected a cross-over point on both solutions' input orders, and combined the input order before the cross-over point of one solution with the input order after the cross-over point of the other solution. Then, both new solutions were subjected to a mutation operator, which mutated a solution with probability 0.05. The specific mutation employed was randomly chosen, and could be the addition or removal of an input, or the change of position of two inputs. The resulting solutions were used to create a new generation.

4.3. Experimental Setup

First, an expert gynecologist discretized the continuous variables into bins and assigned to each variable a cost value between 1 (low cost) and 5 (high cost). The data were splitted into a training and a test part (70-30 ratio).

During input selection, 70% of the training data was used for building Bayesian networks. The remaining 30% was used for fitness evaluation by computing a penalized AUC. The AUC was reduced with $0.003(C-4)$, with C the total input cost and 4 the lower bound of C as the minimum input set size was set at 4. The GA started with a population of 100 randomly chosen solutions, and ran until 1,000 new generations were created. The fittest solution among the last ten generations was used to learn a Bayesian network on the entire training set. This network was then applied to the test set to assess its performance using the AUC. This procedure was repeated 100 times, once with cost optimization (i.e. with a fitness measure that penalizes for the total input cost) and once without. The procedures with and without cost optimization were compared in terms of average total input cost and average test set AUC over the 100 runs. For each of the 100 runs, the part of the training data to be used for fitness evaluation was randomly selected.

4.4. Results and Conclusions

When input cost was not optimized, the total input cost was on average 34 with an average of 14 selected inputs. When incorporating input cost in the GA, the total input cost was on average 20 with an average of 7 selected inputs. Incorporating input cost resulted in the selection of fewer inputs and the avoidance of some high cost inputs. A

good example is the tumor marker CA-125, which is the input with highest cost. When input cost was ignored, this input was selected in 97 out of 100 runs. When input cost was optimized, however, this input was not selected at all.

The average test set AUC was 0.966 when input cost was ignored, and decreases to 0.958 when the input cost was optimized. This decrease in performance can be considered smaller than the gain in total input cost.

The analyses suggest that total input cost can be reduced without substantial decrease in model performance. This is an important observation for obvious reasons.

Cost optimization worked partly as a mechanism to select fewer inputs, which indirectly had an impact on the total input cost. Yet, it were mainly the high cost inputs that were avoided, such that the algorithm also reduced input cost in a direct manner. Finally, note that the test set AUCs obtained in this analysis cannot be compared with those obtained in Section 2 because they were based on another training-test split. Also, not all 1,066 patients were used because CA-125 information was not always available. This issue will be elaborated on in the next section.

5. The Necessity of CA-125 in Ovarian Tumor Diagnostic Models

CA-125 is a widely used tumor marker for ovarian cancer. Keeping in mind that it is a measurement with high cost, it is an important question whether CA-125 is indispensable in diagnostic models for ovarian tumors. Existing literature suggests that CA-125 contains a lot of information. In the previous section, for example, CA-125 was selected as an input in 97 out of 100 runs without input cost optimization. Yet, CA-125 may not be necessary as it was not selected when input cost was optimized. However, the experimental setup hampers strong conclusions concerning CA-125. The analyses are based on the complete cases. The IOTA data set is a very complete data set, but the measurement of CA-125 was not obligatory. As a result, about 25% of the patients have no CA-125 information, and these are ignored in complete case analyses. Some gynecologists who participated in the IOTA study always did (or did not) measure CA-125, while others measured CA-125 less often when the tumor looked clearly benign on ultrasound examination. Hence, caution is due when interpreting the results involving CA-125.

A separate IOTA project focused on the importance of CA-125 for building diagnostic models, using various imputation techniques for the missing CA-125 values such that all patients could be included in the analysis [27]. Models were built that either included or excluded CA-125 to investigate whether excluding CA-125 would lead to a decrease in diagnostic performance.

5.1. Experimental Setup

First, missing CA-125 values were imputed using four different methods: regression imputation (which is a type of conditional mean imputation), expectation-maximization, data augmentation, and nearest neighbor hot-deck. Five situations were considered: four imputation situations depending on the imputation method used, and a fifth situation in which CA-125 was completely ignored. Using the original training set, which was also used for the models in Section 2, a selection of 20 (or 19 in the fifth situation) important inputs were ranked with the ARD algorithm described earlier. ARD input ranking

was performed separately for each situation, and in the first four situations CA-125 was ranked as the most important input by definition.

Next, 100 new random splits of the data into a training and test part were created with stratification for outcome. On each training set and separately for each of the five situations, 18 (or 17 in the fifth situation) Bayesian LS-SVM models with RBF kernel were developed: the first model contained the three most important inputs, the second model contained the four most important inputs, and so on. This implies that, in the first four situations, all models included CA-125 as an input. Models were evaluated on the accompanying test data set, and were evaluated by the AUC. The resulting 100 AUCs were summarized by their mean.

A drawback of this procedure is that, in each situation, one single input ranking was used for each of the 100 training data sets. The reason for this was that repeating the ARD input ranking analysis for each training set was computationally very expensive. In an attempt to overcome this drawback, the ARD input ranking analysis was repeated for 20 training data sets using a Bayesian perceptron model. This model is similar to a Bayesian MLP, but without hidden layer. Therefore, it corresponds to a Bayesian logistic regression model. The ARD analysis is much faster for this kind of model. Note again that, for each training set, an ARD input ranking had to be performed separately for all five situations.

5.2. Results and Conclusions

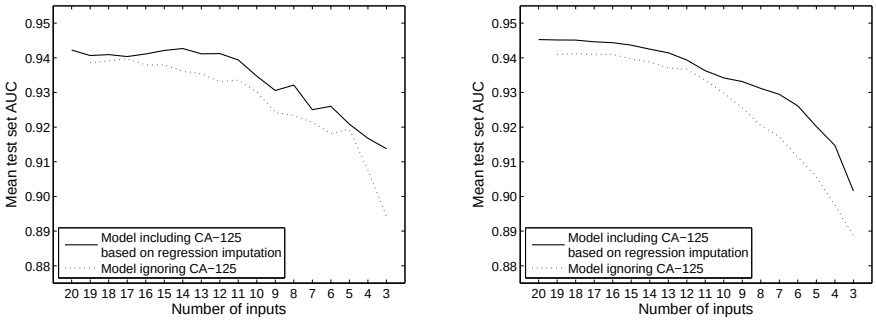
The results of the Bayesian LS-SVMs based on a single input ranking showed that the inclusion of CA-125 results in a minor improvement in AUC that does not seem to justify the high cost of this measurement (Figure 5(a), which considers only the imputation method with the best AUC results). When three inputs were used, the advantage (measured as the difference in mean test set AUC) of using CA-125 was 0.020. When at least four inputs were used, the advantage was never larger than 0.010.

When repeating the ARD input ranking analysis for 20 training data sets using a Bayesian perceptron model, results are similar (Figure 5(b)). The maximum advantage of using CA-125 was observed when four inputs were used (0.017). When at least nine inputs were used, the advantage was never larger than 0.010.

The results confirm what could be hypothesized from the results in Section 4: CA-125 can be replaced by other inputs in diagnostic models. This is a clinically remarkable and important result, that is further corroborated by results on model building using logistic regression (but without missing value imputation) [28], and by results showing that an expert gynecologist's opinion performs better than CA-125 [29].

6. General Discussion

This chapter describes several crucial results with respect to the diagnosis of ovarian tumors. First of all, mathematical models can predict malignancy with a high degree of accuracy, as shown by the performance of the diagnostic models presented in Section 2. Models that construct a linear separation between both tumor types in the input space (e.g. LR or BLS-SVMlin) performed similarly to more flexible models, suggesting that the level of nonlinearity in this classification problem is limited.



(a) (b)

Figure 5. Mean test set AUCs for diagnostic models with or without CA-125 after regression imputation of missing CA-125 values; (a) using Bayesian LS-SVMs with a single input ranking after 100 runs of repeated data splitting, and (b) using Bayesian perceptrons after 20 runs of repeated data splitting with retraining of input ranking.

The OSRE algorithm for rule extraction, as applied to BMLP11-2a, isolated a set of five disjunctive rules that appear to be able to detect malignancies with very high positive predictive value. The rules provide insight in how BMLP11-2a works, and can help to detect tumors that are almost certainly malignant.

Also, well-performing models with a low total input cost were constructed using Bayesian networks. Thus, it appears possible to achieve good performance with a 'cheap' set of inputs. Moreover, extensive analyses suggest that the controversial and costly measurement of the CA-125 tumor marker is not needed. This information can be replaced by other inputs in diagnostic models, and expert gynecologists clearly outperform CA-125 in separating benign from malignant tumors. These are highly relevant observations, as they may result in improvements of the clinical management of ovarian tumors. Less money and time is needed for diagnosis without loosing diagnostic performance, and the psychological impact for the patient and her environment can be reduced. Such practical improvements are what effective CDS should eventually lead to [1].

Two drawbacks of these analyses need to be mentioned. Firstly, it is observed that the performance of the diagnostic models approaches yet does not exceed that of an expert gynecologist. In new phases of the IOTA study group, it will be examined how the models compare to less experienced gynecologists. Models, however, are more flexible because desired sensitivity and specificity levels can be more easily obtained by varying the probability threshold for predicting malignancy. Secondly, the analyses focused on the binary classification while different types of benign and malignant tumors exist which may not always require the same treatment. To address this limitation, multi-class models are being developed that aim to classify tumors as benign, primary invasive, borderline malignant, or metastatic invasive.

Finally, before CDS systems can be disseminated to clinical practice, their quality needs to be confirmed by prospective studies [1]. A first internal prospective evaluation of models from IOTA phase 1 showed very good performance on completely new data from three centers that also participated in phase 1 [30]. For IOTA phase 2, data have

been collected on nearly 2,000 women from 19 clinical centers. Preliminary analyses suggest excellent prospective performance of the models on data from 7 centers that also participated in phase 1 (internal evaluation), but also on data from 12 completely new centers (external evaluation). These are encouraging results when working towards a successful CDS system for ovarian tumor diagnosis.

Acknowledgements

The authors would like to thank the following researchers for enriching collaborations and discussions: Hane Aung, Ioannis Dimou, Paulo Lisboa, Chuan Lu, Ian Nabney, Johan Suykens, and Michalis Zervakis. Ben Van Calster is a postdoctoral researcher funded by the Research Foundation - Flanders (FWO). Olivier Gevaert is supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). Research supported by Research Council KUL: GOA-AMBioRICS, CoE EF/05/006 Optimization in Engineering (OPTEC); FWO: G.0407.02 (support vector machines), G.0302.07 (SVM), G.0341.07 (Data fusion), research communities (ICCoS, ANMMM); IWT: TBM-IOTA3; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO); EU: BIOPATTERN (FP6-2002-IST 508803).

References

[1] J.C. Wyatt and D.G. Altman, Prognostic models: clinically useful or quickly forgotten? *BMJ* **311** (1995), 1539–1541.

[2] D.F. Sittig, A. Wright, J.A. Osheroff, et al, Grand challenges in clinical decision support, *J Biomed Inform* **41** (2008), 387–392.

[3] A. Micheli, A. Sperduti, A. Starita, An introduction to recursive neural networks and kernel methods for cheminformatics, *Curr Pharm Des* **13** (2007), 1469–1495.

[4] F. Baronti, A. Micheli, A. Passaro, A. Starita, *Machine learning contribution to solve prognostic medical problems*, in A.F.G. Taktak and A.C. Fisher (eds.), Outcome prediction in cancer, Elsevier, Amsterdam, 2007, pp. 261–283.

[5] A. Starita and A. Sperduti, *A neural-based system for the automatic classification and follow-up of diabetic retinopathies*, in P.J.G. Lisboa, E.C. Ifeachor, P.S. Szczepaniak (eds.), Artificial neural networks in biomedicine, Springer, London, 2000, pp. 233–247.

[6] A. Jemal, R. Siegel, E. Ward, et al, Cancer statistics, *CA Cancer J Clin* **58** (2008), 71–96.

[7] M.E. Carley, C.J. Klingele, J.B. Gebhart, et al, Laparoscopy versus laparotomy in the management of benign unilateral adnexal masses. *J Am Assoc Gynecol Laparosc* **9** (2002), 321–326.

[8] I. Vergote, J. De Brabanter, A. Fyles, et al, Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma, *Lancet* **357** (2001), 176–182.

[9] D. Timmerman, L. Valentin, T.H. Bourne, et al, Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) group, *Ultrasound Obstet Gynecol* **16** (2000), 500–505.

[10] D. Timmerman, A.C. Testa, T. Bourne, et al, A logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis (IOTA) group, *J Clin Oncol* **23** (2005), 8794–8801.

[11] B. Van Calster, D. Timmerman, I.T. Nabney, et al, Using Bayesian Neural Networks with ARD input selection to detect malignant ovarian masses prior to surgery, *Neural Comput Appl* **17** (2008), 489–500.

[12] B. Van Calster, D. Timmerman, C. Lu, et al, Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods, *Ultrasound Obstet Gynecol* **29** (2007), 496–504.

[13] O. Gevaert, D. Timmerman, B. De Moor, *Optimizing variable order, selection and cost using a genetic algorithm for modeling ovarian masses with Bayesian networks*, Technical Report 08-18, Dept of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Belgium, 2008.

[14] B. Van Calster, I. Nabney, D. Timmerman, et al, The Bayesian approach: a natural framework for statistical modeling, *Ultrasound Obstet Gynecol* **29** (2007), 485–488.
[15] D.J.C. MacKay, Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks, *Netw - Comput Neural Syst* **6** (1995), 469–505.
[16] F. Vivarelli and C.K.I. Williams, Comparing Bayesian neural network algorithms for classifying segmented outdoor images. *Neural Netw* **14** (2001), 427–437.
[17] I.T. Nabney, D.J. Evans, Y. Brulé, et al, *Assessing the effectiveness of Bayesian feature selection*, in D. Husmeier, R. Dybowski, S. Roberts (eds.), Probabilistic modeling in medical informatics and bioinformatics, Springer, London, 2005, pp. 371–390.
[18] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, et al, *Least squares support vector machines*, World Scientific, Singapore, 2002.
[19] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, et al, Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis, *Neural Comput* **14** (2002), 1115–1147.
[20] C. Lu, T. Van Gestel, J.A.K. Suykens, et al, Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines, *Artif Intell Med* **28** (2003), 281–306.
[21] T.A. Etchells, P.J.G. Lisboa, Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach. *IEEE Trans Neural Netw* **17** (2006), 374–384.
[22] M.S.H. Aung, P.J.G. Lisboa, T.A. Etchells, et al, Comparing analytical decision support models through Boolean rule extraction: a case study of ovarian tumor malignancy, *Lect Notes Comp Sci* **4492** (2007), 1177–1186.
[23] D. Timmerman, A.C. Testa, T. Bourne, et al, Simple ultrasound-based rules for the diagnosis of ovarian cancer, *Ultrasound Obstet Gynecol* **31** (2008), 681–690.
[24] D. Heckerman, D. Geiger, D.M. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data, *Mach Learn* **20** (1995), 197–243.
[25] O. Gevaert, F. De Smet, E. Kirk, et al, Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression, *Hum Reprod* **21** (2006), 1824–1831.
[26] C. Huang, A. Darwiche, Inference in belief networks: a procedural guide, *Int J Approx Reas* **15** (1996), 225–263.
[27] I. Dimou, B. Van Calster, S. Van Huffel, et al, *Evaluation of imputation methods in ovarian tumour diagnostic models using generalised linear models and support vector machines*, submitted, 2008.
[28] D. Timmerman, B. Van Calster, D. Jurkovic, et al, Inclusion of CA-125 does not improve mathematical models developed to distinguish between benign and malignant adnexal tumors, *J Clin Oncol* **25** (2007), 4194–4200.
[29] B. Van Calster, D. Timmerman, T. Bourne, et al, Discrimination between benign and malignant adnexal masses by specialist ultrasound examination versus serum CA-125, *J Natl Cancer Inst* **99** (2007), 1706–1714.
[30] C. Van Holsbeke, B. Van Calster, A.C. Testa, et al, Predicting malignancy of ovarian tumors: prospective evaluation of models from the IOTA study, *Clin Cancer Res*, in press, 2008.