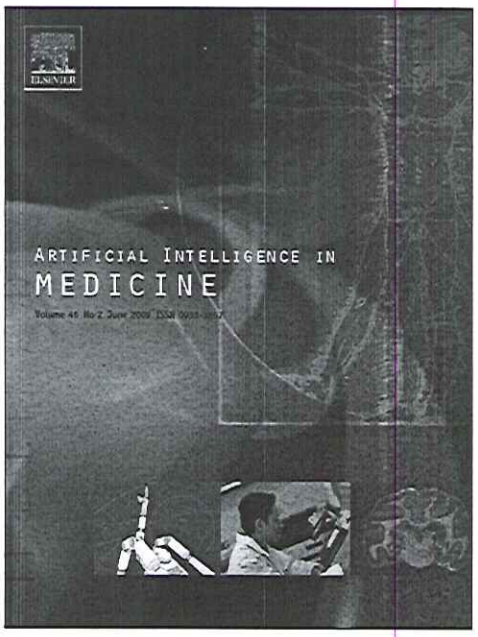
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Artificial Intelligence in Medicine (2009) 46, 139–154

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

An application of methods for the probabilistic three-class classification of pregnancies of unknown location

Ben Van Calster[a,*], George Condous[b], Emma Kirk[c], Tom Bourne[c,d], Dirk Timmerman[d], Sabine Van Huffel[a]

[a] *Department of Electrical Engineering (ESAT-SISTA), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*
[b] *Early Pregnancy, Acute Gynaecology and Advanced Endosurgery Unit, Nepean Hospital, Nepean Clinical School, University of Sydney, Sydney, Australia*
[c] *Imperial College London, Hammersmith Campus, Du Cane Road, London, UK*
[d] *Department of Obstetrics and Gynaecology, University Hospitals Katholieke Universiteit Leuven, Herestraat 49, B-3000 Leuven, Belgium*

Received 5 October 2007; received in revised form 28 November 2008; accepted 1 December 2008

KEYWORDS
Pregnancy of unknown location;
Multi-class classification;
Least squares support vector machines;
Logistic regression;
Kernel logistic regression

Summary

Objective: We developed and compared classifiers to predict the outcome of pregnancies of unknown location (PUL). This is a three-class problem, as the possible outcomes are failing PUL, intra-uterine pregnancy (IUP), or ectopic pregnancy (EP). We focused on probabilistic classification because of the importance of uncertainty information in medical decision making.
Methods and materials: Nine methods were implemented, based on logistic regression (LR), multi-layer perceptrons, least squares support vector machines (LS-SVMs), and kernel logistic regression (KLR). The LR models involved manual checks for the necessity of input transformations or interaction effects. The classifiers were developed on the training set (n = 508) and evaluated on the test set (n = 348). We used two performance measures that only evaluate discriminatory potential, and two that investigate the exact probabilities and/or discriminatory potential. Classifier comparison was carried out using a ranking method.
Results: The classifier based on a combination of binary LR models using pairwise coupling (LR-PC) ranked first or second for each performance measure. LR-PC obtained test set areas under the receiver operating characteristic curve of 0.989, 0.988, and 0.924 for the prediction of failing PULs, IUPs, and EPs, respectively. Multi-class LR, multi-class KLR, and a combination of binary Bayesian LS-SVMs with radial

* Corresponding author. Tel.: +32 16 32 19 25; fax: +32 16 32 19 70.
E-mail address: bvancals@esat.kuleuven.be (B. Van Calster).

0933-3657/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2008.12.003

basis function kernel were always ranked in the top five. Models with low emphasis on nonlinearity were ranked at the bottom. Importantly, LR-PC also performed better than previously developed classifiers based on multi-class LR.
Conclusions: The prediction of PULs was good, most notably for failing PULs and IUPs. Logistic regression models performed remarkably well. Multi-class KLR also performed well, whilst nonlinearity was automatically incorporated and probabilistic outputs were directly given without Bayesian analysis or a combination of binary results. The selected inputs are reasonably objective and easy to obtain in clinical practice. Taken together, this study provided useful decision support tools for implementation in clinical practice.
© 2008 Elsevier B.V. All rights reserved.

1. Introduction

A woman has a pregnancy of unknown location (PUL) when she has a positive pregnancy test in the presence of a negative transvaginal ultrasound scan (TVS) of the pelvis, i.e. the TVS cannot demonstrate the presence of an intra- or extra-uterine pregnancy and there are no signs of miscarriage. PULs are most often failing PULs, but normal intra-uterine pregnancies (IUPs) and ectopic pregnancies (EPs) are also possible outcomes. In this study, we predicted the outcome of PULs using various probabilistic multi-class methods based on logistic regression (LR), multi-layer perceptrons (MLP), least squares support vector machines (LS-SVM), and kernel logistic regression (KLR). The motivation for this study was twofold. Firstly, it was deemed necessary to move beyond existing classifiers for PULs by using techniques originating in the engineering and machine learning communities. Secondly, we were interested in the comparison of several algorithms for probabilistic multi-class classification when applied to a real life medical decision making problem. Both of these issues are now explained in more detail.

Previously, we have predicted the outcome of PULs using multi-class LR (MLR) [1–3]. These studies showed that biochemical markers related to serum human chorionic gonadotrophin (hCG) levels, such as the ratio of hCG levels at 48 and at 0 h (hCG ratio) and the average hCG level, were the key measurements to predict PUL outcome [1,2]. The amount of vaginal bleeding was also useful even though its inclusion only weakly improved the discriminative performance of the model [3]. These models, however, were developed on small data sets using basic statistical methods such as MLR. Even though simple methods already perform very well in many situations [4], it was necessary to apply more advanced mathematical techniques. The LR model is linear in its inputs; nonlinear separation with respect to the original variables can be obtained by manually adding interaction terms, polynomial terms, or transformed versions of a variable to the input set. Advanced techniques such as MLPs [5] or LS-SVMs [6,7] automatically incorporate nonlinearity. Also, in these models it is natural to avoid overfitting of the training data by the use of regularization to keep the estimates of the model parameters small. Traditionally, this is not done often in LR models even though regularized versions exist [8].

Probabilistic multi-class classification is a challenging area of ongoing research. For example, support vector machine (SVM) algorithms are state-of-the-art methods for classification purposes, yet in their standard form they are designed for non-probabilistic binary classification. However, in medical decision making one is often confronted with classification tasks involving more than two outcome classes, and patient classification typically serves the aim of selecting the optimal treatment. In such personalized healthcare settings, accurate probabilistic information regarding the likely nature or outcome of a condition is desired. This information conveys the level of uncertainty with respect to the possible diagnoses, and can thus influence a clinician's decision making process. Therefore, recent research led to machine learning algorithms that yield probabilistic output for multi-class problems. It is useful to compare such algorithms in the context of real life medical decision making problems because such studies can help both clinicians and machine learning researchers move forward [9].

We sought to evaluate the performance of novel probabilistic multi-class algorithms for PUL classification. In this study, we evaluated the following algorithms: standard statistical MLR with manual checks for transformations, interactions, and the linearity in the logit assumption [10], a combination of standard binary LR models with similar checks, a Bayesian MLP [5,11], a Bayesian perceptron, combinations of binary Bayesian LS-SVMs [6,7], and multi-class kernel logistic regression based on LS-SVM theory (MKLR) [12]. An overview of the abbreviations used throughout the paper is given in Table 1.

Table 1 Abbreviations used throughout the manuscript.

Abbreviation	Description
5CV	Fivefold cross-validation
AR	Average rank
ARD	Automatic relevance determination
AUC	Area under the ROC curve
BLSSVMlin-END	Binary Bayesian LS-SVMs with linear kernel combined using END
BLSSVMlin-PC	Binary Bayesian LS-SVMs with linear kernel combined using pairwise coupling
BLSSVMrbf-END	Binary Bayesian LS-SVMs with RBF kernel combined using END
BLSSVMrbf-PC	Binary Bayesian LS-SVMs with RBF kernel combined using pairwise coupling
BMLP	Bayesian multi-layer perceptron
BPER	Bayesian perceptron
END	Ensembles of nested dichotomies
EP	Ectopic pregnancy
hCG	Serum human chorionic gonadotrophin
IUP	Intra-uterine pregnancy
KLR	Kernel logistic regression
LR	Logistic regression
LR-PC	Binary logistic regression models combined using pairwise coupling
LS-SVM	Least squares SVM
M5	The most recent existing classifier [3]
MKLR	Multi-class kernel logistic regression
MLP	Multi-layer perceptron
MLR	Multi-class logistic regression
Pm(κ)	Proportion of times a classifier was ranked amongst the best κ
PUL	Pregnancy of unknown location
RBF	Radial basis function
ROC	Receiver operating characteristic curve
sAUC	Scored area under the ROC curve
SVM	Support vector machine
TVS	Transvaginal ultrasound scan
VUS	Volume under the ROC surface

2. Data

The data consist of women who presented with a PUL at the early pregnancy unit of St. George's Hospital in London (United Kingdom) between June 2001 and October 2004. All women with a positive urinary pregnancy test underwent a TVS. Those women classified as having a PUL also had peripheral blood taken to measure serum hCG and progesterone levels at presentation and 48 h later. Criteria for exclusion from the study population can be found in Refs. [2,3]. All women were followed up with TVS, hCG measurements, and/or laparoscopy until the final clinical status of the pregnancy was established (the outcome). More details can be found in Refs. [2,3].

Fourteen variables were recorded. Firstly, the hCG ratio and the average hCG (IU/L) were computed as described earlier. As a third biochemical index, the average progesterone level (nmol/L) was derived. Both the average hCG and the average progesterone levels were log-transformed by default because they were highly skewed. The woman's age (years) was recorded, as well as the gestational age (days), thickness of the endometrium (mm), character of the midline echo (intact/disrupted), and presence of free fluid in the pouch of Douglas (yes/no). Finally, the following presenting complaints were documented: the level of vaginal bleeding (ordinal variable with values 0 for no vaginal bleeding, 1 for vaginal bleeding without clots, and 2 for vaginal bleeding with clots), the presence of lower abdominal pain, a history of EP (yes/no) or miscarriage (yes/no), unsure date of last menstrual period (yes/no), and maternal anxiety (yes/no). These latter two variables were omitted from further analyses after discussions with the clinicians.

In total, 1003 PULs were included in the data set. Of these, 872 had complete data (87%). Only two variables had more than 1% missing values (gestational age: 83 missing values, 8.3%; average progesterone level: 40 missing values, 4.0%). Likelihood ratio tests suggested that missingness for gestational age was positively related to history of miscarriage and negatively to the presence of lower abdominal pain. Therefore, we decided to continue with the complete cases as bias, if any, was considered to be very small. Sixteen complete cases were persisting PULs and were excluded from the analysis. Of the remaining PULs, 856 were failing PULs (460, 53.7%), 330 were IUPs (38.6%), and 66 were EPs (7.7%). Descriptive statistics are presented in Table 2. The 508 cases examined between June 2001 and July 2003 served as training data (59%): 283 failing PULs, 183 IUPs, and 42 EPs. A part of the training data was used to develop the existing models described in [1–3]. The remaining 348 cases served as test set (41%): 177 failing PULs, 147 IUPs, and 24 EPs.

3. Algorithms, input selection, and model evaluation

We assumed training data points (x_n, y_n), $n = 1, \ldots, N$, with $x_n \in \mathbb{R}^p$ representing the input vector and y_n the ith of k outcome classes (with $k > 2$ and $i = 1$,

Table 2 Descriptive statistics.

Continuous variables	Failing PUL (n = 460)		IUP (n = 330)		EP (n = 66)	
	Median	IQR	Median	IQR	Median	IQR
hCG 0 h (IU/L)	155	456	467	558	531	904
hCG 48 h (IU/L)	76	208	1066	1233	575	936
hCG ratio	0.41	0.31	2.20	0.60	1.14	0.56
hCG average (IU/L)	125	330	755	885	196	472
Progesterone 0 h (nmol/L)	4	8	68	36	29	27
Progesterone 48 h (nmol/L)	3	3	63	39	24	25
Progesterone average (nmol/L)	4	5	67	35	14	22
Endometrial thickness (mm)	8.3	5.9	13.5	5.7	10.3	8.2
Age (years)	31	9	29	10	33	6
Gestation (days)	44	15	32	8	41	11
Ordinal/binary variables	***N***	**%**	***N***	**%**	***N***	**%**
Vaginal bleedings						
No	109	23.7	251	76.1	23	34.8
Yes, without clots	248	53.9	76	23.0	38	57.6
Yes, with clots	103	22.4	3	0.9	5	7.6
Pain at examination	216	47.0	161	48.8	25	37.9
Previous EP	10	2.2	21	6.4	8	12.1
Previous miscarriage	2	0.4	7	2.1	0	0.0
Disrupted midline echo	95	20.7	15	4.5	11	16.7
Free fluid in pouch of Douglas	49	10.7	93	28.2	10	15.2

IQR: interquartile range.

..., k). The true class probability for class i, $P(y = i|\mathbf{x})$, is denoted as π_i, and $\boldsymbol{\pi} = [\pi_1 \ldots \pi_k]^T$.

3.1. Multi-class logistic regression

In MLR, the kth of k classes is used as the reference class. Then, logit equations are constructed for each class i ($i \neq k$) versus the reference class:

$$\begin{cases} \text{logit}(\pi_1) = \log\dfrac{\pi_1}{\pi_k} = \alpha_1 + \boldsymbol{\beta}_1^T\mathbf{x} = z_1 \\ \text{logit}(\pi_2) = \log\dfrac{\pi_2}{\pi_k} = \alpha_2 + \boldsymbol{\beta}_1^T\mathbf{x} = z_2 \\ \vdots \\ \text{logit}(\pi_{k-1}) = \log\dfrac{\pi_{k-1}}{\pi_k} = \alpha_{k-1} + \boldsymbol{\beta}_{k-1}^T\mathbf{x} = z_{k-1} \end{cases}, \tag{1}$$

where the model parameters α_i and $\boldsymbol{\beta}_i$ ($i = 1, \ldots, k-1$) represent the intercept term and the vector of length p consisting of each input's weight in the linear combination. Approaching MLR from a statistical point of view involves checking the assumption of linearity in the logit for each input and checking the necessity of interaction effects [10].

It is assumed that all inputs are linearly related to $\text{logit}(\pi_i)$, given all other inputs in the model. If the assumption is violated, a transformation of the input may help, or possibly the addition of a polynomial effect to the input set. The parameters of all logit equations are estimated simultaneously using maximum likelihood estimation based on a multinomial likelihood function. The true class probability π_i is equal to

$$\pi_i = \frac{\exp(z_i)}{\sum_{j=1}^{k}\exp(z_j)}, \tag{2}$$

where z_k equals 0.

3.2. Combination of binary logistic regression models using pairwise coupling

It is likely that variables will not be useful for distinguishing between each pair of outcomes. Therefore, we created separate binary logistic regression models (with separate input selection) to compare each pair of outcome classes, i.e. the 1-versus-1 approach. This approach allowed to use variables as inputs only in the models they were needed in. The binary models were constructed on the training set using only the cases from the two classes i and j involved ($i \neq j$), and were then applied to all test set cases. The $k(k-1)/2$ binary test set probabilities $\hat{\pi}_{ij}$ were estimates of the true probability of belonging to class i given class i or j (π_{ij}). Then, the $\hat{\pi}_{ij}$ had to be combined to obtain an estimate for $\boldsymbol{\pi}$. This was done using pairwise coupling [13]. In the first pairwise coupling method presented in Ref. [13], $\boldsymbol{\pi}$ is estimated by solving

$$\pi_i = \sum_{j,j\neq i}^{k}\hat{\pi}_{ij}\frac{\pi_i + \pi_j}{k-1},\ \forall i, \text{ where } \sum_{i=1}^{k}\pi_i = 1 \text{ and } \pi_i \geq 0 \tag{3}$$

for each case separately. Eq. (3) is solved using a linear system. This method is abbreviated as LR-PC.

3.3. Bayesian multi-layer perceptron

The fully connected feed-forward MLP is a standard neural network model [5]. We used the tan h transformation function in the hidden layer and the softmax transformation function in the output layer (which has three units) such that the output layer was an estimate of $\boldsymbol{\pi}$. Parameter optimization was done using the Broyden–Fletcher–Goldfarb–Shanno quasi–Newton algorithm [5]. The MLP was approached in a Bayesian sense using MacKay's evidence procedure [5,11]. In this procedure, the prior distribution for the model parameter vector $\mathbf{w}$, which includes all weights and bias terms of the MLP, is Gaussian with mean zero and variance $1/\alpha$. This prior distribution regards small weights more likely than large weights and thus regularizes the network. The hyperparameter α is the regularization parameter that controls the amount of regularization.

The evidence procedure looks for the most probable value for α rather than integrating over it. The most probable value for $\mathbf{w}$ ($\mathbf{w}_{MP}$) is obtained by maximizing the posterior, which, for multi-class classification, is equivalent to minimizing the entropy error function [5] augmented with a regularization term. The posterior distribution for $\mathbf{w}$ is approximated by a Gaussian around its mode $\mathbf{w}_{MP}$. Class probabilities are obtained by integrating the network output using $\mathbf{w}_{MP}$ over the posterior distribution to incorporate uncertainty with respect to the estimated model parameters. In our analyses, consistent priors were used [5]. This means that, rather than a single hyperparameter α, a separate hyperparameter is used for each layer's weights and biases.

Even though the use of the evidence procedure for Bayesian neural networks has been criticized [14,15], it works well for large data sets [14] and good results have been reported [16–18]. This method is abbreviated as BMLP.

3.4. Bayesian perceptron

We also fitted a Bayesian perceptron, which is similar to an MLP without hidden layer. This means that it resembles a Bayesian MLR model. However, no effort was put into checking assumptions or interaction terms. This method is abbreviated as BPER.

3.5. Bayesian least squares support vector machines

As a variant of standard support vector machines (SVMs) [19], LS-SVMs [6,7] require solving a linear system rather than a quadratic programming problem. SVMs are in nature binary classifiers. For (LS-)SVMs, the input space is mapped into a high dimensional feature space using mapping $\varphi: \mathbb{R}^p \to \mathbb{R}^q$. 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. Using y^* as the binary class indication (−1 versus +1), the classifier $y^*(\mathbf{x}) = \text{sign}[\mathbf{w}^T\varphi(\mathbf{x}) + b]$ is obtained by minimizing the cost function

$$\min_{w,b,e}\frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}\gamma\sum_{n=1}^{N}e_n^2 \quad \text{such that } y_n^*[\mathbf{w}^T\varphi(\mathbf{x}_n) + b] = 1 - e_n, \tag{4}$$

with y_n^* the binary class indicator of case n, $\mathbf{w}$ the weight vector of length q, b the bias term, e_n the error variable, and γ a hyperparameter to control the amount of regularization. After taking the Lagrangian of the cost function, the LS-SVM classifier can be reformulated as $y^*(\mathbf{x}) = \text{sign}\left[\sum_{n=1}^{N}\alpha_n y_n^* K(\mathbf{x}, \mathbf{x}_n) + b\right]$, with $\alpha_1, \ldots, \alpha_N$ 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 $K(\mathbf{x}, \mathbf{z}) = \varphi(\mathbf{x})^T\varphi(\mathbf{z})$. The choice of kernel affects how the linear separation in the feature space relates to the input space. We used the linear kernel $\mathbf{x}^T\mathbf{z}$ for constructing linear binary classifiers in the input space and the radial basis function (RBF) kernel $\exp(-\|\mathbf{x} - \mathbf{z}\|_2^2/\sigma^2)$ for nonlinear classifiers. The support values and the bias term are found by solving a linear Karush–Kuhn–Tucker system [6,7].

A drawback of standard LS-SVMs is that they do not have the sparseness property of SVMs. Because for LS-SVMs $\alpha_n = \gamma e_n$, easy cases can have negative support values. We imposed sparseness to the model by repeatedly pruning training data with negative support values [20]. Another problem is that binary LS-SVM classifiers do not give probabilistic output. By using a Bayesian framework based on the evidence procedure as explained in Refs. [20,21], binary class probabilities can be obtained based on the latent output of the LS-SVM, i.e. $\sum_{n=1}^{N}\alpha_n y_n^* K(\mathbf{x}, \mathbf{x}_n) + b$, and the prior class probabilities which are the training set proportions of cases in each class. Whilst other methods to obtain probabilistic output for SVMs or SVM-like formulations exist [22–24], we preferred to use a Bayesian

approach. This method worked well in a comparison study of methods for probabilistic multi-class classification using LS-SVMs [25].

We obtained binary class probabilities using the 1-versus-1 approach (cf. LR-PC). The binary probabilities based on Bayesian LS-SVMs were combined in two ways. The first approach is the pairwise coupling approach described earlier. The second approach used ensembles of nested dichotomies (END) [26]. In this approach, a three-class problem is presented as an ensemble of two nested binary classifiers. First, a classifier is constructed to compare the first two classes with the third class. Second, a classifier is constructed to compare the first with the second class. The multi-class probabilities for the two classes involved in the second step are obtained by multiplying the probability to belong to these classes according to the first classifier with the probabilities obtained in the second step. Three different trees can be constructed for a three-class problem, each of which is often not more important than any other a priori. Therefore, we averaged the results for each tree to obtain a final estimate of π. Both the pairwise coupling and END methods to combine binary classifiers performed well in the comparison study of probabilistic multi-class LS-SVM classifiers mentioned previously [25].

Thus, we constructed binary BLSSVM classifiers using either the linear or RBF kernel, and combined them using either pairwise coupling or END. This means that we implemented four BLSSVM methods, called BLSSVMlin-PC, BLSSVMlin-END, BLSSVMrbf-PC, and BLSSVMrbf-END.

3.6. Multi-class kernel logistic regression

KLR is a method that is closely related to SVMs [27], because essentially they only differ with respect to their loss function. As noted in Ref. [27], KLR has the advantages to directly yield probabilistic output and to be readily extended to MKLR. However, due to the change in loss function, no support vector will be zero and sparseness is lost. Recently, it has been shown that MKLR can be solved in an LS-SVM framework [12]. This method starts from a regularized version of MLR which is solved by optimizing the penalized negative log likelihood function using iteratively regularized re-weighted least squares. By mapping the input space into a high dimensional feature space using a positive definite kernel and applying a model with the structure of an LS-SVM in each iteration, a kernel version of MLR is obtained. See Appendix A for details. We used the RBF kernel for MKLR to automatically obtain a nonlinear separation in the input space.

3.7. Input selection

Input selection was performed separately for different algorithms. Remember that MLR and LR-PC involved manual checks, such that input selection varied between algorithms by definition. For the MLR approach and the three binary LR models for the LR-PC approach, several promising input sets were selected using three criteria. The first two criteria were the Akaike and Bayesian information criteria [28], both of which penalize the log likelihood of the data for the number of inputs used. Roughly, the Akaike criterion has a tendency to overfit by selecting too many inputs, whereas the Bayesian criterion has a tendency to select too few inputs. Thus, an input set with good values for both criteria was desired. The third, less important criterion is the likelihood ratio chi-squared test that can be performed for each input to investigate whether it is linked to the outcome given the other inputs in the model. Then, for each promising input set, the linearity in the logit assumption [10] was checked for all continuous/ordinal inputs. If necessary, inputs were transformed or polynomial terms were added. Next, we checked whether it was necessary to include interaction terms to the input sets. Finally, 20 fivefold cross-validation (5CV) runs were performed on the training set to select the best input set. We stratified for outcome, such that the five folds in each 5CV run had the same class distribution. Using the estimated probabilities for failing PUL, IUP, and EP, each run resulted in five validation set areas under the receiver operating characteristic (ROC) curve (AUC) per class. After 20 runs, this resulted in 100 validation set AUCs per class. These 100 AUCs per class were summarized by their median. The median AUCs were used to guide the selection of the final input set. Extra importance was given to EP because that is the most difficult class.

For the BMLP and BPER approaches, and for the binary Bayesian LS-SVMs, the automatic relevance determination (ARD) method [14,16,29] was used within the context of each model. This is an input selection method that is a ready extension of the Bayesian framework. For Bayesian MLPs, ARD means that a separate hyperparameter α is defined for each input's weights. When an input's α is estimated to be small, the variance of the prior distribution for that input is considered large such that large weights are allowed for that input. This indicates that the input may be important. In this way, all inputs can be ranked from most to least important.

We repeated the ARD analysis ten times with different initializations of the model parameters and computed the median rank of each input over the 10 ARD runs. The input with lowest median rank was dropped and 10 new ARD runs were done using the reduced input set. A final ranking of input importance was constructed based on the order in which inputs were dropped. The ARD analysis was done using a hidden layer with 10 hidden neurons. Then, using cross-validation (20 stratified 5CV runs), we simultaneously determined (a) how many of the most important inputs had to be used, and (b) the number of hidden neurons for the final BMLP model (i.e. the network architecture). For the BPER model, a similar strategy was adopted.

For binary BLSSVMs, ARD input selection was implemented by inserting a diagonal weight matrix $U = \text{diag}\,[u_1, \ldots u_p]$ in the kernel function [29]. We used the RBF kernel for ARD which becomes

$$K(x,z) = \exp\left[\frac{-(x-z)^T U (x-z)}{\sigma^2}\right] = \exp\left[-(x-z)^T \tilde{U} (x-z)\right], \quad (5)$$

where $\tilde{U} = U/\sigma^2$. The weights $\tilde{u}$ are optimized in the Bayesian model. The initial choice for $\tilde{U}$ is diag [1, ... 1]$/\sigma^2$, with σ the optimal RBF kernel parameter for the unweighted Bayesian LS-SVM (fitted on the training set). The input with the smallest weight was dropped and ARD was run again. This procedure resulted in a final ranking of the inputs based on when they were dropped. Twenty runs of stratified 5CV were used to determine how many of the most important units were to be chosen. This was done for all three 1-versus-1 models. The END approach for combining binary probabilities into multi-class probabilities also needed to train 1-versus-2 classifiers. The inputs for these classifiers were those selected for both 1-versus-1 classifiers involving the class that is compared to both other classes.

3.8. Model evaluation

Input selection and model training were done on the training set. The resulting classifiers were then applied to the test set, which was used solely for model testing. Model evaluation is a crucial element of classifier development and needs more scrutiny than it often gets [30–34]. The main performance evaluation measure was the AUC [35]. However, this is not straightforward in multi-class problems. Multi-class extensions of the AUC exist, yet we first looked at each class' separate AUC. Using the test set probabilities for each class, an ROC curve was constructed for each class versus the other classes. The accompanying AUC can be interpreted as the probability to correctly separate a (class i, not class i)-pair of cases. Let $n_1 = 1, \ldots, N_1$ represent the class i cases for which the probability of class i is written as p_{n1}, and $n_2 = 1, \ldots, N_2$ the cases from the other classes for which the probability of class i is written as p_{n2}. Then, the AUC is defined as

$$\text{AUC} = \frac{1}{N_1 N_2} \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} c(n_1, n_2),$$

$$\text{with} \quad c(n_1, n_2) = \begin{cases} 1 \text{ if } p_{n1} > p_{n2} \\ 0.5 \text{ if } p_{n1} = p_{n2} \\ 0 \text{ if } p_{n1} < p_{n2} \end{cases}. \quad (6)$$

Additionally, we used the volume under the ROC surface (VUS) [36] as a multi-class extension of the AUC. Similar to the AUC, the VUS can be interpreted as the probability to correctly identify three random members from each class.

We applied two other relevant performance evaluation measures. First, we computed the scored AUC (sAUC) as weighted version of the AUC [37]. Whereas the AUC uses the estimated probabilities only to rank the cases, the sAUC also incorporates the exact probabilities by adding a weight term to the AUC:

$$\text{sAUC} = \frac{1}{N_1 N_2} \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} c(n_1, n_2)\,(p_{n1} - p_{n2}),$$

$$\text{with} \quad c(n_1, n_2) = \begin{cases} 1 \text{ if } p_{n1} > p_{n2} \\ 0 \text{ otherwise} \end{cases}. \quad (7)$$

Second, we used the average entropy error, which summarizes how well the predicted probability of the correct class approaches unity by averaging each case's negative log of this probability [5]. Thus, whereas the AUC and VUS investigate the discriminatory potential of the classifiers, the sAUC and the entropy investigate the quality of the probabilities and/or the discriminatory potential.

Confidence intervals for performance measures were obtained using the bias-corrected bootstrap method based on 1000 bootstrap samples [38]. Applying different performance measures gives a good picture of how different classifiers compare to each other. It also allows the comparison of the conclusions drawn from the results of each performance measure.

Classifier comparison was done using a ranking technique [39] rather than multiple statistical tests to compare each pair of classifiers. The procedure was as follows. Classifiers were ranked with respect to a performance measure. To obtain information on the variability in the ranks of the classifiers, a bootstrap analysis was performed [39]. One thousand bootstrap samples were drawn from the test set. For each bootstrap sample, the classifiers were ranked with respect to the performance measure (ranks

were averaged in case of ties). This allowed the calculation, over the 1000 bootstraps, of the average rank (AR) as well as the proportion of times the classifier was ranked amongst the best κ ($Pm(\kappa)$) [39].

When evaluating classifiers, often threshold levels on the predicted probabilities are derived to be able to give black-and-white predictions. This allows the computation of the accuracy of a classifier and of sensitivity and specificity levels. However, accuracy estimates resulting from the application of the thresholds are not preferred as measures of performance [30,34]. Therefore, after the classifiers were compared, we derived thresholds for the best classifier only. The thresholds were derived on the training set in collaboration with the clinicians.

4. Results

4.1. Input selection

An overview of selected inputs per algorithm is presented in Table 3. The inputs selected for the Bayesian LS-SVM and MKLR approaches are the same since they are based on the same ARD analyses. Algorithms consisting of a combination of binary models (i.e. LR-PC and all BLSSVM algorithms) used different input sets for the different binary models. The various input selection analyses yielded similar results: hCG ratio, hCG average, progesterone average, vaginal bleedings, age, and history of EP were identified as important variables.

4.2. Comparison of classifiers on the test set

In this section, we refer to a classifier by using the abbreviation of the underlying algorithm (e.g. MLR). Since EP can have serious consequences but is difficult to predict, we focused on this class. Therefore, we ranked the classifiers for the following performance measures: AUC and sAUC for EP (AUC-EP and sAUC-EP), VUS, and entropy.

The AUC and sAUC indices for failing PUL and IUP were not used for ranking purposes, but are reported in Table 4. This table shows that all classifiers had excellent discriminative ability for predicting failing PULs and IUPs. For failing PULs, the AUC was at least 0.985. For IUPs, the AUC was at least 0.984. The test set ROC curves for failing PUL and IUP are shown in Fig. 1 (upper two plots). The different curves were not labeled since they were highly overlapping. For both classes, all ROC curves were nearly perfect.

Results for AUC-EP, sAUC-EP, VUS, and the entropy error are shown in Table 5 and Fig. 2. Table 5 shows the values of each performance measure for each classifier, the 95% confidence intervals around these values, and AR, $Pm(1)$, and $Pm(3)$. Fig. 2 presents the results of the 1000 bootstrap samples using box plots. Our first observation was that the linear classifiers (BPER, BLSSVMlin-PC, BLSSVMlin-END) performed poorest. Concerning the AUC-EP, LR-PC (AUC = 0.924, AR = 2.1, $Pm(1) = 0.38$, $Pm(3) = 0.89$) and MLR had the best ranking. For example, $Pm(1)$ suggested that there is an 80% chance that one of these two is the best of all classifiers. MKLR was ranked third, closely followed by BLSSVMrbf-END. The test set ROC curves for EP are shown in Fig. 1 (bottom two plots). When investigating discriminatory potential between all three classes simultaneously using the VUS, MKLR performed best (VUS = 0.895, AR = 2.4, $Pm(1) = 0.31$, $Pm(3) = 0.80$), followed by MLR and LR-PC.

LR-PC had the best weighted AUC performance for EP (sAUC-EP = 0.416, AR = 1.6, $Pm(1) = 0.57$, $Pm(3) = 0.99$), closely followed by BLSSVMrbf-END. MLR and BLSSVMrbf-PC also performed well,

Table 3 Overview of selected inputs per classifier.

Inputs classifier	hCG ratio	Log (hCG ratio)	Log (hCG avg)	Log (Prog avg)	Vag. bleed.	Age	Age squared	Prev. EP	Log(hCG ratio) × log(prog avg)
MLR		X	X	X	X	X	X		X
LR-PC		X[a,b,c]	X[a,b]	X[a,b,c]	X[c]	X[b]	X[b]	X[c]	X[b]
BMLP	X		X	X	X			X	
BPER	X		X	X	X				
BLSSVMlin-PC	X[a,b,c]		X[a,b]	X[a,b,c]	X[b,c]	X[a,c]		X[c]	
BLSSVMlin-END	X[a,b,c]		X[a,b]	X[a,b,c]	X[b,c]	X[a,c]		X[c]	
BLSSVMrbf-PC	X[a,b,c]		X[a,b]	X[a,b,c]	X[b,c]	X[a,c]		X[c]	
BLSSVMrbf-END	X[a,b,c]		X[a,b]	X[a,b,c]	X[b,c]	X[a,c]		X[c]	
MKLR	X		X	X	X	X		X	

[a] Used in the binary classifier contrasting failing PUL with IUP.
[b] Used in the binary classifier contrasting failing PUL with EP.
[c] Used in the binary classifier contrasting IUP with EP.

Table 4 Performance indices for the classes failing PUL and IUP based on the test set.

Classifier	Failing PUL	IUP
AUC (95% CI)		
MLR	.989 (.978–.995)	.989 (.979–.996)
LR-PC	.989 (.979–.996)	.988 (.977–.996)
BMLP	.988 (.978–.994)	.990 (.981–.996)
BPER	.987 (.976–.994)	.987 (.973–.995)
BLSSVMlin-PC	.985 (.971–.993)	.985 (.969–.995)
BLSSVMlin-END	.985 (.973–.993)	.984 (.967–.994)
BLSSVMrbf-PC	.989 (.981–.995)	.987 (.975–.995)
BLSSVMrbf-END	.990 (.982–.996)	.988 (.979–.996)
MKLR	.989 (.980–.995)	.988 (.979–.995)
sAUC (95% CI)		
MLR	.848 (.807–.885)	.834 (.790–.867)
LR-PC	.839 (.799–.874)	.829 (.794–.861)
BMLP	.584 (.549–.615)	.533 (.506–.559)
BPER	.778 (.738–.813)	.752 (.718–.782)
BLSSVMlin-PC	.839 (.793–.876)	.870 (.831–.904)
BLSSVMlin-END	.834 (.787–.870)	.851 (.812–.884)
BLSSVMrbf-PC	.844 (.804–.882)	.857 (.818–.894)
BLSSVMrbf-END	.853 (.811–.888)	.866 (.828–.901)
MKLR	.817 (.777–.853)	.813 (.775–.849)

CI: confidence interval.

whereas MKLR did not. For the entropy error, LR-PC was the single best classifier (average entropy = 0.228, AR = 1.8, $Pm(1) = 0.60$, $Pm(3) = 0.90$) with $Pm(1)$ and $Pm(3)$ clearly superior to any other classifier. LR-PC was followed by BLSSVMrbf-END and MKLR. Thus, LR-PC and, to a lesser extent, BLSSVMrbf-END produced classifiers with very good probabilities, as judged by sAUC-EP and entropy.

Overall, LR-PC was the best classifier, followed by MKLR, BLSSVMrbf-END, and MLR. BLSSVMrbf-PC and BMLP were not outstanding but performed better than the linear classifiers. To get an idea of the magnitude of the estimated probabilities for each class, we used LR-PC to construct box plots of the estimated probabilities in the test set (Fig. 3). For true failing PULs, the median probabilities were 0.999 for failing PUL, 0.0004 for IUP, and 0.0007 for EP. For true IUPs, the median probabilities were 0.013 for failing PUL, 0.951 for IUP, and 0.031 for EP. For true EPs, the median probabilities were 0.152 for failing PUL, 0.202 for IUP, and 0.541 for EP. Thus, the probabilities given for failing PULs or IUPs were extremely good. For EPs, the probabilities were less excellent but still good. One reason for this result was that EP was by far the least common PUL outcome. Another reason was that this class' profile was sandwiched between the profiles of both other classes, as can be seen in Table 2. Yet, notice that the estimated probability of EP was much larger for true EPs than for failing PULs or IUPs.

Next, thresholds were derived for the LR-PC classifier. Because failing PULs and IUPs were accurately predicted, it was interesting from a clinical viewpoint to find thresholds that identify patients from these two outcome classes with high confidence. For both classes, patients with a probability of 0.8 or more had a chance of more than 95% to belong to that class. The remaining patients were considered at risk for EP. When applying this system to the test set, we found that 45% of the women exceeded the threshold for failing PUL (155/348) and 38% for IUP (131/348). Ninety-eight percent of these predictions were correct (152/155 and 128/131, respectively). Of the remaining 62 women at risk for EP (18%), 19 had an EP (31%), 19 an IUP (31%), and 24 a failing PUL (39%). Thus, more than 80% of the patients were classified with high confidence as either failing PUL or IUP. Such a system allows the follow-up of PULs to be rationalized, by minimizing the number of scans and blood tests for low risk patients, whilst concentrating attention on those who are at risk for EP.

Since this work is an extension of previous work in which we developed MLR models for PULs [1–3], we applied the most recent of these classifiers, M5 [3], to the test set to measure its performance. M5 was compared to LR-PC by computing the performance difference between both classifiers, and constructing a 95% confidence interval around the difference using the bias-corrected bootstrap method based on 1000 bootstrap samples. The

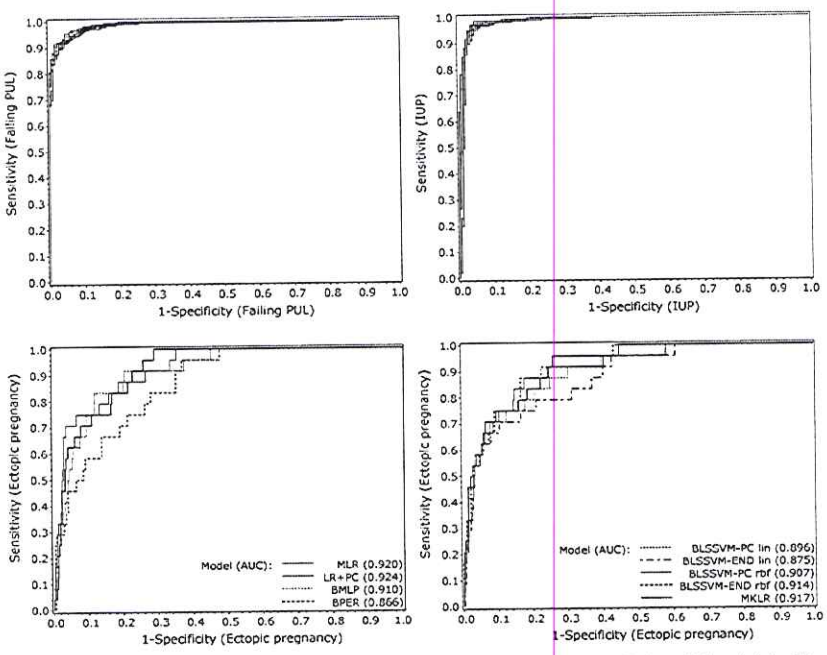


Figure 1 Test set ROC curves. Upper left: ROC curves for failing PUL; upper right: IUP; lower left and right: EP.

results are shown in Table 6. M5's AUCs on the test set were 0.984 for failing PUL, 0.975 for IUP, and 0.886 for EP. All performance measures were better for LR-PC than for M5. The 95% confidence intervals show that LR-PC's advantage over M5 was mainly apparent for the measures that incorporate the exact probabilities: the performance differences for sAUC-IUP, sAUC-EP, and entropy were statistically significant. Concerning standard measures for discriminative ability, only the performance difference for AUC-IUP was statistically significant. The performance difference for AUC-EP was large, but the relatively small number of EPs widened the confidence interval such that strong conclusions could not be drawn.

5. Discussion

In this study, we compared several probabilistic multi-class classifiers for PULs. We used nine algorithms based on LR (including checks of assumptions and checks for interaction terms), MLPs, LS-SVMs, and KLR.

The results can be summarized into six general findings. Firstly, LR-PC yielded the best classifier, followed by MLR, MKLR, and BLSSVMrbf-END. Secondly, MKLR performed very well, whilst nonlinearity was automatically incorporated and multi-class probabilistic outputs were directly given without Bayesian analysis or combination of binary results. MKLR performed as well as, or better than, the algorithms involving the combination of binary Bayesian LS-SVMs with RBF kernel (BLSSVMrbf-PC, BLSSVMrbf-END). As such, the extension of the LS-SVM framework to the area of probabilistic multi-class classification was fruitful. Thirdly, the performance of LR-PC suggests that combining binary probabilistic models with separate input selection may be an interesting alternative for a single multi-class model (cf. MLR). A drawback is that it becomes cumbersome for problems with many classes.

Fourthly, linear approaches (BPER, BLSSVMlin-PC, BLSSVMlin-END) performed poorer than nonlinear

Table 5 Performance indices and model ranking based on the test set.

Classifier	Value (95% CI)	AR Pm(1) Pm(3)
AUC-EP		
LR-PC	.924 (.872–.965)	2.1–0.38–0.89
MLR	.920 (.874–.954)	2.6–0.42–0.74
MKLR	.917 (.864–.962)	3.5–0.08–0.56
BLSSVMrbf-END	.914 (.852–.959)	3.7–0.04–0.36
BMLP	.910 (.853–.954)	4.5–0.04–0.23
BLSSVMrbf-PC	.907 (.835–.958)	4.9–0.05–0.20
BLSSVMlin-PC	.896 (.821–.954)	6.8–0.00–0.01
BLSSVMlin-END	.875 (.795–.945)	8.3–0.00–0.00
BPER	.866 (.799–.923)	8.6–0.00–0.00
VUS		
MKLR	.895 (.821–.942)	2.4–0.31–0.80
MLR	.892 (.832–.935)	3.2–0.32–0.64
LR-PC	.890 (.812–.948)	3.1–0.22–0.59
BMLP	.886 (.813–.932)	4.0–0.09–0.46
BLSSVMrbf-END	.884 (.802–.836)	4.1–0.03–0.31
BLSSVMrbf-PC	.873 (.781–.932)	5.4–0.03–0.16
BLSSVMlin-PC	.862 (.769–927)	6.9–0.00–0.03
BPER	.860 (.779–.914)	7.4–0.00–0.01
BLSSVMlin-END	.853 (.754–.922)	8.4–0.00–0.00
sAUC-EP		
LR-PC	.416 (.318–.509)	1.6–0.57–0.99
BLSSVMrbf-END	.402 (.287–.518)	2.0–0.32–0.97
MLR	.379 (.291–.481)	3.1–0.11–0.54
BLSSVMrbf-PC	.373 (.277–.473)	3.4–0.01–0.50
MKLR	.335 (.262–.421)	5.0–0.00–0.00
BLSSVMlin-PC	.301 (.219–.382)	5.9–0.00–0.00
BLSSVMlin-END	.227 (.160–.289)	7.4–0.00–0.00
BMLP	.221 (.183–.258)	7.6–0.00–0.00
BPER	.171 (.129–.214)	9.0–0.00–0.00
Average entropy		
LR-PC	.228 (.175–.293)	1.8–0.60–0.90
BLSSVMrbf-END	.242 (.176–.318)	3.1–0.19–0.61
MKLR	.245 (.195–.305)	4.0–0.05–0.42
BLSSVMrbf-PC	.249 (.183–.329)	4.7–0.03–0.29
MLR	.253 (.189–.338)	4.5–0.08–0.36
BLSSVMlin-PC	.253 (.180–.349)	4.6–0.05–0.30
BLSSVMlin-END	.261 (.193–.350)	5.8–0.01–0.13
BPER	.286 (.240–.344)	7.7–0.00–0.00
BMLP	.422 (.391–.460)	9.0–0.00–0.00

CI: confidence interval.

ones. It is useful to investigate whether a classification task can be successfully tackled using linear methods because this results in easier classifiers with higher interpretability. It should be noted that the distinction between linear and nonlinear models is artificial up to some point. For example, (a) using a transformation of a variable in a linear model renders the model nonlinear with respect to the original variable; (b) logistic regression belongs to the class of generalized linear models, even though the addition of interaction or polynomial terms introduces nonlinearity in the decision boundary; (c) LS-SVMs with RBF kernel are based on a linear separation in the feature space but not in the input space; and (d) combining linear classifiers using pairwise coupling does not result in an entirely linear multi-class classifier.

Fifthly, amongst the PUL population, failing PULs and IUPs can be predicted with very high level of confidence. However, even though the prediction of the potentially dangerous condition of EP is good, it remains the most difficult class. For true EPs, the median predicted probability of being an EP was 0.54 such that mildly elevated probabilities for EP

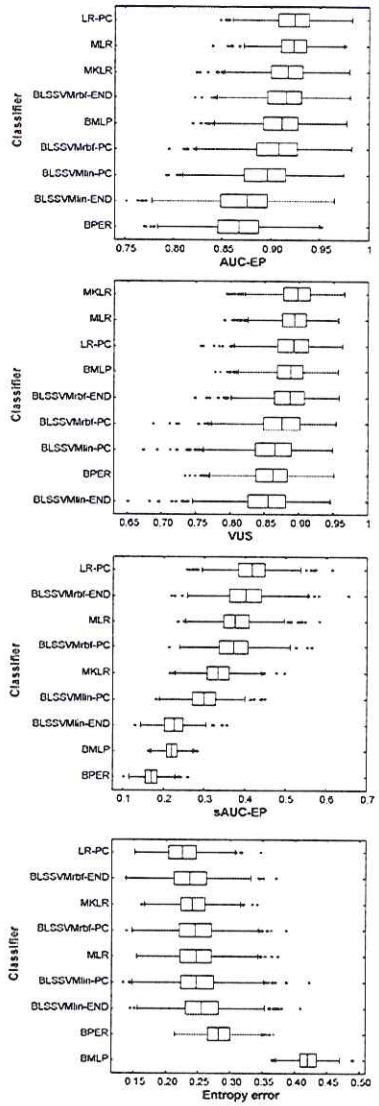


Figure 2 Box plots for each performance measure and classifier, based on 1000 bootstrap samples of the test set.

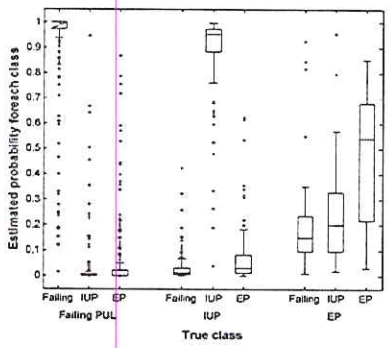


Figure 3 Box plots with LR-PC-based probabilities for each outcome (failing PUL, IUP, EP) for true failing PULs (left), intra-uterine pregnancies (middle), and ectopic pregnancies (right).

should raise concern. We should add here that the LR-PC classifier was also predictive for the at risk patients, i.e. the patients for which the probabilities of a failing PUL and of an IUP were both below 0.8. For example, the at risk patients in the test set that were EPs had a median probability of 0.55 for EP, whereas failing PULs and IUPs had a median probability for EP of 0.35 and 0.22, respectively.

Sixthly, the LR-PC classifier performed better than M5, the most recent existing classifier for PULs [3]. The advantage of LR-PC was mainly related to the performance measures that take the exact predicted probabilities into account: LR-PC produced better probabilities than M5. The discriminatory ability of LR-PC was also higher than that of M5, but the confidence intervals suggested that the difference is small. One problem is that the low number of EPs precluded a reliable investigation of the difference in AUC-EP between both classifiers.

We can divide the performance measures that were used for model evaluation and comparison into two categories: those that merely investigate discriminatory potential (AUC-EP, VUS) and those that investigate the quality of the predicted probabilities, possibly in combination with discriminatory potential (sAUC-EP, Entropy). We consider discriminatory potential to be most important, such that the first category of measures was given most attention. However, for classifiers with similar levels of discriminatory power, the second category of performance measures can be useful. In other analyses that are not discussed in this paper ([40], unpublished data), we observed that the correlation between measures from the first and second category is typically between 0.4 and 0.6, indicating that measures from the second category capture information that is not captured by measures from the first category.

Table 6 Comparison between M5 and LR-PC.

Performance measure	LR-PC	M5	Diff	95% CI[a]
AUC-failing PUL	.989	.984	.005	−.006; .020
AUC-IUP	.988	.975	.013	**.000; .030**
AUC-EP	.924	.886	.038	−.012; .116
VUS	.890	.859	.031	−.036; .103
sAUC-failing PUL	.839	.825	.014	−.018; .041
sAUC-IUP	.829	.773	.056	**.026; .088**
sAUC-EP	.416	.237	.179	**.089; .263**
Entropy	.228	.355	−.127	**−.226; −.042**

[a] Boldface confidence intervals do not contain zero.

Our study has some limitations. Firstly, even though the data set contained several hundreds of PULs, the number of EPs was small. This rendered reliable prediction of this class difficult. It is also known that ultrasound scanning ability can influence the class distribution in PULs. Good ultrasound operators may detect EPs on ultrasound more often, such that they will have fewer such pregnancies in their PUL population. As a result, it remains to be seen whether the classifiers have similar performance when they are applied to PULs collected at other clinical centres. Also, thresholds used to move from probabilities to black-and-white predictions may depend on the centre in which the data are collected. Finally, although we believe that the use of different input selection methods for different algorithms was a good strategy, it hampered an objective comparison between the algorithms that were used to construct the classifiers.

The analyses were performed in close collaboration with gynaecologists, since it was the aim to develop classifiers that can be implemented in clinical practice. The variables that were used in the models are reasonably objective: hCG and progesterone levels, age, level of vaginal bleeding, and history of EP. This suggests that these models should be fairly easy to implement as they do not require a high level of experience to utilize them, and it is unlikely that there will be significant differences in interpretation of the variables used. Even though the AUC for EPs is above 0.90 for LR-PC, EPs represent a small yet heterogeneous group of the PUL population. Therefore, we may find that larger studies improve on the prediction of EP. Meanwhile, research into new markers for EP is ongoing [41,42]. Furthermore, whilst achieving high detection rates for the three outcomes is important, these models can be seen as tools to improve clinical management: if we can predict the failing PULs and IUPs with high confidence, we can reduce the follow up required for these cases at low risk of ectopic pregnancy and more closely monitor the small group of 'at risk' pregnancies that are left. Using this approach we can rationalize management by reducing the number of repeat scans and blood tests in the low risk population, and thereby improve the patients' psychological situation and reduce the financial cost of PUL management [43]. LR-PC isolated more than 80% of patients that have either a failing PUL or an IUP with a very high level of confidence, leaving less than 20% at risk for ectopic pregnancy for which more careful follow is warranted.

At the moment, a multi-centre study examining PUL is recruiting patients (IPULA; International Pregnancy of Unknown Location Analysis) in order to evaluate PUL classifiers on data from a series of new centres, to investigate the centre dependency of our current models when tested on the PUL population and to develop new models. This is an important next step when pursuing the development of a widely applicable classifier that can be used in routine practice. Results from this large prospective study may also contribute to improving the detection of the EP group as a greater number of morphological and behavioural subgroups of ectopic pregnancy will be included in the database. Another very important study that should be carried out involves an interventional trial that assesses the clinical impact of a classifier in terms of the number of blood tests, scans, follow-up visits and financial cost, in the management of PUL. By doing this we hope to demonstrate that the implementation of our models into clinical practice leads to improved patient management and better outcomes [44].

Acknowledgements

We are grateful to the anonymous reviewers for their useful comments that have greatly improved

this paper. We also thank Diana Sima and Peter Karsmakers for constructive discussions and technical assistance. Ben Van Calster is a post-doctoral researcher funded by the Research Foundation–Flanders (FWO). Research supported by *Research Council KUL*: GOA-AMBioRICS, CoE EF/05/006 Optimization in Engineering (OPTEC), IDO 05/010 EEG-fMRI, IOF-KP06/11 FunCopt; *Flemish Government*, *FWO*: projects G.0407.02 (support vector machines), G.0341.07 (Data fusion); *Belgian Federal Science Policy Office*: IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007-2011); *European Union*: BIOPATTERN (FP6-2002-IST 508803), ETUMOUR (FP6-2002-LIFESCIHEALTH 503094), Healthagents (IST–2004–27214), FAST (FP6-MC-RTN-035801).

Appendix A. Multi-class kernel logistic regression

The LS-SVM based version of MKLR [12] starts from a regularized version of MLR which is solved by optimizing the penalized negative log likelihood function (PNLL) using iteratively regularized re-weighted least squares (IRRLS). The PNLL adds a penalty to the negative log likelihood function (NLL) for MLR in order to keep the parameters small:

$$PNLL = NLL + \frac{\gamma}{2}\sum_{l=1}^{k-1}(\beta_l^+)^T\beta_l^+,$$

where $\beta_l^+ = [\alpha_l\ \beta_l]$ contains the intercept, and α_l and β_l are as in Eq. (1). To account for the intercept terms, we define x_n^+ as x_n with a 1 pasted in front. Let $B = [\beta_1^T\ \ \beta_2^T\ \ \cdots\ \ \beta_{k-1}^T]^T \in \mathbb{R}^{(k-1)(p+1)}$. Parameter optimization is performed iteratively by updating the estimated B:

$$B^{(r)} = B^{(r-1)} + s^{(r)},$$

with $B^{(r)}$ the parameter estimates at iteration r, and the step $s^{(r)}$ equal to $-[H^{(r)}]^{-1}g^{(r)}$, where g is the gradient of the PNLL with respect to B, and H is the Hessian. This can be expressed as an IRRLS method.

To formulate this model in an LS-SVM framework, we apply the transformation $\varphi : \mathbb{R}^{p+1} \to \mathbb{R}^q$ to the input space to obtain the feature space with data $\varphi(x_n^+)$. To represent the feature space data, define $\Phi^T \in \mathbb{R}^{(k-1)q\times(k-1)N}$ as

$$\begin{bmatrix} \varphi(x_1^+) & 0 & 0 & \varphi(x_2^+) & 0 & 0 & \cdots & \varphi(x_N^+) & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 & & 0 & \ddots & 0 \\ 0 & 0 & \varphi(x_1^+) & 0 & 0 & \varphi(x_2^+) & & 0 & 0 & \varphi(x_N^+) \end{bmatrix}.$$

Define r_n as

$$[\mathcal{I}(y_n = 1)\cdots\mathcal{I}(y_n = k-1)]^T$$

R as $[r_1\ \ldots\ r_N]^T$, the vector $P^{(r)} \in \mathbb{R}^{(k-1)N}$ with estimated class probabilities as

$$[p_{1,1}^{(r)}\cdots p_{k-1,1}^{(r)}\cdots p_{1,N}^{(r)}\cdots p_{k-1,N}^{(r)}]^T,$$

And the nth block of the blockdiagonal weight matrix $W^{(r)} = \text{blockdiag}(W_1^{(r)},\ldots,W_N^{(r)})$ as

$$W_n^{(r)} = \begin{bmatrix} t_n^{1,1} & t_n^{1,2} & t_n^{1,k-1} \\ t_n^{2,1} & t_n^{2,2} & t_n^{2,k-1} \\ & \ddots & \\ t_n^{k-1,1} & t_n^{k-1,2} & t_n^{k-1,k-1} \end{bmatrix},$$

with

$$t_n^{i,j} = \begin{cases} p_{i,n}^{(r)}(1 - p_{i,n}^{(r)}), \text{ if } i = j \\ -p_{i,n}^{(r)}p_{j,n}^{(r)}, \text{ if } i \neq j \end{cases}$$

By using a positive definite kernel, one can apply IRRLS to the feature space such that one ends up with an iteratively re-weighted LS-SVM (irLS-SVM) algorithm used to obtain π_l as

$$\pi_l = \frac{\exp\left[(\beta_l^+)^T\varphi(x^+)\right]}{\sum_{j=1}^{k}\exp\left[(\beta_j^+)^T\varphi(x^+)\right]},$$

where $\beta_k^T\varphi(x^+) = 0$ since the kth class is the reference class. In irLS-SVM, one iteratively minimizes the cost function

$$\min_{s^{(r)},e^{(r)}} \frac{1}{2}(e^{(r)})^TW^{(r)}e^{(r)} + \frac{\gamma}{2}(s^{(r)} + B^{(r-1)})^T(s^{(r)} + B^{(r-1)}),$$

$$\text{such that } (W^{(r)})^{-1}(R - P^{(r)}) = \Phi s^{(r)} + e^{(r)}$$

where e contains the errors e_n. The problem can be rewritten in its dual form by taking the Lagrangian and defining the first order conditions for optimality. In the dual form, the MKLR model can be obtained by iteratively solving an LS-SVM problem using a linear system [12]. Probabilities for a new data vector x can now be obtained using $(\beta_l^+)^T\varphi(x^+) = \gamma^{-1}\sum_{n=1,n\in D_l}^{N}\alpha_{n,l}K(x_n^+,x^+)$, with D_l denoting the training data belonging to class i, $\alpha_{n,l}$ the support value for the prediction of the ith class for the nth training case and $K(x_n^+,x^+) = \varphi(x_n^+)^T\varphi(x^+)$ a positive definite kernel function to work in the feature space without explicitly constructing it. Thus, one now has $(k\text{-}1)N$ support values α instead of N in binary SVMs. We used the RBF kernel for the MKLR model.

References

[1] Condous G, Okaro E, Khalid A, Timmerman D, Lu C, Zhou Y, et al. The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location. Hum Reprod 2004;19:1900–10.

[2] Condous G, Van Calster B, Kirk E, Haider Z, Timmerman D, Van Huffel S, et al. Prediction of ectopic pregnancy in women with a pregnancy of unknown location. Ultrasound Obstet Gynecol 2007;29:680–7.

[3] Condous G, Van Calster B, Kirk E, Haider Z, Timmerman D, Van Huffel S, et al. Clinical information does not improve the performance of mathematical models in predicting the outcome of pregnancies of unknown location. Fertil Steril 2007;88:572–80.

[4] Hand DJ. Classifier technology and the illusion of progress. Stat Sci 2006;21:1–15.

[5] Bishop CM. Neural networks for pattern recognition. New York: Oxford University Press; 1995.

[6] Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9:293–300.

[7] Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least squares support vector machines. Singapore: World Scientific; 2002.

[8] Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. Biostatistics 2004;5:427–43.

[9] Provost F, Kohavi R. On applied research in machine learning. Mach Learn 1998;30:127–32.

[10] Hosmer DW, Lemeshow S. Applied logistic regression, 2nd edition, New York: Wiley; 2000.

[11] MacKay DJC. Probable networks and plausible predictions–a review of practical Bayesian methods for supervised neural networks. Network Comput Neural Syst 1995;6:469–505.

[12] Karsmakers P, Pelckmans K, Suykens JAK. Multi-class kernel logistic regression: a fixed-size implementation. In: SIJ, Sun R, Brown D, King I, Kasabov N, editors. Proceedings of the 20th International Joint Conference on Neural Networks. Los Alamitos: IEEE Press; 2007. p. 1756–61.

[13] Wu TF, Lin CJ, Weng RC. Probability estimates for multi-class classification by pairwise coupling. J Mach Learn Res 2004;5:975–1005.

[14] Neal RM. Bayesian learning for neural networks. New York: Springer; 1996.

[15] Vivarelli F, Williams CKI. Comparing Bayesian neural network algorithms for classifying segmented outdoor images. Neural Networks 2001;14:427–37.

[16] Nabney IT, Evans DJ, Brulé Y, Gordon C. Assessing the effectiveness of Bayesian feature selection. In: Husmeier D, Dybowski R, Roberts S, editors. Probabilistic modelling in medical informatics and bioinformatics. London: Springer; 2005. p. 371–90.

[17] Van Calster B, Timmerman D, Nabney IT, Valentin L, Testa AC, Van Holsbeke C, et al. Using Bayesian neural networks with ARD input selection to detect malignant ovarian masses prior to surgery. Neural Comput Appl 2008;17:489–500.

[18] Thodberg HH. A review of Bayesian neural networks with an application to near infrared spectroscopy. IEEE Trans Neural Networks 1996;7:56–72.

[19] Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.

[20] Lu C, Van Gestel T, Suykens JAK, Van Huffel S, Vergote I, Timmerman D. Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines. Artif Intell Med 2003;28:281–306.

[21] Van Gestel T, Suykens JAK, Lanckriet G, Lambrechts A, De Moor B, Vandewalle J. A Bayesian framework for least squares support vector machine classifiers. Neural Comput 2002;15:1115–48.

[22] Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola AJ, Bartlett PL, Scholkopf B, Schuurmans D, editors. Advances in large margin classifiers. Cambridge: MIT Press; 2000. p. 61–74.

[23] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Zaiane OR, Goebel R, Hand D, Keim D, Ng R, editors. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press; 2002. p. 694–9.

[24] Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 2001;1:211–44.

[25] Van Calster B, Luts J, Suykens JAK, Condous G, Bourne T, Timmerman D, Van S, Huffel. Comparing methods for multi-class probabilities in medical decision making using LS-SVMs and kernel logistic regression. In: Marques de Sa J, Alexandre LA, Duch W, Mandic D, editors. Artificial neural networks–ICANN 2007 lecture notes in computer science, vol. 4669. Berlin: Springer; 2007. p. 139–48.

[26] Frank E, Kramer S. Ensembles of nested dichotomies for multi-class problems. In: Greiner R, Schuurmans D, editors. Proceedings of the 21st International Conference on Machine Learning. New York: ACM Press; 2004. p. 305–12.

[27] Zhu J, Hastie T. Kernel logistic regression and the import vector machine. J Comput Graph Stat 2005;14:185–205.

[28] Burnham KP, Anderson DR. Model selection and inference: a practical information-theoretic approach. New York: Springer; 1998.

[29] van Gestel T, Suykens JAK, de Moor B, Vandewalle J. Automatic relevance determination for least squares support vector machines classifiers. In: Verleysen M, editor. Proceedings of the European Symposium on Artificial Neural Networks. Evere: d-side Publications; 2001. p. 13–8.

[30] Provost F, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. In: Shavlik JW, editor. Proceedings of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann; 1998 . p. 445–53.

[31] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73.

[32] Adams NM, Hand DJ. Improving the practice of classifier performance assessment. Neural Comput 2000;12:305–11.

[33] Hand DJ. Measuring diagnostic accuracy of statistical prediction rules. Stat Neerl 2001;55:3–16.

[34] Ling CX, Huang J, Zhang H. AUC: a statistically consistent and more discriminating measure than accuracy. In: Gottlob G, Walsh T, editors. Proceedings of the 18th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann; 2003. p. 519–26.

[35] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

[36] Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Med Decis Making 2000;20:323–31.

[37] Wu S, Flach Q, Ferri C. An improved model selection heuristic for AUC. In: Kok JN, Koronacki J, Lopez de Mantaras R, Matwin S, Mladenic D, Skowron A, editors. Machine learning–ECML 2007 lecture notes in computer science, vol. 4701. Berlin: Springer; 2007. p. 478–89.

[38] Chernick MR. Bootstrap methods: a practitioner's guide. New York: Wiley; 1999.

[39] Pepe MS, Longton G, Anderson GL, Schummer M. Selecting differentially expressed genes from microarray experiments. Biometrics 2003;59:133–42.

[40] Van Calster B, Van Belle V, Condous G, Bourne T, Timmerman D, Van Huffel S. Multi-class AUC metrics and weighted alternatives. In: Liu D, Kozma R, editors. Proceedings of the 21st International Joint Conference on Neural Networks. Los Alamitos: IEEE Press; 2008. p. 1391–7.
[41] Condous G, Kirk E, Syed A, Van Calster B, Van Huffel S, Timmerman D, et al. Do levels of serum cancer antigen 125 and creatine kinase predict the outcome in pregnancies of unknown location? Hum Reprod 2005;20:3348–54.
[42] Kirk E, Papageorghiou A, Van Calster B, Condous G, Cowans N, Van Huffel S, Timmerman D, Spencer K, Bourne T. The use of serum inhibin A and activin A levels in predicting the outcome of 'pregnancies of unknown location'. Hum Reprod; in press.
[43] Kirk E, Condous G, Van Calster B, Van Huffel S, Timmerman D, Bourne T. Rationalizing the follow-up of pregnancies of unknown location. Hum Reprod 2007;22:1744–50.
[44] Wyatt JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? BMJ 1995;311:1539–41.