

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

PREDICTIVE DIAGNOSTIC MODELS FOR GYNECOLOGIC APPLICATIONS WITH FOCUS ON MULTI-CLASS CLASSIFICATION

Promotoren:
Prof. dr. ir. S. Van Huffel
Prof. dr. D. Timmerman

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Ben VAN CALSTER

Maart 2008



KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

PREDICTIVE DIAGNOSTIC MODELS FOR GYNECOLOGIC APPLICATIONS WITH FOCUS ON MULTI-CLASS CLASSIFICATION

Jury:
Prof. dr. ir. J. Berlamont, voorzitter
Prof. dr. ir. S. Van Huffel, promotor
Prof. dr. D. Timmerman, promotor
Prof. dr. ir. J.A.K. Suykens
Prof. dr. A.-M. De Meyer
Prof. dr. I.T. Nabney (Aston University)
Prof. dr. T. Bourne (University of London)

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Ben VAN CALSTER

U.D.C. 681.3*I21 Maart 2008

© Katholieke Universiteit Leuven – Faculteit Ingenieurswetenschappen

Arenbergkasteel, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

D/2008/7515/29

ISBN 978–90–5682–920–9

Voorwoord

Toen Sabine me in 2004 een mailtje stuurde met de vraag of ik niet aan een doctoraat wilde beginnen bij haar, met onderzoek dat in het verlengde lag van de thesis die ik schreef voor de Master of Science in Statistics, had ik wel zin, maar de gedachte om als psycholoog aan de slag te gaan bij de Faculteit Ingenieurswetenschappen vond ik toch merkwaardig. Maar intussen, drie en een half jaar later, is gebleken dat ingenieurs en andere harde wetenschappers best compatibel kunnen zijn met psychologen. En dus ben ik blij dat ik dit doctoraat heb gemaakt. Zoals steeds kruisen heel wat mensen hierbij je pad, om je te helpen met je onderzoek of om je eraan te herinneren dat er ook andere dingen zijn.

Eerst en vooral ben ik mijn promotoren dank verschuldigd. Professor Sabine Van Huffel en professor Dirk Timmerman wil ik bedanken voor hun hulp, begeleiding, en de vrijheid die ze gaven. Professor Dirk Timmerman wil ik ook bedanken voor zijn bijdrage tot de zeer aangename en vruchtbare samenwerking tussen statistici en gynaecologen. Van beiden voelde ik dat ze vertrouwen hadden in mij, en mede door hen is dit doctoraat geworden tot wat het is. Verder wil ik graag professoren Johan Suykens en Anne-Marie De Meyer bedanken voor hun inbreng als leden van mijn begeleidingscommissie, professoren Ian Nabney en Tom Bourne voor hun bereidheid te zetelen in mijn examencommissie, en professor Jean Berlamont voor zijn bereidheid mijn examencommissie voor te zitten.

In de loop van mijn doctoraatsonderzoek heb ik met vele mensen mogen samenwerken, dankzij hen heb ik veel geleerd en kunnen werken op vele interessante onderwerpen: de mensen van de departementen Oncologie en Vrouw en Kind van de U.Z. Leuven (prof. Dirk Timmerman, prof. Patrick Neven, prof. Frederic Amant, prof. Hans Wildiers, prof. Ignace Vergote, Caroline Van Holsbeke, Ekatarini Domali, Lani Morales, Wouter Hendrickx), de gynaecologen van St. George's Hospital in Londen (prof. Tom Bourne, prof. George Condous, Emma Kirk, Cecilia Bottomley, Asma Khalid, Zara Haider, Linda Tan), de mensen van IOTA (prof. Dirk Timmerman, prof. Lil Valentin, prof. Tom Bourne, prof. Antonia Testa), de collega's van SISTA (prof. Johan Suykens, Kristiaan Pelckmans, Peter Karsmakers, Fabian Ojeda, allen bedankt voor de hulp in de wondere wereld van de least squares support vector machines, Olivier Gevaert, Anneleen Daemen, Frank De Smet), de partners binnen Biopattern (prof. Ian Nabney, prof. Michalis Zervakis, prof. Paulo Lisboa, Ioannis Dimou, Hane Aung), prof. Bea Van den Bergh, prof. Tim Smits,

prof. Lonneke van de Poll, Jo Røislien, Gerard Moulin-Romsée, en uiteraard de administratieve en financiële medewerkers van SCD (Ida Tassens, Ilse Pardon, Lut Vander Bracht, Péla Noé, Bart Motmans).

Met de mensen van onze onderzoeksgroep BioMed heb ik graag samengewerkt, de bureau gedeeld en over andere dingen dan het werk gebabbeld: Chuan, Lieveke, Wim, Anneleen, Jean-Michel, Pieter, Mieke, Ivan, Andy, Bart, Arjan, Teresa, Diana, Maarten, Vanya, Jan, Jean-Baptiste, Mariya, Anca, Katrien, Wouter, Steven, Joachim, Maria, Dominique, Bogdan en Gokcen.

Zoals het hoort, heb ik de laatste jaren geprobeerd om veel te bewegen: samen met gelijkgestemden (Tim, Sven, Steffen, Filip, Jeroen, Raf, Niels en mijn vader) ging ik al eens zwemmen, fietsen of lopen. Af en toe het verstand op nul zetten doet deugd.

Dan zijn er natuurlijk mijn ouders, die al heel mijn leven hun best doen om hun zoon te helpen en kansen te geven. Hopelijk hebben zij ook het gevoel dat hun inspanningen vruchten hebben afgeworpen.

Ik wil ook graag mijn vrienden bedanken. Ik heb veel aan jullie en ben blij dat ik jullie ken! Graag verontschuldig ik me dan ook voor die keren dat jullie meer vriendschap gaven dan terugkregen...

En om in goed gezelschap te eindigen, wil ik graag mijn twee vrouwen bedanken voor alles wat ze voor mij doen en betekenen. Ilse, ik zie je graag, bedankt voor alles, en samen blijven we ervoor gaan! Marentje, jij bent een ongelooflijke schat! Ik kan me geen leven meer indenken zonder jou. Papa is helemaal weg van zijn kleine parel en de knuffels die ze geeft!

For data and funding support, I am indebted to the International Ovarian Tumor Analysis group, St. George's Hospital Medical School, the Multidisciplinary Breast Centre from the University Hospitals Leuven, Prof. Bea Van den Bergh, and the EU projects BIOPATTERN (FP6-2002-IST 508803) and eTUMOUR (FP6-2002-LIFESCIHEALTH 503094).

Abstract

Building upon the ideas of evidence-based medicine, clinical decision support systems can be very helpful tools in clinical practice. A vast range of tools is available for building such systems, partly fed by the growing field of computational intelligence. Decision support systems can assist (but not replace) medical personnel when facing important patient-related decisions. However, it turns out that it is not trivial to implement qualitative decision support systems into everyday clinical practice. These systems need to emerge from intense collaboration between their developers and end-users such that the systems fill up clinical needs. Also, the systems need to be rigorously evaluated to show their functionality and to convince clinicians to implement them.

In this thesis, mathematical models were developed for the diagnosis of ovarian tumors and pregnancies of unknown location. In both situations, an accurate diagnosis is necessary such that optimal treatment decisions can be made. The focus was on probabilistic models since uncertainty information is vital in medical decision making. Both conditions represent multi-class classification problems, which are less straightforward than binary problems. The models are based on logistic regression, Bayesian least squares support vector machines, Bayesian multi-layer perceptrons, and kernel logistic regression.

The models were built in collaboration with gynecologists, and resulted in accurate predictions. The ovarian tumor models were based on multi-center data and have successfully passed prospective internal and prospective external evaluation. The models for pregnancy of unknown location were based on single center data and have passed a first internal evaluation. However, a large multi-center study is ongoing, aiming for a thorough validation and for the development of new models for pregnancies of unknown location.

Korte inhoud

Het *evidence-based medicine* principe indachtig, is het duidelijk dat systemen ter ondersteuning van klinische beslissingen zeer nuttige hulpmiddelen kunnen zijn in de klinische praktijk. Er zijn ontelbare toepassingen voor dergelijke systemen, deels door de vooruitgang in het domein van de computationele intelligentie. Deze systemen helpen clinici bij het nemen van belangrijke beslissingen, maar kunnen hen uiteraard niet vervangen. Het blijkt echter niet eenvoudig om goede systemen te implementeren in de klinische praktijk. Het is belangrijk dat dergelijke systemen gestoeld zijn op een intense samenwerking tussen de systeemontwikkelaars en de uiteindelijke gebruikers, zodat ze beantwoorden aan de reëele noden van deze gebruikers. Ook is het essentieel dat de systemen uitgebreid getest worden om een goed beeld te krijgen van hun performantie.

In deze thesis werden wiskundige modellen ontwikkeld voor de diagnose van ovariumtumoren en zwangerschappen van onbekende lokatie. Een snelle en accurate diagnose is voor beide condities nodig om een zo goed mogelijke beslissing inzake de behandeling te kunnen nemen. Er werd gewerkt met probabilistische modellen aangezien de mate van onzekerheid met betrekking tot een bepaalde diagnose een invloed heeft op de beslissingen van de behandelende arts. Het gaat in beide situaties over meerklasse classificatie, wat minder eenvoudig is dan binaire classificatie. De modellen zijn gebaseerd op logistieke regressie, Bayesiaanse kleinste kwadraten support vector machines, Bayesiaanse meerlaagse perceptrons, en kernel logistieke regressie.

De modellen werden ontwikkeld door nauw samen te werken met gynaecologen, en resulteerden in goede en betrouwbare voorspellingen. De modellen voor ovariumtumoren zijn gebaseerd op multicentrische data en hebben met succes prospectieve interne en prospectieve externe validaties doorstaan. De modellen voor zwangerschappen van onbekende lokatie zijn gebaseerd op data van één centrum en hebben een eerste interne evaluatie doorstaan. Op dit moment loopt er een multicentrische studie met de bedoeling de modellen voor zwangerschappen van onbekende lokatie uitvoerig te testen en om nieuwe modellen op te stellen.

Nederlandse samenvatting

Predictieve diagnostische modellen voor gynaecologische toepassingen met nadruk op meerklasse classificatie

Inleiding en algemene doelstellingen

Door opleiding en ervaring zijn artsen vaak in staat om adequate beslissingen te nemen aangaande de diagnose en behandeling van hun patiënten. Desondanks kunnen systemen ter ondersteuning van klinische beslissingen (*clinical decision support systems* of CDS systems) nuttig zijn. Ze kunnen bijvoorbeeld de relaties tussen patiëntkarakteristieken en differentiaaldiagnoses of optimale behandelingen expliciteren. CDS systemen kunnen bij uitstek nuttig zijn bij complexe problemen of voor artsen met weinig ervaring. Het doel van CDS systemen is het assisteren van medisch personeel bij het maken van beslissingen door patiëntkarakteristieken te linken aan, algemeen gezegd, een kennissysteem zodat patiënt-specifieke adviezen of beslissingen gegeven kunnen worden [88, 155, 197, 474, 475].

CDS systemen kunnen bijvoorbeeld gebruikt worden om afwijkingen vast te stellen in signalen zoals ECG metingen, voor diagnose en prognose, om behandelingen te plannen en op te volgen, om medicatie voor te schrijven, of om relevante informatie snel en automatisch te zoeken in de bestaande literatuur [473, 197, 88]. Het is belangrijk te beseffen dat dergelijke systemen louter dienen om medisch personeel te helpen, niet om ze te vervangen. Het is duidelijk dat dergelijke systemen beïnvloed worden door evoluties in elektronica (computers, laptops), telecommunicatie (lokale netwerken, internet), computer- en ingenieurswetenschappen, en statistiek. Dit ziet er veelbelovend uit, maar in de praktijk blijken CDS systemen niet zo snel ingang te vinden als men zou verwachten of wensen [87, 88, 155, 263, 474]. Belangrijke redenen hiervoor zijn dat het ontwikkelen van dergelijke systemen niet altijd afgesteld is op de klinische realiteit [43, 87, 88, 155], en dat een uitgebreide evaluatie ervan niet vaak wordt uitgevoerd [263, 474]. Aangaande het eerste punt kan vermeld worden dat systemen klinische noden moeten opvullen, dat ze snel en gebruiksvriendelijk moeten zijn, en dat ze zo transparant mogelijk moeten zijn [43, 88, 155]. Ze moeten ook mooi passen in het dagelijkse werkpatroon van de artsen.

Het belangrijkste doel van deze thesis is het bekomen van kennissystemen die kunnen leiden tot goede CDS systemen met de uiteindelijke bedoeling ze effectief te implementeren in de klinische praktijk. Het gaat meer specifiek om het ontwikkelen van predictieve diagnostische modellen voor ovariumtumoren en zwangerschappen van onbekende lokatie. Bijkomend werden ook analyses uitgevoerd om kennis te vergaren (*knowledge discovery*) ter beschrijving van bepaalde problemen en situaties. Meer bepaald werd onderzoek uitgevoerd naar borstkanker en het effect van prenatale angst van de moeder op depressieve symptomatologie bij adolescenten. Deze analyses worden slechts summier beschreven in de tekst.

Data

De data over ovariumtumoren zijn afkomstig van de *International Ovarian Tumor Analysis* (IOTA) groep. De dataset uit de eerste fase van IOTA bevat 1066 patiënten die gerecruteerd werden in negen centra uit vijf Europese landen (Italië, Frankrijk, België, Zweden en het Verenigd Koninkrijk). Deze dataset is opgesplitst in een groep van 754 patiënten die gebruikt werd om modellen te ontwikkelen, en een groep van 312 patiënten die gebruikt werd om de performantie van de modellen te evalueren. Tijdens het ontwikkelen van de modellen liep fase Ib, wat resulteerde in een prospectieve dataset met data van 507 patiënten afkomstig uit drie centra van IOTA fase I. Deze data werden gebruikt voor een prospectieve interne evaluatie van de modellen. Deze fase werd gevolgd door fase II, waarvan de dataverzameling op dit moment afgesloten is. Het checken en corrigeren van de data loopt op dit moment op zijn einde. De finale dataset zal informatie bevatten over bijna 2000 patiënten afkomstig uit 20 centra uit Italië, België, Zweden, Tsjechië, Polen, het Verenigd Koninkrijk, Canada en China. Aan deze fase werken niet alleen centra mee die ook deel uitmaakten van fase I, maar ook nieuwe centra. Daarom kan op deze dataset een prospectieve interne en een prospectieve externe evaluatie gebeuren. De data over zwangerschappen van onbekende lokatie werden verzameld in het Londense St. George's Hospital (Verenigd Koninkrijk). De dataset bevat informatie over 1003 patiënten met een zwangerschap van onbekende lokatie.

Specifieke doelstellingen

Met betrekking tot de diagnose van ovariumtumoren waren de specifieke doelstellingen het ontwikkelen van modellen die pre-operatief voorspellen of de tumor goedaardig of kwaadaardig is (binaire classificatie), het ontwikkelen van modellen die pre-operatief voorspellen of de tumor goedaardig, primair invasief, borderline kwaadaardig, of metastatisch invasief is (meerklasse classificatie), het prospectief intern en prospectief extern testen van de performantie van deze predictieve modellen, en het onderzoeken van het belang van het eiwit CA-125 in de voorspelling. Deze analyses maken deel uit van een groter geheel, gecoördineerd door de IOTA groep: andere onderzoekers hadden reeds binaire predictieve modellen gemaakt

[266, 20], en het onderzoek gaat verder dan het werk in deze thesis, onder andere met het verder testen van de modellen en het opstellen van nieuwe modellen gebaseerd op andere data.

De specifieke doelen aangaande het onderzoek naar zwangerschappen van onbekende lokatie waren het ontwikkelen van modellen die, aan de hand van gegevens van twee consultaties (dag 0 en dag 2), voorspellen of het gaat om een falende zwangerschap, een normale intra-uteriene zwangerschap of een ectopische zwangerschap (meerklasse classificatie), het testen van de modellen op een apart deel van de dataset, en het onderzoeken welke van de beschikbare metingen vermoedelijk voldoende zijn om een goede voorspelling te maken. Over dit onderwerp is, naar analogie met de IOTA groep, recent ook een internationale samenwerking opgestart (IPULA of *International Pregnancy of Unknown Location Analysis*). Deze studiegroep is op dit moment begonnen met het verzamelen van een grote multicentrische dataset die gebruikt zal worden om de modellen die in het kader van deze thesis ontwikkeld werden te testen en om nieuwe modellen te maken.

Enkele aspecten van meerklasse classificatie

Eerst werden echter enkele aspecten van meerklasse classificatie van naderbij bekeken. In de thesis wordt enkel gewerkt met probabilistische modellen die voor elke klasse een kansschatting geven. Dit is nuttig voor medische applicaties, omdat onzekerheid over een diagnose uiteraard een invloed heeft op de beslissing inzake de behandelingsstrategie. (Dit impliceert uiteraard dat de kansschattingen van de modellen accuraat zijn.) In het domein van de '*machine learning*' worden nog steeds nieuwe methoden voorgesteld om tot meerklasse kansschattingen te komen, aangezien dit een niet-triviale uitbreiding is van binaire classificatie. Er zijn bijvoorbeeld methoden waarbij de output van binaire probabilistische modellen wordt gecombineerd om tot meerklasse kansschattingen te komen (zie bijvoorbeeld [11]). Dergelijke methoden veronderstellen dat het meerklasse probleem eerst wordt opgesplitst in een reeks binaire problemen, bijvoorbeeld het ontwikkelen van modellen die elke klasse vergelijken met elke andere klasse (*1-vs-1*) of die elke klasse vergelijken met alle andere klassen (*1-vs-All*). Andere methoden combineren binaire niet-probabilistische modellen, of produceren tegelijkertijd kansschattingen voor alle klassen die sommeren tot 1 (*all-at-once*). Binaire niet-probabilistische modellen zoals kleinste kwadraten support vector machines (LS-SVMs, [373, 371]) kunnen binaire kansschattingen opleveren door ze bijvoorbeeld in een Bayesiaans kader te plaatsen [441] of door de output te transformeren [324, 477].

Gebruikmakend van de data waarop de predictieve modellen voor ovariumtumoren en zwangerschappen van onbekende lokatie werden ontwikkeld, werden een aantal LS-SVM gebaseerde methoden voor probabilistische meerklasse classificatie vergeleken [420]. De beste methode om tot binaire kansschattingen te komen bleek het gebruik van een Bayesiaans kader te zijn. Deze binaire kansschattingen werden

het beste gecombineerd aan de hand van de paarsgewijze koppelingsmethoden uit [472], waarbij 1-vs-1 kansschattingen gekoppeld worden, of aan de hand van 'combinaties van geneste dichotomieën' (*ensembles of nested dichotomies* of END). Bij END [157] wordt het meerklasse probleem opgesplitst in een boom van geneste dichotomieën. Voor alle dichotomieën uit de boom worden binaire probabilistische modellen ontwikkeld waarvan de kansschattingen gecombineerd worden door ze te vermenigvuldigen. Bijvoorbeeld, als er vier klassen zijn (a,b,c,d), kan men dit probleem eerst opsplitsen in de dichotomie a versus (b,c,d). De groep (b,c,d) kan vervolgens opgesplitst worden in de dichotomie c versus (b,d), en de groep (b,d) vervolgens in b versus d. De meerklasse kansschatting voor klasse d wordt dan bekomen door de kansen van de drie geneste dichotomieën te vermenigvuldigen. Men kan uiteraard verschillende boomstructuren opstellen, die a priori evenwaardig lijken. END werkte goed indien alle mogelijke boomstructuren toegepast werden, en vervolgens per klasse het gemiddelde van de bekomen kansschattingen berekend werd. Het gaat er vooral om dat je de resultaten voor meerdere boomstructuren middelt en niet zozeer dat je exhaustief alle boomstructuren toepast aangezien het aantal mogelijke structuren exponentieel toeneemt met het aantal klassen. Naast deze combinaties van binaire kansschattingen bleek LS-SVM gebaseerde meerklasse kernel logistieke regressie (MKLR) [220], een all-at-once techniek, bijzonder goede resultaten op te leveren. Het voordeel van MKLR ten opzichte van de END methode is dat ze veel minder complex is.

Een tweede aspect van meerklasse classificatie gaat over maten om de discriminatieve performantie te evalueren. Hierbij gaat men na hoe goed de modellen in staat zijn om de verschillende klassen te voorspellen, waarbij dan de kansschattingen voor een bepaalde klasse hoger zijn voor gevallen (patiënten) die tot die klasse behoren dan voor andere gevallen. Een veelgebruikte maat om discriminatieperformantie te evalueren voor binaire classificatiemodellen is de oppervlakte onder de ROC curve (*area under the receiver operating characteristic curve* of AUC). Kortweg gezegd kan men de AUC interpreteren als de kans dat het model twee toevallig gekozen gevallen (een van elke klasse) correct onderscheidt aan de hand van de gegeven kansschattingen. Een perfect model heeft een AUC van 1, een waardeloos model heeft een AUC van 0.5. Twee interessante uitbreidingen van de AUC handelen over het veralgemenen van de AUC naar meerklasse problemen [290, 181], en over het betrekken van de kansschattingen zelf in de AUC maat (gewogen AUC maten) [154, 470]. In de traditionele AUC worden de kansschattingen (of meer algemeen de numerische output van een model) enkel gebruikt om gevallen te ordenen. Maar, als de kansschatting voor klasse a voor een model groter is voor een geval van klasse a dan voor een geval van de andere klasse, is het ook interessant om te bekijken hoeveel groter de kansschatting is. Dergelijke gewogen AUC maten zijn interessant om te gebruiken bij modelselectie en bij modelevaluatie. Modellen met een grotere gewogen AUC maat zouden robuustere modellen kunnen zijn en kunnen gezien worden als modellen met betere kansschattingen. Door beide uitbreidingen te combineren komt met tot gewogen meerklasse AUC maten die kunnen gebruikt worden voor

meerklasse problemen. In deze thesis stellen we enkele maten voor en passen we ze toe op de evaluatie van modellen voor zwangerschappen van onbekende lokatie [414]. De meeste voorgestelde maten waren heel sterk gecorreleerd, zodat ze hetzelfde lijken te meten. Deze maten werden vervolgens toegepast op alle meerklasse modellen die in de thesis beschreven worden. Onze mening is dat dergelijke gewogen maten heel nuttig zijn, maar dat ze, wat modelevaluatie betreft althans, ondergeschikt zijn aan ongewogen AUC maten. Indien ongewogen maten geen onderscheid maken tussen verschillende modellen, kunnen gewogen maten mogelijks wel een voorkeur voor een bepaald model suggereren. Indien ongewogen AUC maten wel duidelijk een model verkiezen boven de andere modellen, lijken gewogen maten ons minder relevant.

Binaire en meerklasse modellen voor ovariumtumoren

Binaire predictieve modellen voor ovariumtumoren met behulp van de IOTA fase I dataset werden reeds in 2005 ontwikkeld: twee logistieke regressiemodellen (LR1 en LR2) [20, 394], drie LS-SVMs en drie relevance vector machines (RVMs) [266, 426]. In deze thesis hebben we het arsenaal aan modellen uitgebreid met twee Bayesiaanse meerlaagse perceptrons en een Bayesiaans perceptron [428, 427], en met twee logistieke regressiemodellen: een LASSO (*least absolute shrinkage en selection operator*) model en een gewoon model met enkel objectief te meten inputs. De variabelen die in minstens een van de binaire modellen gebruikt worden, zijn leeftijd, voorgeschiedenis van ovariumkanker, huidig gebruik van hormoonbehandeling, pijn tijdens het echografisch onderzoek, de aanwezigheid van ascites, maximale diameter van de eierstok, de maximale diameter van de tumor, de maximale diameter van de grootste soliede component van de tumor, het aantal papillaire structuren, aanwezigheid van bloedcirculatie in de papillaire structuren, onregelmatige binnenwand van de cyste, aanwezigheid van akoestische schaduwen op het echografisch beeld, een volledig soliede tumor, een uniloculaire tumor, een multiloculair-soliede tumor, de Doppler kleurenscore, en het vermoeden dat de tumor van ovariale oorsprong is. Alle modellen hadden goede performantie op de test set, met AUC waarden tussen 0.92 en 0.95. Er was geen model dat duidelijk beter was dan de rest op basis van de AUC en gewogen AUC resultaten. De lineaire en niet-lineaire modellen presteerden even goed. Aangezien de drie Bayesiaanse LS-SVM modellen de beste gewogen AUC waarden hadden, zou de LS-SVM met lineaire kernel gekozen kunnen worden. Echter, de inputs die door dit model gebruikt werden zijn niet optimaal. Bijvoorbeeld, dit model gebruikt het vermoeden dat de tumor van ovariale oorsprong is, een variabele die erg subjectief is en die de artsen liever niet als input in het model willen. Ook het gebruik van de maximale diameter van de eierstok was vreemd voor sommige clinici. Deze variabele is heel sterk gerelateerd aan de maximale diameter van de tumor. Andere interessant modellen zijn het logistieke regressiemodel met 12 inputs (LR1) [394] of het Bayesiaanse perceptron. Een andere interessante observatie is dat de modellen niet in staat waren om beter te presteren dan het subjectieve oordeel van de arts. Hierbij moet opgemerkt worden dat de artsen in fase I van IOTA ervaren zijn, zodat de modellen wel beter kunnen

scoren dan minder ervaren artsen (zie bijvoorbeeld [391]).

Zes meerklasse modellen voor ovariumtumoren werden ontwikkeld: een meerklasse logistiek regressiemodel (MLR), twee modellen die bestaan uit een combinatie van 1-vs-1 logistieke regressiemodellen met behulp van paarsgewijze koppeling (LR-PC en LR-PC2), een combinatie van 1-vs-1 Bayesiaanse LS-SVMs (LSSVM-PC), een combinatie van 1-vs-1 kernel logistieke regressiemodellen (KLR-PC), en een meerklasse kernel logistiek regressiemodel (MKLR). Het combineren van binaire modellen terwijl een meerklasse model voorhanden is, kan zijn nut hebben doordat men inputselectie kan doen voor elk binair model apart. Op die manier kan men ervoor zorgen dat inputs enkel gebruikt worden waar ze echt nodig zijn. Voor de laatste drie modellen werden de inputs geselecteerd met behulp van twee methoden: ARD (*automatic relevance determination*) [442] en een methode voor LS-SVMs met lineaire kernel aan de hand van rang-een updates (R1U methode) [310]. De ARD methode werd toegepast op LS-SVMs met zowel lineaire als radiale basisfunctie (RBF) kernels. Op basis van cross-validatie analyses op de training set bleek dat de R1U methode duidelijk betere resultaten opleverde, zodat deze methode gebruikt werd om de finale input sets te bekomen voor elk 1-vs-1 model. Voor MKLR is slechts één input set nodig, die bestond dan uit alle inputs die in minstens één van de 1-vs-1 modellen gebruikt werden op basis van R1U. De inputs voor MLR en LR-PC werden geselecteerd op basis van technieken binnen het kader van logistieke regressieanalyse, met de nodige aandacht voor eventuele transformaties van variabelen, en voor het eventuele belang van kwadratische effecten en interactie-effecten. Voor LR-PC2 werden de inputs gebruikte die geselecteerd werden met behulp van R1U.

De variabelen die voor minstens een van de meerklasse modellen gebruikt werden, zijn leeftijd, voorgeschiedenis van ovariumkanker, de aanwezigheid van tumors op beide eierstokken, de aanwezigheid van ascites, de maximale diameter van de tumor, de maximale diameter van de grootste soliede component van de tumor, het aantal papillaire structuren, aanwezigheid van bloedcirculatie in de papillaire structuren, onregelmatige binnenwand van de cyste, aanwezigheid van akoestische schaduwen op het echografisch beeld, een volledig soliede tumor, en een uniloculaire tumor. Merk op dat de meerklasse modellen een aantal eerder subjectieve variabelen niet gebruiken, zoals de Doppler kleurenscore, pijn tijdens het echografisch onderzoek, en het vermoeden dat de tumor van ovariale oorsprong is. Alle meerklasse modellen hadden goede performantie op de test set, met AUC waarden voor goedaardige tumoren en voor primair invasive tumoren tussen 0.93 en 0.94, voor borderline tumoren tussen 0.80 en 0.85, en voor metastatisch invasieve tumoren tussen 0.90 en 0.92. Vooral de performantie voor de moeilijke en kleine klasse van borderline tumoren is veelbelovend. Het LR-PC2 model heeft algemeen genomen de beste performantie, een bevinding die suggereert dat de R1U methode voor inputselectie goed werkt. Er kon opnieuw vastgesteld worden dat het LS-SVM model goede resultaten behaalt voor gewogen AUC maten. Modellen waar binaire modellen gecombineerd werden presteerden goed, wat dit naar onze mening een attractieve

methode maakt. Meer dan bij een all-at-once model wordt hier bovendien duidelijk in welke zin welke input belangrijk is.

Prospectieve evaluatie van de modellen voor ovariumtumoren

De binaire en meerklasse diagnostische modellen werden vervolgens prospectief getest op nieuwe data. De IOTA Ib dataset omvatte 507 patients van drie centra die ook deelnamen aan IOTA fase I. Deze data werden gebruikt om de modellen prospectief intern te evalueren. IOTA fase II, daarentegen, recruteerde patiënten bij zeven IOTA I centra en bij een aantal nieuwe centra, zodat een meer uitgebreide prospectieve interne zowel als prospectieve externe validatie mogelijk was. Aangezien de data van IOTA fase II op dit moment nog gecheckt worden, kon hier slechts een preliminaire dataset met 1916 patiënten gebruikt worden. Data voor deze patiënten waren volledig en juist. De evaluatie van de modellen op deze datasets liet zien dat de binaire en de meerklasse modellen zowel de prospectieve interne als prospectieve externe evaluatie heel goed doorstonden. De resultaten voor de prospectieve externe validatie waren zelfs een beetje beter dan die voor de prospectieve interne validatie. Voor binaire classificatie kan best een linear model zoals LR1, ObjLR, BLSSVM-lin of BPER11 gebruikt worden. ObjLR heeft het voordeel dat de subjectieve variabelen pijn en de Doppler kleurenscore niet gebruikt worden, BLSSVM-lin heeft het nadeel dat de gebruikte inputs niet optimaal zijn. Voor meerklasse classificatie is LR-PC2 de beste optie. Het is belangrijk te vermelden dat de meerklasse modellen ook in nieuwe centra goede AUCs behaalden voor borderline en metastatische tumoren, twee kleine en moeilijke klassen.

Het belang van CA-125 voor de diagnose van ovariumtumoren

Naast het ontwikkelen van predictieve modellen waren we ook geïnteresseerd in het belang van CA-125 bij het diagnosticeren van ovariumtumoren. We hebben deze informatie, die in de medische wereld gezien wordt als een belangrijke marker voor ovariumkanker, niet gebruikt in de predictieve modellen. De voorkeur ging uit naar modellen zonder CA-125, en het belang van deze variabele zou apart onderzocht worden. In [266] werd bovendien preliminaire evidentie gevonden, gebruik makende van de training data van IOTA fase I, dat CA-125 waarschijnlijk niet noodzakelijk is in predictieve modellen. Een eerste vraag was dan ook of CA-125 waarden werkelijk niet noodzakelijk zijn in mathematische modellen voor ovariumtumoren. Daarom werd in eerste instantie een logistiek regressiemodel ontwikkeld waar CA-125 per definitie gebruikt werd [387]. Hiervoor werden de patiënten uit de training set gebruikt waarvoor CA-125 waarden beschikbaar waren. Dit model werd vergeleken met LR1, het bekomen model wanneer CA-125 per definitie niet mocht geselecteerd worden als input voor het model, op de test set patiënten met CA-125 informatie. Hetzelfde werd gedaan voor de subgroepen van pre- en postmenopauzale patiënten afzonderlijk [387]. Het model zonder CA-125 was in beide gevallen opnieuw LR1,

ondanks het feit dat LR1 ontwikkeld werd op alle training set patiënten (i.e. pre- en postmenopauzale patiënten), zodat LR1 dus niet noodzakelijk een optimaal model is voor pre- of postmenopauzale patiënten afzonderlijk. De rationale hiervoor was als volgt: indien het nieuw bekomen model met CA-125 voor een subgroep op de test set niet beter presteert dan LR1, dat mogelijks geen optimaal model is, hoeft geen tijd besteed te worden aan het ontwikkelen van een nieuw model zonder CA-125 voor deze subgroep. Het bleek dat in geen enkele situatie (alle, premenopauzale, of postmenopauzale patiënten) het model met CA-125 beter was dan LR1.

Omdat CA-125 waarden niet beschikbaar waren voor ongeveer een kwart van de patiënten uit IOTA I, en omdat het nuttig is om de analyse over te doen met behulp van meer complexe modellen dan logistieke regressiemodellen, werd een vervolgstudie opgezet [132, 131]. Hier werden de ontbrekende CA-125 waarden geïmputeerd met behulp van vier technieken (*expectation-maximization*, *data augmentation*, regressie-imputatie en *hot-deck* imputatie) zodat vier 'opgevulde' versies van deze variabelen bekomen werden. Vervolgens werden modellen met en zonder CA-125 ontwikkeld met behulp van LS-SVMs of Bayesiaanse perceptrons. Zowel voor de LS-SVM als de Bayesiaanse perceptron modellen werden vier modellen met CA-125 ontwikkeld: een per opgevulde versie. De bekomen modellen werden opnieuw geëvalueerd op de test set. De resultaten waren gelijklopend aan die van de initiële studie: modellen met CA-125 waren niet beter dan modellen zonder CA-125.

Een tweede vraag was wat het verschil was tussen de performantie van CA-125 en de subjectieve opinie van de arts in het voorspellen van kwaadaardigheid. Hier bleek dat de subjectieve opinie van de arts veel beter in staat was ovariumtumoren te beoordelen dan om het even welke drempelwaarde voor CA-125 [424]. Dit bleef onveranderd indien enkel naar pre- of postmenopauzale patiënten gekeken werd. Er waren twee goedaardige histologieën met een verhoogde mediaan voor CA-125 in vergelijking met de andere goedaardige histologieën: endometriomas en abcessen. Anderzijds hadden borderline kwaadaardige tumoren (van stage I) een lagere mediaan in vergelijking met andere soorten kwaadaardige tumoren. Voor deze drie groepen tumoren bleek de arts correct in de diagnose van goed- of kwaadaardigheid in 90% van de gevallen, terwijl CA-125 amper 38% van deze tumoren goed beoordeelde (hier werd kwaadaardigheid voorspeld als CA-125 de drempelwaarde van 30 U/ml bereikte). Ook bleek dat een combinatie van subjectieve opinie en CA-125 niet tot betere resultaten leidde dan subjectieve opinie alleen [410].

Predictieve modellen voor zwangerschappen van onbekende lokatie

Meerklasse modellen werden ontwikkeld om een vroege diagnose te stellen van zwangerschappen van onbekende lokatie. Initiële meerklasse logistieke regressiemodellen werden ontwikkeld op een deel van de dataset (376 patiënten) [91, 90]. Een eerste model maakte een voorspelling aan de hand van de hCG ratio (hCG

waarde op 48 uur gedeeld door hCG waarde bij presentatie) en de gemiddelde hCG waarde. Het model bevatte ook het kwadratische effect van de hCG ratio. Voor een tweede model werd ook gekeken naar klinische informatie (zoals de mate van vaginale bloeding en verschillende pijnindices) om na te gaan of deze informatie de voorspelling kan verbeteren. Het finale model omvatte de hCG ratio, de gemiddelde hCG waarde, het kwadratische effect van de hCG ratio, en de mate van vaginale bloeding. De AUCs van beide modellen op het testgedeelte van de data waren 0.90 en 0.91. Dus, klinische informatie lijkt niet cruciaal, ook al werd de mate van vaginale bloeding geselecteerd in het tweede model. Het is belangrijk om te vermelden dat de progesteron waarden niet in rekening gebracht werden door de hoge correlatie tussen de hCG ratio en progesteron.

Een meer uitgebreide analyse werd uitgevoerd op de volledige dataset bestaande uit gegevens van 1003 vrouwen met een zwangerschap van onbekende lokatie. De modellen die ontwikkelde werden waren een meerklasse logistiek regressiemodel (MLR), een combinatie van 1-vs-1 logistieke regressiemodellen met behulp van paarsgewijze koppeling (LR-PC), een Bayesiaans meerlaags perceptron, een Bayesiaans perceptron, de combinatie van binaire Bayesiaanse LS-SVM modellen aan de hand van paarsgewijze koppeling of END (als kernel functie voor de binaire LS-SVMs werd ofwel de lineaire kernel ofwel de RBF kernel gebruikt, zodat in totaal vier meerklasse LS-SVM modellen ontwikkeld werden), en een meerklasse kernel logistiek regressiemodel (MKLR). Dit keer werden de progesteron waarden wel meegenomen, meer bepaald door middel van hun gemiddelde.

Inputselectie gebeurde apart voor de verschillende modellen maar leidde tot erg gelijklopende selecties. De hCG ratio, de gemiddelde hCG waarde, de gemiddelde progesteron waarde, de mate van vaginale bloeding, leeftijd, en voorgeschiedenis van ectopische zwangerschappen bleken de belangrijkste variabelen te zijn. Voor beide logistieke regressiemodellen werd de hCG ratio log-getransformeerd, werd ook een kwadratisch effect van leeftijd gebruikt, en werd een interactie tussen de hCG ratio en de gemiddelde progesteron waarde gebruikt. De test set AUCs voor falende zwangerschappen lagen rond 0.99 voor alle modellen. Voor intra-uteriene zwangerschappen lagen ze tussen 0.98 en 0.99, en voor ectopische zwangerschappen tussen 0.87 en 0.92. Opnieuw bleek dat LS-SVM modellen goede resultaten behaalden voor de gewogen AUC maten. Algemeen genomen had het LR-PC model de beste performantie. Andere modellen die goed presteerden waren MLR, de combinatie van binaire LS-SVM modellen met RBF kernel via END, en MKLR, hoewel dit laatste model matige resultaten haalde voor de gewogen AUC maten. De lineaire modellen presteerden slechter dan de niet-lineaire modellen, wat reeds gesuggereerd werd door het belang van een kwadratisch effect en een interactie-effect in de logistieke regressiemodellen. Voorts werd eens te meer duidelijk dat falende en intra-uteriene zwangerschappen erg goed te voorspellen zijn, maar dat het voorspellen van ectopische zwangerschappen moeilijker is terwijl dat de belangrijkste klasse is. Deze modellen hadden ook een betere performantie dan de initiële

meerklasse logistieke regressiemodellen.

Ten laatste werd onderzocht of hCG informatie voldoende was voor de diagnose van zwangerschappen van onbekende lokatie, dan wel of progesteron informatie en eventueel nog andere informatie ook echt nodig zijn. Dergelijke informatie werd wel belangrijk bevonden door de inputselectie analyses, maar uiteindelijk is de AUC performantie belangrijker. Daarom werden meerklasse kernel logistieke regressiemodellen ontwikkeld met drie verschillende input sets: enkel hCG informatie, hCG en progesteron informatie, en het volledige model met hCG en progesteron informatie, mate van vaginale bloeding, leeftijd en voorgeschiedenis van ectopische zwangerschappen. De dataset werd 100 keer opgesplitst in een training set en een test set, en op elke training set werden de drie MKLR modellen ontwikkeld waarna ze werden toegepast op de test set. Dit resulteerde per model in 100 AUC waarden voor elke klasse. Deze 100 waarden werden samengevat via de mediaan. Er waren weinig verschillen tussen de drie modellen wat betreft de mediaan AUCs voor falende en intra-uteriene zwangerschappen. De mediaan AUC voor ectopische zwangerschappen was 0.89 voor de MKLR met enkel hCG informatie, 0.91 voor de MKLR met hCG en progesteron informatie, en 0.93 voor de MKLR met alle inputs. Op basis van de 100 test set resultaten werden ook gemiddelde ROC curves opgesteld. Deze curves laten zien dat de mediaan AUC voor ectopische zwangerschappen een beetje groter is voor het volledige model in vergelijking met het model dat enkel hCG en progesteron informatie gebruikt, maar de verschillen tussen de gemiddelde ROC curves situeren zich op minder belangrijke plaatsen van de curve. Daarom is de conclusie dat hCG informatie alleen reeds heel veel informatie bevat om een diagnose te stellen, dat progesteron informatie toegevoegde waarde heeft, maar dat extra informatie waarschijnlijk niet noodzakelijk is. Daartegenover staat dat de gebruikte extra informatie eenvoudig te bekomen is.

Conclusies en verder onderzoek

Uit de analyses blijkt dat ovariumtumoren en zwangerschappen van onbekende lokatie accuraat kunnen gediagnosticeerd worden door de predictieve modellen. Lineaire modellen bleken even goed te werken als niet-lineaire modellen voor ovariumtumoren, maar niet voor zwangerschappen van onbekende lokatie. Het lijkt erop dat logistieke regressiemodellen, indien ze zorgvuldig worden opgesteld, kunnen resulteren in predictieve modellen met performantie gelijkwaardig aan de performantie van geavanceerde modellen zoals LS-SVMs. Logistieke regressiemodellen zijn eenvoudiger, maar het nadeel is dat de inputselectie (o.a. evalueren of er kwadratische effecten of interactie-effecten nodig zijn) erg doordacht moet gebeuren en dus veel tijd vraagt. De Bayesiaanse LS-SVM modellen hadden door de band genomen echter bijzonder goede resultaten voor gewogen AUC maten. Dit is naar onze mening een belangrijk argument voor het gebruik van dit soort modellen.

Vaak werden zowel all-at-once meerklasse modellen als combinaties van binaire

modellen gebruikt. De tweede soort van modellen kunnen ervoor zorgen dat inputs enkel gebruikt worden waar ze nodig zijn. Onze resultaten tonen aan dat het combineren van binaire modellen goede resultaten oplevert, minstens even goed als de resultaten voor de meerklasse modellen. Het nadeel is dat combinaties van binaire modellen meer werk vragen dan meerklasse modellen.

Belangrijke resultaten wat betreft het uitbouwen van de diagnostische modellen tot CDS systemen zijn de goede prospectieve performantie van zowel de binaire als de meerklasse modellen voor de diagnose van ovariumtumoren. Dit gold zowel voor de prospectieve interne evaluatie, als voor prospectieve externe evaluatie. Dit zijn belangrijke bevindingen voor de klinische praktijk.

Wat verder onderzoek betreft, zijn er voor beide gynaecologische toepassingen plannen om het belang van nieuwe markers te onderzoeken, zoals hormonale informatie of indices uit proteomics en genomics analyses. Daarnaast is het interessant om na te gaan in welke mate combinaties van modellen door middel van *ensemble*- of *fusion*-technieken de performantie kan verbeteren. Voor de bestaande modellen zal onderzocht worden of er verschillen tussen klinische centra zijn wat betreft de optimale omzetting van de kansschattingen in een zwart-wit voorspelling. Dit is belangrijk om de kwaliteit van de zwart-wit voorspellingen te verbeteren.

Met betrekking tot de modellen voor ovariumtumoren, is het belangrijk om een uitgebreide evaluatie van de modellen door te voeren op de data van IOTA fase II. Ook kunnen de data van fase I, Ib en II samengevoegd worden tot een dataset met gegevens over ongeveer 3500 patiënten. Deze dataset kan gebruikt worden om nieuwe, nog meer betrouwbare modellen te maken. Die kunnen dan getest worden op de data van IOTA fase III, die binnenkort opgestart wordt. Die dataset zal ook gebruikt worden om een aantal tweedelijnstesten te evalueren die toegepast kunnen worden op patiënten waarvoor de predictieve modellen onzeker zijn. Ten laatste is het heel nuttig om de modellen te vergelijken met artsen die minder ervaring hebben, en om na te gaan of de modellen tastbare voordelen bieden in de praktijk. Voorbeelden van voordelen zijn het reduceren van morbiditeit, financiële kosten en hospitalisatieduur, of het verbeteren van de prognose voor mensen met ovariumkanker.

Voor zwangerschappen van onbekende lokatie zal verder onderzoek gebruik maken van de IPULA dataset. Die kan gebruikt worden om de predictieve modellen te evalueren en om nieuwe modellen te maken gebaseerd op multicentrische data.

Glossary

Acronyms and abbreviations

5CV	Five-fold cross-validation
AIC	Akaike information criterion
AOT	Area of triangle
AR	Average rank
ARD	Automatic relevance determination
AUC	Area under the ROC curve
BIC	Bayesian information criterion
BLSSVM	Bayesian least squares support vector machine
BMLP	Bayesian multi-layer perceptron
BPER	Bayesian perceptron
CDS	Clinical decision support
CI	Confidence interval
CV	Cross-validation
END	Ensembles of nested dichotomies
EP	Ectopic pregnancy
GAM	Generalized additive models
hCG	Beta human chorionic gonadotrophin
IOTA	International Ovarian Tumor Analysis
IPULA	International Pregnancy of Unknown Location Analysis
IUP	Intra-uterine pregnancy
KDD	Knowledge discovery in databases
KLR	Kernel logistic regression
LASSO	Least absolute shrinkage and selection operator
LR	Logistic regression
LR+	Positive likelihood ratio
LR−	Negative likelihood ratio
LS-SVM	Least squares support vector machine
MKLR	Multi-class kernel logistic regression
MLP	Multi-layer perceptron
pAUC	Partial area under the ROC curve
PC	Pairwise coupling

Pm(κ)	Estimated probability of being among the κ best models
prAUC	Probabilistic area under the ROC curve
PUL	Pregnancy of unknown location
R1U	Input selection using rank-one updates for LS-SVMs
RBF	Radial basis function
RMI	Risk of malignancy index
ROC	Receiver operating characteristic curve
RVM	Relevance vector machine
sAUC	Scored area under the ROC curve
Sens75	Sensitivity at a specificity of 0.75 (or 75%)
SVM	Support vector machine
TL	Total length between probability vectors and class corners
TVS	Transvaginal sonography
VIF	Variance inflation factor
VUS	Volume under the surface
wVUS	Weighted volume under the surface

Contents

Voorwoord **i**

Abstract **iii**

Korte inhoud **v**

Nederlandse samenvatting **vii**

Glossary **xix**

Contents **xxi**

1 General Introduction **1**
1.1 Overview of the knowledge discovery process 1
1.1.1 Machine learning, statistics, and related concepts 1
1.1.2 A small taxonomy of learning algorithms 3
1.1.3 The process of applying computational algorithms to real-world problems 4
1.1.4 Applications in medicine: aims and considerations 6
1.2 Aims of this thesis 8
1.2.1 Prediction of ovarian tumor malignancy 8
1.2.2 Prediction of pregnancies of unknown location with a focus on ectopic pregnancy 9
1.2.3 Focus on probabilistic predictions 10
1.2.4 Secondary aims 10
1.3 Structure of the thesis and overview of personal contributions 12

2 Classification algorithms and model evaluation **17**
2.1 A note on overfitting and input selection 17
2.2 Methods for supervised classification 18
2.2.1 Logistic regression 19
2.2.2 Multi-layer perceptrons 27
2.2.3 (Least squares) support vector machines 32
2.2.4 Relevance vector machines 37
2.2.5 Kernel logistic regression 38

2.2.6 Software . . . 40
2.3 Evaluating model performance . . . 41
2.3.1 Performance measures . . . 41
2.3.2 Evaluating classifiers or algorithms . . . 45
2.3.3 Model comparison . . . 46
2.3.4 Calibration . . . 48
2.3.5 Prospective validation . . . 48
2.4 Conclusion . . . 49

3 Ovarian tumor and PUL diagnosis: background, aims, and data 51
3.1 Previous research . . . 51
3.1.1 Ovarian tumors . . . 51
3.1.2 Pregnancies of unknown location . . . 55
3.2 Available data . . . 60
3.2.1 Ovarian tumors: the IOTA group . . . 60
3.2.2 PULs: from St. George's to IPULA . . . 68

4 Issues in probabilistic multi-class classification 71
4.1 Comparison of methods for multi-class probabilities . . . 71
4.1.1 Introduction . . . 71
4.1.2 Obtaining multi-class probability estimates . . . 72
4.1.3 Experimental Setup . . . 77
4.1.4 Results and Discussion . . . 78
4.2 Weighted multi-class AUC metrics . . . 81
4.2.1 Introduction . . . 81
4.2.2 Weighted AUC metrics . . . 82
4.2.3 Weighted multi-class AUC metrics . . . 83
4.2.4 A real world application . . . 87
4.2.5 Discussion and conclusions . . . 91

5 Diagnostic models for ovarian tumors 95
5.1 Introduction . . . 95
5.2 Binary diagnostic models . . . 96
5.2.1 Logistic regression models . . . 96
5.2.2 Advanced models . . . 102
5.2.3 Model evaluation on the test set . . . 108
5.2.4 A simple alternative: simple rules for immediate use . . . 113
5.3 Multi-class diagnostic models . . . 115
5.3.1 Logistic regression models . . . 115
5.3.2 Advanced models . . . 119
5.3.3 Model evaluation on the test set . . . 124
5.4 Discussion and conclusions . . . 127

6 Prospective evaluation of diagnostic classifiers for ovarian tumors 131
6.1 IOTA phase I: prospective evaluation of old models 131
6.1.1 The experimental setup 131
6.1.2 Results 132
6.2 IOTA phase Ib: prospective internal evaluation of IOTA models . . . 133
6.2.1 Binary classifiers 133
6.2.2 Multi-class classifiers 138
6.3 IOTA phase II: prospective internal and external validation 141
6.4 Conclusions 145

7 The role of CA-125 for ovarian tumor diagnosis 147
7.1 Introduction 147
7.2 The importance of CA-125 in mathematical models 148
7.2.1 Round one: Logistic regression models based on complete cases 148
7.2.2 Round two: LS-SVMs and perceptron models based on imputed data sets 154
7.3 CA-125 versus subjective ultrasound evaluation 159
7.3.1 Introduction and experimental setup 159
7.3.2 Results and discussion 159
7.4 Conclusions 162

8 Diagnostic models for pregnancies of unknown location 163
8.1 Introduction 163
8.2 Initial logistic regression models 164
8.2.1 M4: the model using hCG information 164
8.2.2 M5: does clinical information matter? 165
8.2.3 Test set performance and conclusions 166
8.3 New diagnostic models using a large set of PULs 166
8.3.1 Algorithms and input selection 167
8.3.2 Model evaluation 169
8.3.3 Results 169
8.3.4 Does hCG information suffice or do we need more? 176
8.4 Discussion and conclusions 179

9 Conclusions and further research 181
9.1 Overview of results 181
9.2 Future research 184

Appendix: Description of IOTA variables 187

Bibliography 192

Publication list 223

Curriculum Vitae 237

Chapter 1

General Introduction

Abstract – This introductory chapter starts with a general overview of the process in which mathematical algorithms are applied to real-world data in order to acquire knowledge to solve a particular problem. The focus is on real-world medical data. The problem that lies at the basis of the knowledge acquisition process can generally be aimed at either prediction or description. In prediction problems, we seek for knowledge that can be used to predict elements of newly encountered situations. In description problems, we seek knowledge that is able to explain a given situation or phenomenon. After this overview, the aims of this thesis as well as own contributions are presented.

1.1 Overview of the knowledge discovery process

The application of machine learning, pattern recognition, statistics, and computer science to medical research questions has been a growing area of research. They can help clinicians to identify structure in the data they collected from patients. They can help to identify relationships between markers and disease and as such can help clinicians in diagnosis, prognosis, and personalized healthcare. We will now elaborate shortly on this practice of application. We will discuss (i) the research areas delivering computational algorithms that can be applied to real-world data, (ii) the ideal nature of the application process including the factors determining its success or failure, and (iii) the specific practice of applying computational intelligence for answering medical research questions.

1.1.1 Machine learning, statistics, and related concepts

First, what is meant with terms such as machine learning, pattern recognition, statistics, and computer science? How do these concepts relate to each other? What activities do they entail? Various definitions of these concepts have been given [128, 138, 150, 244, 279, 284]. It is a fact that there is overlap in their meaning. We believe it is a worthwhile exercise to look for a framework in which all of these

concepts can be placed in order to see similarities and differences between them.

A good point to start is the concept of Knowledge Discovery in Databases (KDD) [59, 150]. KDD is the general process aimed at the discovery of useful and interesting knowledge and patterns from data. The very first step of knowledge discovery is to get acquainted with the domain of application. We have to understand the field in which we want to gain knowledge and we need to collect all information (prior knowledge) that is necessary to start the process. Then, we need to obtain a data set to extract our knowledge from. The data set needs to be clean and preprocessed such that it is ready to work on. This step may entail noise or artifact removal and missing data handling. A next step is that of data reduction and projection. We may transform variables, create new variables based on existing ones, and apply dimensionality reduction or variable selection techniques such that the number of effective variables is decreased. At this point, the data mining step can start. Data mining can be defined as the application of specific algorithms to extract patterns from data [150] or, more generally, as the use of data to improve subsequent decision making [283]. We need to choose a data mining method (e.g. classification or regression) and a specific algorithm within the chosen method. As Fayyad and colleagues [150] point out, the choice of algorithm should also reflect whether the end-user is interested in the description of revealed relationships and patterns, or also in prediction of future cases using the revealed patterns. The chosen algorithm is then applied to the data, which is followed by the crucial tasks of interpretation and consolidation. The patterns discovered by the algorithm need to be explained and, if possible, visualized, and the new knowledge needs to be consolidated by implementing it in a system for further use (e.g. clinical decision support systems as explained in section 1.1.4) or by documenting it (e.g. by publishing in scientific journals). The KDD process may obviously have loops, and different steps are not always clearly discernable in practice.

Concepts such as pattern recognition, statistics, machine learning, computer science, and artificial intelligence, are mainly related to the data mining phase of the KDD process. When considering definitions of pattern recognition, it is our impression that it mainly refers to one possible strand of the data mining phase in the KDD process. Duda and colleagues [138] define it as the act of taking in raw data (pattern) and making an action based on the pattern's category. Therefore, pattern recognition usually deals with the development of a system to classify patterns into categories. Such a system can be obtained using a set of patterns for which the category is known and the aim is to find a system that classifies patterns into the correct category (supervised learning), or a set of patterns without category information for which a list of possible categories is derived based on relationships between the set of available patterns (unsupervised learning). Duda and colleagues [138] present the overall process to be followed in order to obtain a pattern recognition system. Five steps are identified: data collection, input selection, algorithm selection, classifier training, and classifier evaluation. This process closely

resembles the KDD process.

Different types of algorithms exist to extract knowledge from data. Michie and colleagues [279] distinguish the following types: statistical algorithms, machine learning algorithms, and neural networks. Because neural networks are often considered as a class of machine learning algorithms, we will not regard them as a separate category. Michie and colleagues [279] report an interesting, and in our opinion crucial, distinction between statistical and machine learning approaches. Statistical approaches usually assume human intervention with respect to issues such as input selection, variable transformation, and algorithm application. Statistical approaches are typically based on an underlying model that is valid only if some assumptions hold. It is the responsibility of the statistician to check the assumptions. On the other hand, machine learning approaches are designed to operate with minimal human intervention. Machine learning can be defined as a domain that focuses on the development of algorithms that are able to learn from experience [244]. Mitchell [283, 284] sees machine learning as a part of computer science that emerged at the intersection of computer science and statistics. Whereas computer science is concerned with building machines to solve problems, and statistics with reliable inference from data using modeling assumptions, machine learning is concerned with making machines able to make inference from data themselves in the most efficient way.

Machine learning, as well as statistics, plays an important role in data mining. These research domains develop algorithms that can be used in the data mining phase of a KDD process. However, the term data mining also refers to the particular interest in methods to extract knowledge from massive data sets. Some of these algorithms are appropriate for pattern recognition purposes, while others focus on regression problems or dependency analyses. Since knowledge is extracted from data using computers and mathematical theories, statistical and machine learning algorithms are one example of artificial intelligence (AI). Honavar [191] describes AI as the domain that works towards building and understanding intelligent systems. It focuses on computational models having intelligent behavior (perception, reasoning, learning). In other words, it deals with the computerization of human intelligence in successfully performing various tasks [191, 253].

1.1.2 A small taxonomy of learning algorithms

Introductions to various machine learning and statistical algorithms can be found in, among others, [51, 114, 138, 186, 217, 298, 371, 453].

Learning algorithms can be ordered in several ways. One ordering distinguishes between supervised, unsupervised, semi-supervised, transductive, and reinforcement learning. In supervised learning, the algorithm receives training data consisting of n samples for which both the input measurements $\mathbf{x}_n$ and the output measurement

y_n are available. The algorithm learns a function to link the inputs with the output. If $\mathbf{x}_n$ and y_n are generated by, respectively, the random variables $\mathbf{X}$ and Y, with joint probability density $p(\mathbf{X}, Y)$, then supervised learning can be seen as a density estimation problem where one is interested in the conditional density $p(Y|\mathbf{X})$ [186]. One is interested in a function that generalizes well to new data rather than in a function that fits the training data as good as possible (generalization versus overfitting). In the latter case, the algorithm also models the specific random noise of the training data such that the function will perform poor when applied to new samples. In unsupervised learning, the algorithm only has the input measurements (i.e. unlabeled data). The algorithm fits a model to determine the organization or clustering in the data. In semi-supervised learning, the algorithm receives both labeled and unlabeled data and learns a function to link the inputs with the output. Usually most of the data are unlabeled. In transductive learning, the algorithm receives a set of labeled samples and a set of unlabeled samples for which the output needs to be predicted. The aim of transductive learning is to improve model performance by adding the inputs of the cases for which the output is to be predicted to the training data. In this type of learning, one is not interested in the function that links the inputs to the output. Finally, reinforcement learning involves learning how to act in a given environment. An action is followed by either positive or negative feedback that helps learning.

In another ordering, one distinguishes between algorithms for classification, regression, clustering, dependency modeling, and summarization [150]. In classification algorithms, a function is learned to map the inputs to one of several classes. Binary classification involves two classes while multi-class classification involves three or more classes. The output represents the class membership of a sample. In regression algorithms, the inputs are mapped to a real-valued output. Clustering algorithms work with unlabeled data and try to find underlying structure in the data by looking for the best way to divide the training samples into clusters. Dependency modeling involves learning the dependencies between a given set of measurements. Such algorithms can also perform classification or regression tasks by feeding both the inputs and the output to the algorithm (e.g. [162]). Finally, summarization algorithms look for a summarizing description of a set of data. Density estimation techniques are an example of this category.

1.1.3 The process of applying computational algorithms to real-world problems

Applying algorithms stemming from machine learning or statistics to specific research problems is important. In our opinion, only successful applications can justify the usefulness of the broad research area of computational intelligence. Moreover, applications can help to give this research area a human face, by pointing at and helping to resolve practical issues that hampered successful application of computational algorithms [333]. A constant interaction between applied and

fundamental research is vital [333, 66].

It is important to note that an application study does not merely involve feeding some data to some algorithm [345]. It is a complex and iterative process that is supposed to be embedded in the domain of application, and that involves multiple agents [345, 150, 178, 66]. The KDD process is a good example of the lay-out of an application process. Similar process designs were proposed by Saitta and Neri [345] and Brodley and Smyth [66]. All of these process designs stress the iterative nature of the process. Observations at any point in the process may force one to go back in the process to revise previous steps. Three different roles or agents are typically discerned: there is the model developer whose main task concerns the computational aspects of the project, the end-user who should use the results of the process in everyday practice, and the domain expert who assists the model developer such that all analyses are matched with relevant domain-specific knowledge. These three roles are not always represented by three different persons (or groups), it is possible that one person represents more than one role. For example, the end-user and the domain expert roles may often be played by the same person or group of persons. The necessity of end-user involvement is also stressed by Hand [178]. Embedding the analysis process in the specific domain of application is also crucial in all process designs. When applying algorithms to a data set without knowing of relevant domain knowledge, results tend to be suboptimal since they will not be tuned towards the domain-specific problems and vagaries.

To summarize, all process designs start with understanding the problem, specification of the project goals, and acquisition of relevant domain knowledge (see also the description of the KDD process). The next step involves data collection, the success of which also depends on specific domain-related issues. The data then have to be cleaned, preprocessed, and reduced. When the data are ready for use, the specific algorithm to apply to the data has to be chosen and used. The results have to be interpreted and their performance evaluated. The final phase entails the implementation of the newly gathered knowledge into real-life practice, eventually resulting in routine use. Lasting in-field use is the only way to truly prove effectiveness of the extracted knowledge.

Several comments are given concerning the choice of algorithm. Brodley and Smyth [66] state that an algorithm is characterized by three components: the functional form of the models that are being used by the algorithm, estimation (concerning the criterion to measure the quality of parameter estimates), and search (the process used to find parameter estimates). The first component seems most important for algorithm selection. It encompasses the important issue of understandability of an algorithm's results. This is often very important for an end-user, but it may come at the expense of performance. Therefore, a better performing but less interpretable algorithm may be preferred. Also, linking algorithms to specific data and problems is still an open area. One choice could entail that several

algorithms are used and their results combined [345]. Finally, the choice of algorithm can be influenced by the project's goals such as hypothesis-confirming versus hypothesis-generating [178], description or explanation of underlying processes versus prediction of a condition [178, 150].

Saitta and Neri [345] focus on the difference between 'ready-to-use' data versus 'real-world' data by mentioning that applied research is not merely feeding ready-to-use data to some algorithm. Ready-to-use data such as the benchmark data sets from the UCI repository [27] must be distinguished from real-world data. Testing algorithms on UCI data sets is not wrong, it can result in useful insights about the algorithms. But ready-to-use data sets do not correctly reflect real life: one does not know exactly how the data were collected and what preprocessing steps were undertaken, what inclusion criteria were, what was done with missing values, and domain knowledge is ignored. They are useful in theoretical research, not for applications.

1.1.4 Applications in medicine: aims and considerations

The last step of the KDD process involved the consolidation of the newly obtained knowledge. In medical research, as well as in other research areas, there are two main options for consolidating knowledge: documenting and implementing. The results may only be written down for publication such that it is made public for those interested. The documentation of findings may lead to new research questions, or may be picked up by people who can put it into practice (e.g. in policy making).

Implementing obtained results in medical research results in clinical decision support systems (CDS systems, also called computerized decision support systems even though it is usually yet not necessarily computerized). CDS systems aim to help clinicians with decision making by matching patient characteristics to a knowledge base in order to give patient-specific recommendations [88, 155, 197, 474, 475]. CDS systems can for example be used to detect abnormalities in patient signals such as ECG recordings, help in patient diagnosis or prognosis, assist in therapy planning and monitoring, prescribe medication, or retrieve relevant information in the literature [473, 197, 88]. The first CDS systems emerged in the fifties (e.g. [296, 190]), and numerous systems have followed ever since [281]. Miller [281] states that CDS systems use an imperfect model of the highly complicated and not yet fully understood process of medical diagnosis. This may suggest that the CDS system can replace the medical expert, but this is obviously not the case. They serve to assist the experts by suggesting a diagnosis based on knowledge extracted from the data on which they were developed. Both the suggested diagnosis (or the system's estimated probability of a diagnosis) and the knowledge on which the system is based can help the clinician in making better decisions. It is obvious that the nature of CDS systems has evolved alongside evolutions in electronics (e.g. the introduction of desktop computers, laptop computers), telecommunication (e.g.

local networks, internet facilities), machine learning and statistics (e.g. development of new algorithms such as support vector machines).

Yet, despite the clearly promising nature of CDS systems and the numerous available CDS systems in the literature, it has been questioned why their use does not spread as fast as might be expected [87, 88, 155, 263, 474]. Two main reasons have been suggested for the gap between CDS system development and clinical use. First, the development of CDS systems was not always guided by clinical needs (e.g. [43, 87, 88, 155]). As explained in the previous sections, interaction with the end-user and a good understanding of the problem are vital for the future success of a CDS system. We need to examine for what tasks clinicians need assistance, and what kind of assistance they want. Based on their own experience, Bates and colleagues [43] state that CDS should fill up clinical needs, and should be fast, simple and easy to use. Further, it should nicely fit the clinician's workflow such that it does not take too much additional time [43, 88, 155]. These issues can improve the 'clinical credibility' of a CDS system, which is an important factor [474]. Note that the use of complex mathematical models such as artificial neural networks or support vector machines, for which it may be difficult to understand how the model uses patient characteristics to arrive at a conclusion, may hamper the clinical credibility. However, a possible improvement in performance may compensate for the lack of understandability of the model. Moreover, there is research on rule extraction methods to make such models more understandable [263]. Nevertheless, some clinicians are skeptical towards any CDS system because they cannot believe that a mathematical system can perform better than humans. Nevertheless, reports of the opposite exist in the literature [474].

Second, a rigorous evaluation of CDS systems is not always performed (e.g. [263, 474]). Wyatt and Altman [474] state that there should be evidence of the CDS system's accuracy, generality, and effectiveness. For example, a system developed to give diagnostic assistance should be accurate as evidenced by a high sensitivity (a high percentage of people with a condition should be predicted to have the condition) and high specificity (a high percentage of people without a condition should be predicted not to have that condition). Evidence of accuracy should be obtained from a large test data set. Evidence of generality means that the system should be tested in different clinical centres and at different points in time. Characteristics of the CDS system's target population may differ geographically, and may change over time [180]. Therefore, a CDS system developed on a particular set of data may eventually become outdated, leading to decreasing performance. Consequently, the ideal yet unrealistic situation would be constant quality control to check for changes in performance. Finally, evidence of effectiveness should ideally be based on large clinical trials in which the clinical effectiveness is tested. Examples of clinical effectiveness are a decrease in unnecessary investigations, hospitalization time, or financial costs, improved prognosis, or an increase of patient satisfaction when the CDS system is used together with the expert opinion of the clinician.

Today's model of clinical practice is based on 'evidence-based medicine' [344]. In this model, patients are treated by a combination of the clinical expertise of the clinician and the appropriate use of the latest evidence from medical research. Even though this is a widely accepted model for clinical practice, it turns out that the research-based evidence is not always used when managing patients [43, 360]. CDS systems are considered useful instruments to improve the integration of evidence in the workflow of clinicians. As such, they are an important tool in the clinical decision making process.

1.2 Aims of this thesis

In general terms, the aim of the thesis is to apply knowledge discovery processes to medical problems aimed at prediction (i.e. medical decision making) and, to a lesser extent, description. The medical decision making problems involve two medical conditions: ovarian tumors and pregnancies of unknown location. The description problems deal with breast cancer and the role of prenatal maternal anxiety in the development of emotional dysregulation among adolescents.

For the projects on medical decision making (prediction), mathematical algorithms are applied to data sets in order to obtain a model that links the measurements to the outcome. The outcome is one of several classes (e.g. benign versus malignant ovarian tumors). The resulting models are labeled classifiers, mathematical models, diagnostic models, or predictive models, and these terms are used interchangeably.

1.2.1 Prediction of ovarian tumor malignancy

The term tumor refers to an abnormal growth of tissue that can be either benign or malignant. The term cancer is reserved for malignant tumors. Ovarian cancer is a lethal cancer. More specifically, it is the most lethal of all cancers affecting the female reproductive system. In 2004, 14.716 deaths due to ovarian cancer were reported in the United States [210]. Optimal treatment of ovarian tumors is important, but it depends on the type of tumor. For example, benign masses can often be treated conservatively or with minimally invasive surgery [73]. Adequate treatment decisions have beneficial consequences for the patient, such as improved prognosis after treatment for ovarian cancer [454]. However, histopathological examination of the tumor, such as an exploratory laparotomy, is the only way to obtain certainty about the type of tumor. The result of this is that suboptimal surgery is applied in suboptimal conditions to many cancerous tumors. This can be detrimental for the patient [222, 454, 146]. Therefore, early preoperative discrimination between benign and malignant tumors is crucial.

Recently, the International Ovarian Tumor Analysis (IOTA) group [395] initiated the collection of large multi-center data sets containing data from women with at least one adnexal (i.e. ovarian, para-ovarian, or tubal) tumor. The main aims concerned the development and prospective evaluation of mathematical models to predict tumor type. Four types (classes) of adnexal tumors were distinguished: benign, primary invasive, borderline, and metastatic invasive. The latter three classes are malignancies. Borderline ovarian cancers are primary invasive cancers of 'low malignant potential', representing less aggressive cancers that are less life-threatening. At first, the development of mathematical models dealt with the discrimination between benign and malignant tumors. It is also useful, however, to discriminate between all four types of malignancy. For example, borderline malignancies are less aggressive than other malignancies, and this difference can be important for deciding upon the best treatment for a specific patient. As is clear from the previous sections in this chapter, mathematical models need to be rigorously evaluated on new data in order to investigate if and how the classifiers can be of help in clinical practice.

Other, but related, aims of the IOTA group are the prospective evaluation of previously developed models by other research groups, the comparison of model performance with the performance of the expert clinicians who predict the tumor type based on the ultrasound images, and the investigation of the cancer antigen 125 (CA-125) tumor marker. CA-125 is a widely used marker for ovarian cancer. Elevated serum CA-125 levels are seen as a cause for concern, even though CA-125 is also raised in some benign conditions such as endometriosis while it may be low for some types of ovarian cancer [404]. The usefulness of the CA-125 tumor marker is a topic of discussion but nevertheless it is often used as a tool for diagnosis. IOTA aimed at investigating whether CA-125 is a necessary measurement for building classifiers and the investigation of the performance of CA-125 alone for predicting tumor malignancy.

1.2.2 Prediction of pregnancies of unknown location with a focus on ectopic pregnancy

A pregnancy of unknown location (PUL) can be defined as a situation in which there is a positive pregnancy test with no signs of intra- or extra-uterine pregnancy on ultrasound, and without remnants of miscarriage [102, 106]. Such pregnancies usually turn out to be one of three types. The majority represent failing pregnancies, which could be either ectopic or intra-uterine, but is never seen on ultrasound. Another large group of PULs are early intra-uterine pregnancies (IUPs). A third, smaller group of PULs represent ectopic pregnancies.

Because an ectopic pregnancy is located outside the uterus, there is a risk that it ruptures [211]. Therefore, such a pregnancy can be dangerous for the mother. Ectopic pregnancy is still an important cause of maternal deaths that

are directly related to pregnancy [257, 22, 224]. This makes the early and accurate diagnosis of ectopic pregnancy, also among pregnancies of unknown location, very important. Early detection also improves the effective non-invasive treatment of ectopic pregnancies. An expectant, wait-and-see approach has been demonstrated to be safe for women with an ectopic PUL [71, 174, 35]. This approach can be a serious psychological and financial burden, however, and the dangerous and unpredictable nature of ectopic pregnancies further emphasizes the need for early outcome prediction of this subgroup of PULs, such that the well-being of this group of women can be guaranteed.

The prediction of the outcome in a PUL population, however, is a great challenge in early pregnancy today. We focused on the development of multi-class diagnostic models in order to predict the pregnancy outcome. These methods have advantages relative to existing heuristic low-level approaches such as threshold rules or subjective interpretation by gynecological experts.

1.2.3 Focus on probabilistic predictions

We intend to produce probabilistic models to predict the type of ovarian tumor or to predict the outcome of PULs. Information regarding uncertainty is crucial for applications concerning medical decision making. When doctors need to make treatment decisions they want to take into account the degree of (un)certainty concerning the patient's specific condition. For example, assume that the clinician blindly follows the prediction of the mathematical model. The model makes black-and-white predictions and for a certain patient predicts that the tumor is benign. The clinician will be leaning towards minimally invasive treatments. But, assume a probabilistic model would consider the probability of malignancy to be 30%. The clinician may consider this probability too high to choose a low-level treatment.

1.2.4 Secondary aims

Breast cancer research

Breast cancer occurs mainly in women, but men can be affected too. Breast cancer is the most common type of cancer among women [210]. In the United States, 40.954 women died of breast cancer in 2004, making it the second most lethal of all cancers. The lifetime risk of a women to die from breast cancer is estimated to be 1 in 33 (3.0%) [18]. Important biomarkers of breast cancer, and pathways of tumor growth, are the hormone receptors for estrogen and progesterone (ER and PR) and the human epidermal growth factor receptor 2 (HER-2). When the cancer's receptors for estrogen or progesterone are overexpressed, it is called ER-positive or PR-positive. Such cancers are sensitive to these hormones, which stimulate cancer cell growth. This means that the cancer is hormone-dependent and prospers in a hormone-rich environment [63].

HER-2 is an oncogene producing a glycoprotein. Overexpression of the gene (HER-2 positive cancer) results in excess protein production. This is an important element in the pathogenesis of cancer [148]. Such cancers tend to be aggressive. All three biomarkers are known to have independent prognostic value [148]. Also, the fact that they are interrelated seems to be a mechanism underlying therapy resistance [314]. Thus, investigating how these markers are interrelated is necessary. Since there is suspicion that the associations are age-related [194], we used nonparametric regression techniques to assess the associations between the markers and the extent to which these are age-related.

Another important characteristic of breast cancers is their lymph node involvement. Lymph fluid is transported around the body through lymph vessels and lymph nodes. It is known that the presence of cancer cells in the underarm (axillary) lymph nodes (i.e. lymph node positive breast cancer) increases the likelihood of cancer spread to other parts of the body. The more nodes involved, the more aggressive the cancer. Therefore, we investigated independent risk factors for lymph node involvement of breast cancer.

Prenatal maternal anxiety, the HPA-axis, and emotional problems in adolescence

The 'developmental origins of health and disease' hypothesis states that the risk of developing certain diseases is not only determined by genetic and life-style factors, but also by one's early life environment [163]. Research indicates that early life causes of disease may come into play already during the prenatal and perinatal period (e.g. [36, 209, 327, 461]). The idea is that early environmental cues program systems that are under development [461]. The consequence is that systematic or stable dysregulations arise in the developing system, thus predisposing one for altered behavioral, cognitive, and/or somatic functioning.

One central nervous system subject to programming is the hypothalamic-pituitary-adrenal (HPA) axis. Among other activities, this axis controls levels of the stress hormone cortisol in relation to daily events and stress cues [208, 461]. Early life environmental aspects seem to dysregulate HPA-axis functioning, which in turn may affect emotional processing and may increase one's responsiveness to stress (e.g. [461, 308, 216]). Therefore, HPA-axis functioning may act as a mediator between early life stress and later depression. Early life stress is related to depression [188], but the intermediary role of the HPA-axis has hardly been studied. We investigated the effect of prenatal maternal anxiety on the offspring's HPA-axis functioning, and whether HPA-axis functioning mediated an effect of prenatal maternal anxiety and the offspring's depressive symptomatology at the age of 14 to 15 years.

Mood disorders such as depression are related to altered HPA-axis functioning [81]. Different disorders have been linked with different circadian cortisol profiles [123, 188]. We questioned whether such links also exist in nonclinical populations.

Therefore, we investigated the association between the circadian cortisol profile and symptoms of depressive mood, anxiety, and aggressive behavior in a nonclinical sample of adolescents.

1.3 Structure of the thesis and overview of personal contributions

We will now present a chapter-by-chapter overview of the contents of the thesis and of personal contributions. For each topic, we refer to all relevant journal papers and conference contributions that we published.

The methods from statistics and machine learning that were used in this thesis are described in Chapter 2. This chapter also discusses aspects of model evaluation. Chapter 3 starts with introductory information on ovarian tumors and pregnancies of unknown location. This introduction to both classification problems starts with clinical background knowledge of the medical condition, continues with a description of the research aims in the field, and ends with an overview of existing work on the topic. The second part of Chapter 3 describes the data that were used to pursue the research aims.

Chapter 4 discusses two topics related to probabilistic multi-class classification. First, a comparison study is presented in which several methods to obtain multi-class probabilities are applied and their performance compared [420]. All the methods applied are based on least squares support vector machines. A comparison study can point at the methods that are likely to give the best results, such that we can choose the best methods when developing classifiers in the following chapters. This work was published in *Lecture Notes in Computer Science*. Second, ideas towards weighted multi-class AUC metrics are presented [414]. The AUC index quantifies the amount of discrimination between classes that is obtained by a classifier: it measures to what extent the model is able to distinguish the classes of interest. It does not, however, take into account the specific probabilities generated for cases from different classes. Of course, we want the model to generate high probabilities for cases to belong to their true class, and low probabilities to belong to another class.

Chapter 5 gives an overview of mathematical models that were derived to classify different types of ovarian tumors. The initial classification problem deals with distinguishing between benign and malignant (i.e. cancerous) tumors. Some models for this problem were developed by colleagues [394, 20, 266, 426, 392], with minor contributions from our part. To compete with the basic logistic regression model [394], we developed more advanced models such as Bayesian neural networks [428, 427, 419], and an L_1-regularized logistic regression model. The neural network models are in press with *Neural Computing and Applications*. Finally, a logistic regression model using only objective inputs was developed. This is useful because

the use of objective measurements is likely to enhance a model's robustness and generalizing ability. Together with the paper discussing the models based on LS-SVMs and RVMs [426], we published an opinion paper on Bayesian statistics in *Ultrasound in Obstetrics and Gynecology* [421]. As an alternative to mathematical models that is easier to use by clinicians in everyday practice, we developed simple rules that clearly point at either benignity or malignancy [393]. The results of this study are in press with *Ultrasound in Obstetrics and Gynecology*. To extend the binary classification, multi-class models were constructed in order to differentiate benign, primary invasive, borderline malignant, and metastatic invasive ovarian tumors [431, 429, 430]. No such models were hitherto developed, even though the reliable identification of borderline and metastatic tumors would be a good step forward for clinical practice.

Chapter 6 presents three prospective evaluation studies of existing and newly derived mathematical models [445, 390, 447, 430, 443, 446, 444]. As delineated in Chapter 1, this is an important step towards the development of clinical decision support systems. The performance of the subjective assessment of the expert ultrasound operator is also investigated and compared with the performance of the models.

Chapter 7 presents the research on the CA-125 tumor marker. In today's clinical practice, this is a widely used marker when confronted with an ovarian mass. Elevated serum CA-125 levels, as obtained by a blood test, clearly suggest malignancy. However, there is controversy with respect to the necessity and usefulness of measuring CA-125, since it can lead to incorrect diagnoses. Moreover, the measurement is time consuming and expensive. Therefore, a thorough examination of the role of CA-125 for diagnosing ovarian tumors is crucial. We investigated the necessity of CA-125 for mathematical models by comparing models that use CA-125 with models ignoring CA-125 [387, 389, 412, 388]. This work was published in *Journal of Clinical Oncology*. Because CA-125 was not measured for all patients in the data set, we re-analyzed the data after applying several imputation techniques to fill up the missing values [132, 131]. Next, we compared the performance of CA-125 with that of an expert's subjective assessment after ultrasound examination of the tumor [413, 425, 424, 410, 409]. Some of the results of this analysis were published in *Journal of the National Cancer Institute*, and were reported on in the media [349].

Chapter 8 describes the mathematical models for PULs. No mathematical models to simultaneously classify the three types of PULs were developed before. Two initial models were derived using logistic regression analysis. First, using hormonal, demographic, and ultrasound data, a model was derived that only used information based on human chorionic gonadotrophin (hCG, the pregnancy hormone) levels [91, 96, 94]. Next, we investigated the added value of clinical information [90]. These models were published in *Ultrasound in Obstetrics and*

Gynecology and *Fertility & Sterility*, respectively. Then, in order to improve the quality and reliability of the diagnosis, mathematical models were derived using more complex algorithms and a larger data set [415]. Algorithms used are logistic regression, neural networks, least squares support vector machines, and kernel logistic regression. Awaiting a larger multi-centric prospective evaluation of all mathematical models, a prospective evaluation of one of the initial logistic regression models on data from other clinical centers was carried out [225, 240], as well as an prospective interventional study to evaluate the clinical advantages of using the mathematical model in practice [228, 236, 235]. The results of the interventional study were published in *Human Reproduction*.

Chapter 9 summarizes the work presented in the previous chapters, presents conclusions and important topics for further research. Figure 1.1 visualizes the chapter-by-chapter structure of the thesis.

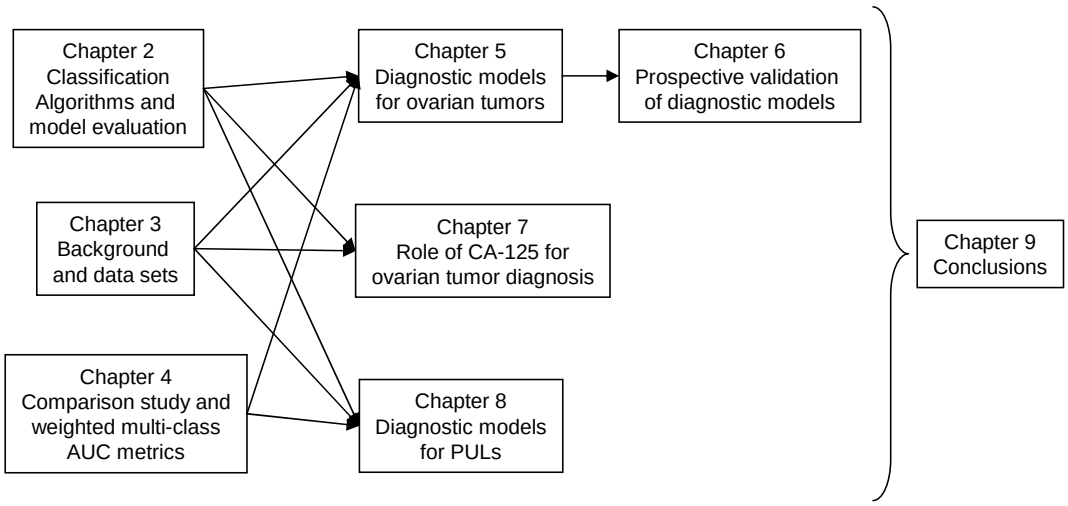


Figure 1.1: Overview of chapters.

The research on breast cancer and on the associations between prenatal maternal anxiety, the HPA-axis, and emotional problems in adolescence are not addressed in depth in this thesis. This work does not involve the development of mathematical models for predictions, but rather for description (cf. Chapter 1 for the distinction between these two purposes). For reasons of clarity of exposition and text coherence, we restrict the focus of this thesis to prediction. Therefore, we confine ourselves to a summary of our contributions to these areas in the next two paragraphs.

Concerning clinical results, we discovered that triple positive breast cancers (i.e. the receptor status for estrogen, progesterone, and HER-2 is positive) have more likely lymph node involvement [417]. This work is in press with *Breast Cancer*

Research and Treatment. Continuing this analysis, we used generalized additive models to visualize the nonlinear relationship between age and lymph node status [463]. We observed in our data that this nonlinear relationship is mainly apparent in small tumors, and this was confirmed in a prospective analysis using data from a center in Eindhoven [465, 464]. With the use of the same technique for nonlinear modeling, we reported on the nonlinear relationship between age and receptor status for estrogen, progesterone, and HER-2 [299, 300]. This latter analysis is in press with *Breast Cancer Research and Treatment*. The probability of an estrogen receptor positive breast cancer increases with age up to approximately age 50 after which the probability stagnates. For the progesterone receptor, the probability of it being positive increases up to age 55 after which the probability shows a modest decrease with age. However, these relationships seemed to apply only for HER-2 negative breast cancers. For HER-2 positive breast cancers, the probability of estrogen or progesterone receptor positivity shows a steady decrease with age. Also, the probability of a HER-2 positive cancer decreases with age only for cancers that are positive for both the estrogen and progesterone receptors. The result of these findings is that triple positive breast cancers appear at an earlier age than other breast cancers. On a statistical note, these results show that checking the linearity in the logit assumption in logistic regression can discover nonlinearity that is best addressed using a piecewise logistic model [418, 422]. These models should be part of a statistician's arsenal to tackle nonlinearity when using logistic regression analysis.

With respect to the effect of prenatal anxiety, we analyzed longitudinal data on 58 mother-child pairs analyzed with regression analysis for repeated measurements. We obtained evidence that prenatal maternal anxiety may increase the risk for depressive symptomatology in adolescence [437, 433, 438, 434, 435]. This effect was only found in girls, and the analysis suggests that HPA-axis dysregulation plays a mediational role in this effect. Thus, maternal anxiety may cause malfunctioning of the HPA-axis in girls, resulting in an increased neurobiological vulnerability of depressive mood in adolescence (aged 14 to 15 years in our data). The HPA-axis dysregulation was characterized by a flattened diurnal cortisol profile where cortisol was mildly lowered at awakening but elevated in the evening. The mediational role of the HPA-axis in the link between early life development and onset of disease later in life was not demonstrated before. These findings were published in *Neuropsychopharmacology*. The existence of mediation was assessed using both the classic approach by Baron and Kenny [41] and the recently suggested bias-corrected bootstrap method [271]. Additionally, we linked adolescents' cortisol profile at age 14-15 with indices of emotional distress such as anxiety, depressed mood, and aggression [436, 432]. Using k-means cluster analysis and indices to assist in choosing the optimal number of clusters such as the Calinski-Harabasz index [72], the silhouette value [340] and the Gap statistic [383], we discerned four clusters of adolescents based on their cluster profile. One cluster contained only three subjects and was characterized by very high evening cortisol. Notwithstanding the small sample size, we observed increased emotional distress in this group of adolescents:

there were more indications of anxiety and depressed mood. Alternatively, these data were analyzed using longitudinal regression analysis. It was observed that a flattened cortisol profile, as described above, was linked with more depressive symptoms and anxiety. When entering both depressive symptoms and anxiety in a model, the link between depressive symptoms and the flattened cortisol profile disappeared. This suggests that depressive symptoms are linked with a flattened profile because anxiety is a core part of depression. We did not find links between cortisol and aggression.

Finally, some of our contributions dealt with statistical consulting activities that do not fall within the scope of this thesis. These analyses do not deal with ovarian tumors or PULs, and/or do not entail mathematical modeling for either description or prediction. Therefore, we will merely list our contributions by topic: pregnancies of unknown location [100, 97, 92, 229, 107, 98, 230, 239], the availability of ultrasound in the acute gynecology unit (AGU) [175], treatment of ectopic pregnancy [93, 95, 238, 233, 234, 237], pelvic examination of women undergoing laparoscopic total hysterectomy [108], diagnosis of endometrial polyps and submucous fibroids [177], chlamydia trachomatis infections in women attending an AGU or an early pregnancy unit [176, 226], arthralgia following aromatase inhibitor based treatment in breast cancer [289], detection of sentinel lymph nodes in breast cancer [127, 291], breast cancer prognosis using an improved version of the Nottingham prognostic index [411], or prognosis of endometrial stromal sarcoma [17, 16]. Other activities dealt with an application of least squares support vector machines for function estimation to sports data [423].

Chapter 2

Classification algorithms and model evaluation

Abstract – In this chapter, we describe various algorithms for supervised classification based on generalized linear models, neural networks, support vector machines, and Bayesian approaches thereof. Both models for binary and multiclass classification are presented. Attention is also devoted to other aspects such as preprocessing and input selection. Finally, model evaluation and model comparison are discussed. Both of these issues are as important as the core task of classifier development.

2.1 A note on overfitting and input selection

It is not difficult to construct a model that fits the training data extremely well, but this is not the point of model building. The aim is to model the underlying structure in the data and not the idiosyncracies or specific vagaries - the noise - associated with the training data set. This is important for research aimed at prediction or explanation. In case of prediction, we need a model that also fits well when applied to new cases. Alternatively, when one aims to explain a phenomenon, fitting noise is obviously not improving our understanding of the phenomenon under study. Modeling training data noise is called overfitting. On the other hand, failing to uncover a part of the data's underlying structure is called underfitting. Neither of these is desirable, therefore we have to look for a simple model (avoiding overfitting) that still fits the data well (avoiding underfitting). This relates to the principle of parsimony. Basically, we are interested in a optimal trade-off between bias and variance [67, 186, 159]. Bias refers to the average closeness of a model's prediction to the true outcome, variance refers to the sensitivity of the predictions to the training data. An underfitted model will have large bias but small variance. An overfitted model will have small bias but large variance such that it does not generalize well to new data. To prevent overfitting, issues of regularization (keeping model parameters small to obtain a simple model) and input selection become relevant. Even though adverse effects of irrelevant inputs may be largely suppressed

by effective regularization, dropping irrelevant features is parsimonious and is a kind of regularization because it fixes the effect of these inputs at zero. Some algorithms perform both at the same time [382, 304].

In the context of linear models, it has been suggested that one should have around ten training data cases (or, in case of a binary outcome, cases belonging to the smallest class) per input considered (including polynomial and interaction terms) [184, 316, 365]. In the machine learning literature, similar recommendations exist [361]. Even though one can assume that linear models are less prone to severe overfitting than more advanced models such as neural networks or support vector machines, overfitting is an important issue when the dimensionality of the input space is large as compared to the size of the training data set (cf. the curse of dimensionality). This is a common situation in bioinformatics research [361]. Overfitting also becomes more of an issue when considering, in linear models, nonlinear inputs such as interaction terms. Model selection is therefore necessary in order to achieve a parsimonious model that models the underlying structure.

Input selection makes a model parsimonious, facilitates interpretation, reduces financial, time, and storage costs related to its application, and alleviates the curse of dimensionality such that predictions become better and more reliable [173]. They are usually categorized into two categories: filter methods and wrapper methods [242, 243]. Sometimes, embedded methods are distinguished as a third method [173]. Filter methods perform input selection as a kind of preprocessing step, independent of the classification algorithm. The inputs are filtered before turning to the algorithm. Wrapper methods, on the contrary, are wrapped around the algorithm. The performance of various input sets are evaluated using the algorithm itself, such that the usefulness of inputs is investigated relative to the classification algorithm. Embedded methods perform input selection and model training simultaneously.

2.2 Methods for supervised classification

The data D to be feeded to the algorithm consist of N samples with an input vector $\mathbf{x}_n = [x_1 \ ... \ x_p]^T \in \mathbb{R}^p$, $n = 1, ..., N$, and an outcome y_n indicating to which of the K outcome classes the sample belongs. we would like to make a distinction between the measurements available in the data set (the variables), and the inputs that are fed to the algorithm. The set of p inputs are the result of an input selection process using all available measurements (or variables). Thus, depending on the adopted input selection procedure, inputs may consist of an interaction between two measurements. If $K = 2$, the problem is one of binary classification. When $K > 2$, we talk about multi-class classification. The input vector can be considered a sample from the random variable $\mathbf{X}$ and the outcome a sample from the random variable Y. The data samples, then, are sampled from the bivariate distribution $p(\mathbf{X}, Y)$. We want to find a model (a classifier) that makes a prediction for Y given $\mathbf{X}$. We are mainly interested in probabilistic classifiers. Such classifiers' output yields the posterior

probability of belonging to class k given input vector $\mathbf{x}$, $P(Y = k|\mathbf{x})$, $k = 1, ..., K$, such that $\sum_{k=1}^{K} P(Y = k|\mathbf{x}) = 1$.

2.2.1 Logistic regression

Binary logistic regression

In standard binary logistic regression [193, 5], the probability of outcome class 1 given the inputs, $P(Y = 1|\mathbf{x}) = \pi_1(\mathbf{x})$, is based on a linear combination of inputs. To ensure that the output of the model can be interpreted as a probability, the model uses the linear combination of inputs to predict the logit transformation of $\pi_1(\mathbf{x})$:

$$\log\left(\frac{\pi_1(\mathbf{x})}{1-\pi_1(\mathbf{x})}\right) = \alpha + \boldsymbol{\beta}^T\mathbf{x} = z, \tag{2.1}$$

where α is the intercept, and $\boldsymbol{\beta} = [\beta_1 \ ... \ \beta_p]^T$ is a vector indicating the weight of each input in the linear combination. In this formulation, the log of the odds of class 1 are modeled. The corresponding (predicted) probability of class 1 will lie between 0 and 1, as desired. The predicted probability is obtained as

$$\pi_1(\mathbf{x}) = \frac{\exp(z)}{1+\exp(z)}. \tag{2.2}$$

The vector of model parameters $\boldsymbol{\theta}$ consists of the intercept term and the weight vector $\boldsymbol{\beta}$: $\boldsymbol{\theta} = [\alpha \ \beta_1 \ ... \ \beta_p]^T$. An estimate for $\boldsymbol{\theta}$ ($\hat{\boldsymbol{\theta}}$) is obtained by maximizing the likelihood of the data: it is the value of $\boldsymbol{\theta}$ for which the likelihood of observing the training data is maximal. Since all observations are independent and the distribution for the outcome Y is a Bernoulli distribution, the log of the likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{n=1}^{N}\Big[y_n \log \pi_1(\mathbf{x}_n) + \big(1-y_n\big)\log\big(1-\pi_1(\mathbf{x}_n)\big)\Big]. \tag{2.3}$$

The estimate $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood is found using Newton-Raphson. In this method, the likelihood function (i.e. the log-likelihood as a function of $\boldsymbol{\theta}$) around the initial estimate for $\boldsymbol{\theta}$ is approximated by a concave parabola. The value of $\boldsymbol{\theta}$ where the parabola has its maximum becomes the new estimate. This boils down to a type of weighted least squares fitting. Due to the nonconstant variance of Y, cases are weighted differently. These weights can vary over iterations. Therefore, the optimization procedure is also called iteratively re-weighted least squares: one iteratively minimizes the sum of squared weighted differences between the true outcome and its predicted probability.

The logistic regression results can be interpreted using odds ratios. The odds ratio gives the increase in the odds of class 1 for each unit increase in the input. When there are two or more inputs in the model, the odds ratio is to be interpreted conditional on the presence of the other inputs. The odds ratio is computed as $\exp(\hat{\beta}_i)$. The odds of class 1 equals $\hat{\pi}_1/\hat{\pi}_0$. The odds ratio simply is the ratio of the odds for two different situations (such as a unit increase in an input). Odds ratios lie between 0 and infinity, and a value of 1 means that the odds of class 1, and by consequence $\hat{\pi}_1$, does not depend on the input under scrutiny. When the odds ratio is smaller (larger) than 1, the odds of class 1 decreases (increases) for increasing values of the input. Even though they are widely used in the medical literature, the odds ratio has been criticized as a tool to represent logistic regression results [343, 303], mainly because they are often misunderstood as being relative risks but also because odds ratios can behave awkward in some specific situations. The relative risk is the ratio of π_1 values, which is indeed more easily understood.

Multi-class logistic regression

When the outcome has more than two classes, the binary model has to be extended. The most common method uses baseline category logit equations [5, 193]. In this model, the K^{th} class is used as the reference class. Then, $K-1$ logit equations are constructed, comparing each class k ($k \neq K$) to the reference class:

$$\begin{cases} \text{logit}\big(\pi_1(\mathbf{x})\big) = \log\left(\frac{\pi_1(\mathbf{x})}{\pi_K(\mathbf{x})}\right) & = \alpha_1 + \boldsymbol{\beta}_1^T\mathbf{x} = z_1 \\ \text{logit}\big(\pi_2(\mathbf{x})\big) = \log\left(\frac{\pi_2(\mathbf{x})}{\pi_K(\mathbf{x})}\right) & = \alpha_2 + \boldsymbol{\beta}_2^T\mathbf{x} = z_2 \\ \ldots \\ \text{logit}\big(\pi_{K-1}(\mathbf{x})\big) = \log\left(\frac{\pi_{K-1}(\mathbf{x})}{\pi_K(\mathbf{x})}\right) & = \alpha_{K-1} + \boldsymbol{\beta}_{K-1}^T\mathbf{x} = z_{K-1}, \end{cases} \tag{2.4}$$

where the model parameters $\boldsymbol{\alpha} = [\alpha_1 ... \alpha_{K-1}]^T$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 ... \boldsymbol{\beta}_{K-1}]^T$ with elements $\boldsymbol{\beta}_k = [\beta_{k,1} ... \beta_{k,p}]^T$ represent the intercept terms and the weight of each input in the respective linear combination. Using the softmax function, the predicted probability of belonging to class k is computed as

$$\pi_k(\mathbf{x}) = \frac{\exp(z_k)}{\sum_{k=1}^{K}\exp(z_k)}, \tag{2.5}$$

where z_K is equal to 0. The model parameters are estimated by maximizing the multinomial log-likelihood [193].

Important issues in logistic regression

Apart from data inspection prior to data analysis, which is important for any kind of data analysis, important issues in logistic regression analysis include

multicollinearity, model selection and outlier detection. As is usually done in the existing literature, these will be described for the binary logistic regression model, but the extension to the multi-class case is not always straightforward.

Multicollinearity This term refers to the presence of dependencies between measurements in the data set. These dependencies may involve two or more variables, therefore the inspection of intercorrelations between variables is necessary yet not sufficient. The presence of strong multicollinearity can yield unstable parameter estimates. Consider the extreme case in which two variables have a perfect linear correlation [298]. In that case, all points on the scatter plot between these variables fall on a line and one dimension is lost. If one then fits a three-dimensional hyperplane to predict an outcome, e.g. the log odds of $P(Y = 1)$ in logistic regression, infinitely many planes yield identical fit since the data points fall into a two-dimensional subsection of the three-dimensional space made up by both variables and the outcome. Thus, in case of strong multicollinearity, the standard errors of the parameter estimates can be inflated: a different sample of data can result in very different estimates.

Obviously, intercorrelations are unavoidable. Multicollinearity is said to exist if the dependencies between the variables are large. In that case, some actions to avoid unstable estimates can be taken. One option is to perform a principal component analysis to represent the available variables by means of a few uncorrelated principal components, with highly correlated variables having similar loadings on the principal components. This is usually not a good idea in medical research settings, since clinicians prefer a good model using few measurements. Another option is to drop some measurements from further analyses in order to reduce the level of multicollinearity. Even though highly correlated variables contain to a large extent the same information, it is possible that the unique (i.e. uncorrelated) part of a dropped variable is useful to predict the outcome. Despite the intercorrelations, dropping variables thus entails a risk of losing useful bits of information.

The existence of multicollinearity is typically investigated through the variance inflation factors (VIFs) [298]. The VIFs refer to the degree of inflation of the variances of the parameter estimates caused by the linear relation of that predictor to one or more other predictors. The VIF of a variable is obtained by regressing that variable on all other variables. It equals $(1 - R^2)^{-1}$, where R^2 is the proportion of variability in the variable under scrutiny that is explained by all other variables. For logistic regression models, the VIFs should be computed through the use of weighted ordinary least squares regression [10]. When a VIF is high, one may consider to drop that variable from further analysis. The question is, of course, when a VIF is too high. Often a value of 10 is seen as a threshold, even though some argue for 2.5 in the case of logistic regression [10].

Model selection With model selection, we refer to the selection of inputs, including the checks for the linearity in the logit assumption and the checks for interaction effects. There are many approaches to deal with this aspects of model building. It has been suggested to start with input selection, to continue with checking the assumption of linearity in the logit of the selected inputs, and to finish with checking for useful interactions [193]. This order is logical and useful yet not absolute. It is possible to go back to previous steps or, for example, to check for a quadratic effect of a variable that did not make it to the set of selected inputs. Checks for linearity in the logit and interactions may result in transformations of selected inputs, or in the addition of new inputs such as polynomial effects or interaction effects.

Input selection can be done using stepwise selection or automated techniques such as forward, backward, or best subset selection. In standard stepwise selection, variables are iteratively added to and removed from the input set. The best variable according to a chi-square statistic is added, if it meets the selection criterion (to be set by the user). Then, the worst variable according to a chi-square statistic is removed, if it satisfies the removal criterion. The process stops when no more variable can be added or removed. Such automated techniques have been criticized to have bad performance [13, 67, 125, 184, 365, 382]. The main arguments are that confidence intervals of estimated model parameters are too small, estimates of model parameters are biased high, noise variables tend to be selected, different selections can arise due to minor changes in the data set, severe multiple testing is performed leading to poor interpretability of p-values, and focus is given to finding the one and only best model such that model uncertainty is ignored. Moreover, it motivates analysts to not think about the underlying problem.

We would like to make a few sideline remarks on these criticisms. First, it is often argued that model selection should primarily be based on prior knowledge about the problem and the measurements [184]. Obviously, this is an important aspect of input selection. If one only relies on 'data-driven' input selection, it can happen that important measurements are not selected, for example because their influence is not statistically significant due to lack of power. It has been shown that exclusion of important but nonsignificant predictors can reduce discriminative ability between the outcome classes in new patients [365]. However, new knowledge about a particular problem needs to be generated. This requires data-driven analyses. Therefore, we prefer input selection approaches that combine prior knowledge with data-driven results. Second, too narrow confidence intervals and biased estimates resulting from not taking into account model uncertainty seems to be a result of all methods that lead to the selection of a single set of inputs [78]. Note that if the aim of the analyses is to find a predictive model (prediction) rather than the description of an underlying mechanism (description or explanation) (cf. previous chapter), ignoring model uncertainty is less crucial.

Third, the option of fitting a full model is not desirable when dealing with medical decision making problems. A good model that is useful for clinicians is one with good performance and few inputs. This increases understandability, user friendliness, and it saves resources by requiring less measurements. Moreover, it can lead to overfitting which results in poor performance of a predictive model when applied to new cases. Fourth, even though overfitting is an important problem and always needs attention, we believe that it is not a huge problem for simple linear models such as logistic regression. Overfitting in logistic regression can result from parameter estimates that are biased high, the use of noise variables, and the liberal inclusion of data-driven polynomial or interaction terms (terms adding nonlinearity to the decision boundary between both classes). Our opinion is that the most important issue is to avoid the inclusion of noise inputs (be it single variables or polynomial or interaction terms). Biased high parameter estimates are mainly a problem when training sample size is small relative to the number of measurements that are available for selection. Again, the difference between prediction or description as the main aim of the research is important. If one is interested in description, a remedy for biased high model estimates can be interesting.

An interesting point is that model specification (e.g. input selection) and model fitting to obtain parameter estimates is almost always performed on the same data. Thus, after input selection, we fit the model to the same data under the assumption that the model specification was known in advance. The truth is that we fit a model that already proved useful on the same data. Thus, no matter how data-driven input selection was done, parameter estimates and their confidence intervals will tend to be biased [78, 198]. Ideally, both tasks are performed on different sets of data. Of course, this is often difficult to translate into practice due to limited resources. Again, this problem is inherent to all data-driven input selection methods.

Other tools to assist data-driven input selection are the use of criteria based on information theory [67]. Well known criteria are the Akaike Information Criterion (AIC) [6] and the Bayesian Information Criterion (BIC) [355]. The AIC is computed as

$$\text{AIC} = -2\ell(\theta) + 2p, \tag{2.6}$$

whereas BIC is given by

$$\text{BIC} = -2\ell(\theta) + p\log(N). \tag{2.7}$$

Both criteria penalize the log-likelihood for the number of terms in the model and/or the number of cases used. The AIC penalty term seems simple and arbitrary but it has a strong theoretical foundation [67]. The BIC adds a penalty in order to find the true dimension of the true model where a true model is assumed to exist and its dimension is assumed to be low. This criterion was derived mainly with the aim

of prediction as opposed to explanation. It appears, however, that AIC tends to select too many inputs whereas BIC tends to select too few [67]. Therefore, when using these criteria one may be interested in models having low values for both AIC and BIC rather than in a model with the lowest possible value for any one of these criteria. These criteria can be used to compare non-nested models whereas likelihood ratio hypothesis tests are only valid for comparing nested models (i.e. one model is a special case of another model). Variants on these well known criteria as well as other information criteria exist, see [67].

Other methods that are interesting in the context of input selection are shrinkage methods [366]. Such methods shrink the estimated regression coefficients to avoid biased high estimates. This is mainly an issue when the input dimension is high relative to the training sample size. Shrinkage using the least absolute shrinkage and selection operator (LASSO) [382] combines shrinkage with input selection by using L_1 regularization. For linear regression, the LASSO minimizes the squared error loss function while satisfying $\sum |\beta| \leq t$, with t determining the amount of shrinkage. For logistic regression (and generalized linear models in general), the LASSO technique performs parameter optimization using regularized maximum likelihood fitting as follows [315]:

$$\hat{\theta}(\lambda) = \arg\min[-\ell(\theta) + \lambda\|\theta\|_1], \tag{2.8}$$

where $\ell(\theta)$ is the appropriate log-likelihood function and $\lambda > 0$ is the regularization parameter to be tuned. Other shrinkage methods, such as penalized maximum likelihood fitting using L_2 regularization (e.g. [139, 252, 480]) or linear shrinkage [448] do not perform input selection, yet one will force regression estimates to be small. See [407] for a comparison of some methods, and [304] for a discussion of L_1 and L_2 regularization.

In case of model building to find predictive classifiers for ovarian tumors or pregnancies of unknown location, we used a combination of tools to arrive at a final selection of inputs. Automated techniques such as stepwise or backward selection were used as exploratory tools to get information on possibly interesting measurements/variables. As far as it was available, a priori domain knowledge was also used, for example knowledge regarding the subjectivity of measurements, the difficulty of obtaining the measurements, and the likelihood of association with the outcome. It should be noted that, for ovarian tumors and PULs, many measurements were recorded because the clinicians wanted to know which measurements could be interesting for obtaining better predictions. As such, the model building phase was partly exploratory in nature. The AIC and BIC criteria were also recorded and a model with good values for both was favored. The number of inputs (cf. the principle of parsimony) also guided our choice for a particular model. Using all of these tools, several interesting input sets were created. After checks for the linearity in the logit assumption and the necessity of interaction terms (see below), the various input sets

were compared using cross-validation on the training data. The performance index of interest in the cross-validation analyses was the area under the receiver operating characteristic (ROC) curve (see below). The cross-validation results were used to obtain the final input set.

For a promising input set (i.e. before checking for interactions), the linearity in the logit assumption needs to be checked. This is not often done (or at least not often reported on) in applied medical research using logistic regression [223]. The logistic model formulation assumes that each input is linearly related to the log odds, conditional on other predictors in the model. If the assumption is violated, a transformation or the addition of a polynomial term may remedy the problem. Several methods exist for checking whether this assumption holds in case the input is continuous (see e.g. [193]). One simple approach adds the product of the input with its log-transformation to the model [57]. If the effect of this product term is strong, the input may not be linear in the logit. A more interesting and elaborate approach is to apply a semiparametric model using the theory of generalized additive models (GAM) [187]. In case of a binary outcome (classes 0 and 1), the nonparametric additive model is written as

$$\log\left(\frac{P(Y=1|\mathbf{x})}{1-P(Y=1|\mathbf{x})}\right) = \alpha + \sum_{i=1}^{p} f_i x_i, \tag{2.9}$$

where f_i is an arbitrary function that is estimated by a smooth regression function based on methods such as cubic smoothing splines or locally weighted regression (loess) [83, 187]. These methods have a smoothing parameter that regulates the amount of smoothing. This parameter can for example be estimated using generalized cross-validation [458]. For checking the linearity in the logit assumption, a semiparametric model can be used:

$$\log\left(\frac{P(Y=1|\mathbf{x})}{1-P(Y=1|\mathbf{x})}\right) = \alpha + \sum_{i=1}^{p-1} \beta_i x_i + f_p x_p, \tag{2.10}$$

such that the effects of all but one input are entered parametrically (as in standard logistic regression) and the effect of one input, for which the assumption of linearity is checked, is entered nonparametrically. The estimated smooth function for the effect of that input can be used to see whether linearity is satisfied or a whether transformation or a polynomial effect may be preferred.

Finally, checks for interactions should be done. It is often advised to look mainly at interactions that are plausible a priori. As a consequence of the curse of dimensionality, overfitting is more easily accomplished for higher-order effects. Depending on the analysis problem, it is of course possible to check for interesting interactions in a data-driven way. We would suggest to be cautious when considering

to add interaction terms that are discovered in this way. One could for example check with clinicians to listen to their opinion on the effect, and one could for example use cross-validation or similar methods to gain more information on the usefulness of the interaction effect.

Model fit and outlier detection When the model is built, we have to check its fit. This entails many issues, such as investigating whether the model describes the outcome data well (through 'goodness-of-fit' measures) and checking if there are outliers that had a big influence on parameter estimation (through 'influence diagnostics'). We will not go into detail on this issue and will mainly refer to relevant literature. For example, a very useful step-by-step approach for detecting outliers is described in [193]. Well known goodness-of-fit (GOF) indices are the Deviance and Pearson GOF, which are based on the residuals. Using chi-squared tests, it can be statistically tested whether the null hypothesis, stating that the model is correct, should be rejected or not. However, these tests only hold if the different covariate patterns (i.e. the distinct versions of $\mathbf{x}$ observed in the data) have enough instances. When continuous variables are present, there are usually many covariate patterns and the test distributions do not hold any longer. Another GOF test was developed to overcome this [255]. In this test, the training set cases are divided into 10 groups of approximately equal size based on the model's predicted $P(Y = 1|\mathbf{x})$. The cases with the smallest probabilities are put in group 1, and so on. By summing the probabilities of the cases in each group, one gets an estimate of the number of cases with $Y = 1$ the model expects in that group. A chi-squared test is used to compare the expected with the observed numbers.

The R^2 coefficient of determination in regression analysis estimates the proportion of variability in the outcome variable explained by a regression model. Alternatives for logistic regression have been derived [276, 285, 295]. Usually the McFadden R^2 index is suggested for use:

$$R^2 = 1 - \frac{\ell(\theta_M)}{\ell(\theta_0)}, \tag{2.11}$$

where $\ell(\theta_M)$ is the log-likelihood of the model, and $\ell(\theta_0)$ is the log-likelihood of the intercept only model. The McFadden R^2 has a clear interpretation as the proportional reduction in error, and thus as the proportion of explained variation if 'variation' is conceptualized in a general sense [276, 295]. It can be seen as an indication of the level of association between the set of input variables on the one hand, and the outcome on the other hand.

2.2.2 Multi-layer perceptrons

Standard multi-layer perceptrons

The fully connected feedforward multi-layer perceptron (MLP) is a standard neural network model [50]. We will focus here on MLPs with one hidden layer: the outcome Y is modeled by connecting the input vector $\mathbf{x}$ with the H units in the hidden layer which are themselves connected with the outcome units. If $K = 2$, there is one output unit, $k' = 1$. If $K > 2$, there are K output units, $k' = 1, ..., K$. More specifically, the activation z_h of a hidden neuron h ($h = 1, ..., H$) is modeled as

$$z_h = f\left(\sum_{i=1}^{p} w_{ih}x_i + b_h\right), \tag{2.12}$$

where w_{ih} is the weight of the path between input i and hidden neuron h, b_h is the bias term of hidden neuron h, and f is the transfer function that transforms the linear combination of the inputs. In the present study the tanh transfer function is used. The output unit activation $y^*_{k'}$ is modeled as

$$y^*_{k'} = g\left(\sum_{h=1}^{H} w_{hk'}z_h + b_{k'}\right), \tag{2.13}$$

where w_h is the weight of the path between hidden neuron h and the output unit, b is the bias term of the output unit, and g is the output transfer function. For binary classification problems the logistic sigmoidal function can be used. Although other link functions can be used, this is the canonical choice and also implies that we are generalizing a logistic regression model. For multi-class problems, $g(\cdot)$ will typically be the softmax function (cf. equation (2.5). See Figure 2.1 for graphical presentation of an MLP.

The vector of model parameters, usually indicated by $\mathbf{w}$, consists of all weights and bias terms. An estimate for $\mathbf{w}$ ($\hat{\mathbf{w}}$) can be obtained by minimizing the cross-entropy error function E for binary classification, or the entropy error function in case of multi-class classification; this is the negative log-likelihood of the data when using the logistic (softmax for multi-class classification) output function, which corresponds to a Bernoulli (multinomial for multi-class classification) output distribution $p(Y|\mathbf{x})$:

$$E_{bin} = -\sum_{n=1}^{N}\Big[y_n \log y^*_{k'}(\mathbf{x}_n, \hat{\mathbf{w}}) + \big(1 - y_n\big)\log\big(1 - y^*_{k'}(\mathbf{x}_n, \hat{\mathbf{w}})\big)\Big], \tag{2.14}$$

$$E_{mc} = -\sum_{n=1}^{N}\sum_{k'} I_{(y_n=k)} \log y^*_{k'}(\mathbf{x}_n, \hat{\mathbf{w}}), \tag{2.15}$$

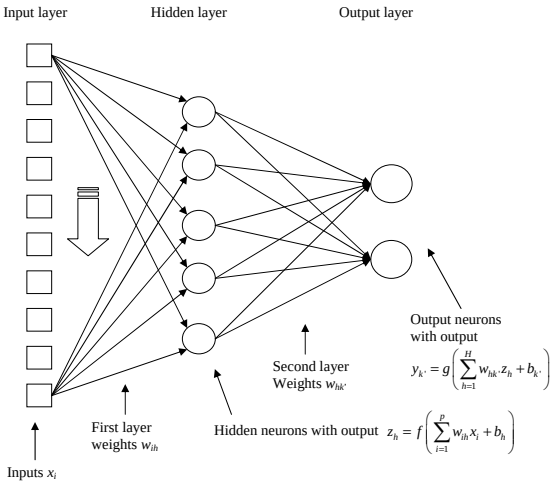


Figure 2.1: Visualization of a multi-layer perceptron.

where $y^*_{k'}(\mathbf{x}_n, \hat{\mathbf{w}})$ is the output of outcome unit $y^*_{k'}$ based on input vector $\mathbf{x}_n$ and $\hat{\mathbf{w}}$, and I is an indicator function with value 1 if $y_n = k$ and 0 otherwise. The output activation using $\hat{\mathbf{w}}$ is interpreted as the class posterior probability $P(Y = k|\mathbf{x})$. To regularize the network, the actual error function minimized is the sum of the negative log-likelihood and a regularization term intended to keep the weights small and the class boundaries smooth:

$$E_{reg} = E + \frac{\alpha}{2}\hat{\mathbf{w}}^T\hat{\mathbf{w}}, \tag{2.16}$$

where $\hat{\mathbf{w}}^T\hat{\mathbf{w}}$ is a weight decay term that avoids fitting the training data peculiarities by keeping the weights small, and α is the regularization constant that controls the amount of regularization. This constant has to be tuned, for example using cross-validation.

Many methods exist to optimize the objective function E_{reg} such that the estimated class probabilities are close to the true class outcome. Optimization algorithms for neural networks typically involve the iterative minimization of an error function. Each iteration starts with computing the derivative of the error function with respect to the weights and biases. This is done using error back-propagation, in which the derivative $\frac{\partial D}{\partial \mathbf{w}}$ is obtained by propagating the errors backward through the network [342, 50]. Then, in a second step, the derivatives are used to update the weights and biases. A variety of algorithms has been suggested for this step, the simplest of which is called gradient descent (or steepest descent) [342]. More advanced methods exist, such as Newton, Levenberg-Marquardt, quasi-Newton, and conjugate gradient [50]. In this thesis, quasi-Newton optimization using the Broyden-Fletcher-Goldfarb-Shanno formula [50]. The optimization for MLPs is

nonconvex, meaning that many local minima of E_{reg} exist in the weight space. In practice, therefore, a network is often trained using different initial values for the weights such that the weight space is sufficiently explored. Another disadvantage of MLPs is that the number of hidden neurons has to be determined.

Bayesian multi-layer perceptrons

ML solutions for parameter estimation problems such as the MLP weights and biases do not take into account uncertainty in the estimates. Bayesian analysis accounts for uncertainty by looking for a posterior probability distribution for $\mathbf{w}$ rather than a point estimate. See [48, 161] for general texts on Bayesian statistics, and [258, 259, 362, 421] for discussions of the Bayesian approach versus the traditional statistical (i.e. frequentist) approach within the healthcare domain. The posterior distribution is obtained by combining a prior distribution with information in the data D. The prior distribution reflects the prior knowledge of $\mathbf{w}$. Usually, the prior is rather vague or noninformative since we have no strong belief in what the approximate value of the model parameters should be. The data update the prior to give the, usually more informative, posterior distribution through Bayes' rule:

$$p(\mathbf{w}|D,\mathcal{H}) = \frac{p(D|\mathbf{w},\mathcal{H})p(\mathbf{w}|\mathcal{H})}{p(D|\mathcal{H})}, \tag{2.17}$$

where $\mathcal{H}$ represents the model type (including its assumptions), $p(\mathbf{w}|\mathcal{H})$ is the prior distribution of $\mathbf{w}$, $p(D|\mathbf{w},\mathcal{H})$ is the data-based likelihood of $\mathbf{w}$ under $\mathcal{H}$, and $p(D|\mathcal{H})$ is the normalization term representing the evidence of the particular model $\mathcal{H}$ (i.e., the probability density of D conditional on $\mathcal{H}$, obtained by integrating over the model parameter space). In most subsequent equations, the conditioning on $\mathcal{H}$ is omitted. Finally, model predictions are made by integrating over the posterior distribution:

$$p(Y|\mathbf{x},D) = \int p(Y|\mathbf{x},\mathbf{w})p(\mathbf{w}|D)\mathrm{d}\mathbf{w}, \tag{2.18}$$

with point predictions being the expected value of this distribution:

$$E(Y|\mathbf{x},D) = \int yp(Y|\mathbf{x},\mathbf{w})p(\mathbf{w}|D)\mathrm{d}\mathbf{w}. \tag{2.19}$$

This procedure requires the computation of complex integrals that are very often not analytically solvable (i.e. have a closed form solution). One Bayesian approach, the evidence procedure [269, 270, 381], is to approximate the posterior distribution by a Gaussian. This method generates a point estimate for some of the hyperparameters, and approximates the posterior distribution of $\mathbf{w}$ by a Gaussian that is centered on the most likely weights with a variance given by the inverse

Hessian (matrix of second order partial derivatives of the error function). This Gaussian approximation is then used to compute the required integrals for (2.18) and (2.19). Because the method optimizes the hyperparameters rather than integrating them out, it is related to maximum likelihood estimation, and is sometimes known as an ML II approach [46]. (Fully Bayesian approaches typically involve Markov chain Monte Carlo [297] or variational methods [202].)

In Bayesian methods, hyperparameters are also involved. In a Bayesian MLP, the prior distribution for $\mathbf{w}$ is Gaussian with zero mean and a hyperparameter α representing the inverse variance of the prior. While other choices of prior are possible, this is easy to interpret and convenient to analyze. This hyperparameter is related to the complexity of the network because a prior with larger variance can give rise to larger weights which tends to make the decision boundary more curved. Therefore, α is used to regularize the MLP by restricting model complexity. In fact, it is more adequate to use consistent priors [50]. Rather than using one single hyperparameter α for $\mathbf{w}$, a separate α can be used for each layer's weights and for each layer's biases. For an MLP with one hidden layer, this means that four hyperparameters are used for the prior distribution for $\mathbf{w}$.

To find the most probable model parameter values $\mathbf{w_{MP}}$, the posterior distribution is maximized which is similar to maximizing the product of the likelihood and the prior distribution (the normalization term can be ignored here), cf. equation (2.17). After taking the negative logarithm, this boils down to minimizing E_{reg} as in equation (2.16). The evidence approach to Bayesian modeling now tries to find optimal values for the hyperparameter (α_{MP}) rather than integrating over it. Therefore this is not a full Bayesian methodology. The complete formula to obtain the posterior distribution for $\mathbf{w}$,

$$p(\mathbf{w}|D) = \int \int p(\mathbf{w}|\alpha, D)p(\alpha|D)\mathrm{d}\alpha, \tag{2.20}$$

reduces to

$$p(\mathbf{w}|D) \approx p(\mathbf{w}|\alpha_{MP}, D). \tag{2.21}$$

The evidence procedure approximates this distribution using a Gaussian density around mode $\mathbf{w_{MP}}$ that is obtained by minimizing E_{reg}. What happens then in the evidence procedure is to a) initialize the weights and the hyperparameter, b) to train the network in order to minimize E_{reg} using standard optimization algorithms (such as quasi-Newton or conjugate gradient methods), c) update the hyperparameter α, d) iterate steps b and c until convergence. To make a final prediction in the sense class probabilities, we need to take the network output based on the input vector $\mathbf{x}$

and $\mathbf{w_{MP}}$ and integrate over the posterior weight distribution to take into account the uncertainty with respect to $\mathbf{w}$ (output moderation or marginalization) [269, 452].

$$P(\mathcal{H}|D) = \frac{p(D|\mathcal{H})p(\mathcal{H})}{p(D)}. \tag{2.22}$$

Since $P(D)$ does not depend on the model, and the prior distribution $P(\mathcal{H})$ is usually considered to be equal for different models, models can be compared by looking at their evidence. In general, the evidence takes into account E_{reg} with penalties for complex models and for models that have to be finely tuned. The evidence thus tries to find an equilibrium between several criteria. The higher the evidence the better the model. It is known, however, that the evidence can also be very misleading with respect to which model should be chosen and that caution is necessary [273, 319].

Some of the advantages of a Bayesian framework for MLPs are that weight decay is done automatically, no separate validation set is needed (as for example in cross-validation), it entails a method for model comparison which automatically applies Occam's razor, and it is able to apply input selection using automatic relevance determination (see below). The evidence procedure for Bayesian MLPs, however, has been criticized and mainly appears to give satisfactory results for medium to large data sets [297, 457]. Notwithstanding this, good results have been reported [293, 381]. More detailed information about the evidence procedure for Bayesian MLPs can be found in, among others, [50, 270, 292].

Input selection using automatic relevance determination

The Bayesian framework for MLPs can be readily extended for input selection purposes, using the technique of automatic relevance determination (ARD) [270, 297]. For Bayesian MLPs, ARD means that a separate hyperparameter α_i is defined for each input i's weights. When α_i is estimated to be small, the variance of the prior distribution for input i 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. In this thesis, the ARD analysis is typically repeated ten times with different initializations of the model parameters. Then, the median rank of each input over the ten ARD runs is computed. The input with lowest median rank is dropped and ten new ARD runs are done. This procedure allows us to make a final ranking of input importance by recording the order in which inputs are dropped. The ARD analysis was usually done using a hidden layer with 10 hidden neurons, to allow eventual nonlinearity express itself. Then, using cross-validation, it was determined how many of the most important inputs should be used and what number of hidden neurons would be optimal for the final BMLP model (i.e. the architecture of the network). It is sometimes argued that tuning the number of hidden neurons is not necessary given effective regularization and enough hidden neurons for the specific problem at hand [334, 199]. In practice

however, with small or medium-size data sets some variation in performance is not unusual. Moreover, a sparse network is beneficial for prospective use [180]. Good performance of ARD in the context of Bayesian MLPs has been reported [293].

2.2.3 (Least squares) support vector machines

Least squares support vector machines

Support vector machines (SVMs) are very flexible sparse models for regression and classification [449, 450, 114, 372, 1]. SVMs are developed for binary classification, even though multi-class extensions exist [462, 254]. In SVMs, the input space (the multidimensional scatter plot of the input variables) is transformed into a high-dimensional feature space using mapping $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}^q$ (See Figure 2.2). The feature space can be infinite dimensional. In the feature space, a hyperplane is sought to separate both classes by finding an equilibrium between maximization of the margin between both classes (this corresponds to regularization) and minimization of the number of misclassifications. Maximizing the margin between between the closest instances of each class corresponds to regularization by keeping the model parameters small. Least squares support vector machines (LS-SVMs) [373, 371] are a variant of SVMs that can be solved by a linear system rather than by a quadratic programming problem. Coding outcome class membership as -1 versus $+1$, the LS-SVM classifier in the primal space

$$y(\mathbf{x}) = \text{sign}\Big[\mathbf{w}^T\varphi(\mathbf{x}) + b\Big], \tag{2.23}$$

where $y(\mathbf{x})$ is the crisp class prediction, is obtained by minimizing the objective function

$$\begin{aligned} \min_{\mathbf{w},b,\mathbf{e}} \ \mathcal{J}_P(\mathbf{w},\mathbf{e}) \ &= \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma\sum_{n=1}^{N} e_n^2 \\ \text{such that} \qquad & y_n\Big[\mathbf{w}^T\varphi(\mathbf{x}_n) + b\Big] = 1 - e_n, \ n = 1, ..., N \end{aligned} \tag{2.24}$$

where the first term in equation (5.7) is responsible for margin maximization, the second term is responsible for minimization of misclassifications, $\mathbf{w}$ is the weight vector of length q, b is the bias term, e_n is the error variable, and γ is a hyperparameter to control the amount of regularization.

Rescaling of the problem ensures that the latent output of the classifier, i.e. the term between square brackets in equation (2.23), is 0 for points located on the hyperplane, and $+1$ (or -1) for points on the margin at the side of class $+1$ (or -1). Differences with SVMs are the quadratic loss function and the fact that the constraints are equality rather than inequality ($\geq$) constraints. Thus, the right hand side of the constraint in equation (5.7) contains the target value 1 on which an error

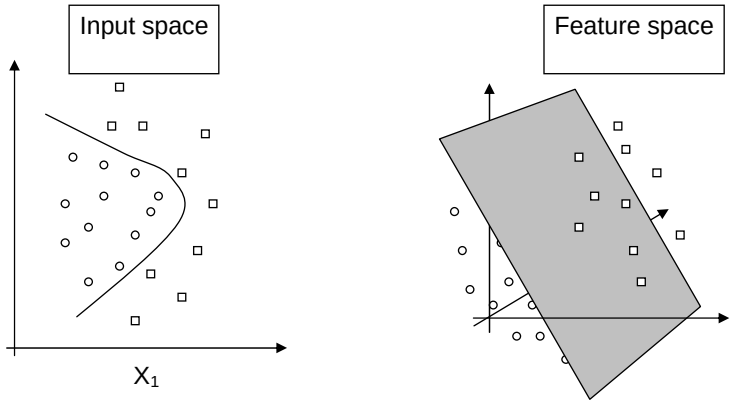


Figure 2.2: Graphical representation of the underlying rationale for support vector machines.

e_n is allowed to tolerate misclassifications. This means that the algorithm will try to locate each training case on the margin at the correct side of the hyperplane (i.e. y_n equals the latent output), while allowing deviations. The inequality constraints in SVMs mean that the 1 is seen as a threshold value, such that the algorithm tries to locate each case at least on the margin. These two adjustments in the LS-SVM formulation greatly simplify parameter optimization. For this constrained optimization problem, the Lagrangian $\mathcal{L}$ is constructed as

$$\mathcal{L}(\mathbf{w}, b, e, \alpha) = \mathcal{J}_P(\mathbf{w}, e) - \sum_{n=1}^{N} \alpha_n \Big(y_n \Big[\mathbf{w}^T \varphi(\mathbf{x}_n) + b \Big] - 1 + e_n \Big), \tag{2.25}$$

where α_n are the Lagrange multipliers ($n = 1, ..., N$). The conditions for optimality then yield

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \rightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \varphi(\mathbf{w}_n) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0 \\ \frac{\partial \mathcal{L}}{\partial e_n} = 0 \rightarrow \alpha_n = \gamma e_n, \ n = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_n} = 0 \rightarrow y_n \Big[\mathbf{w}^T \varphi(\mathbf{x}_n) + b \Big] - 1 + e_n = 0, \ n = 1, ..., N. \end{cases} \tag{2.26}$$

This results in the following linear optimization problem in the dual space, to be solved in $\boldsymbol{\alpha} = [\alpha_1 ... \alpha_N]^T$ and b:

$$\left[\begin{array}{c|c} 0 & \mathbf{y}^T \\ \hline \mathbf{y} & \Omega + I/\gamma \end{array} \right] \left[\begin{array}{c} b \\ \hline \boldsymbol{\alpha} \end{array} \right] = \left[\begin{array}{c} 0 \\ \hline 1_N \end{array} \right], \tag{2.27}$$

where $\mathbf{y} = [y_1...y_n]^T$, $1_N = [1...1]^T$, and $\Omega \in \mathbb{R}^{N\times N}$ is a matrix with elements $\Omega_{nm} = y_n y_m \varphi(\mathbf{x}_n)^T \varphi(\mathbf{x}_m)$ with $n, m = 1, ..., N$.

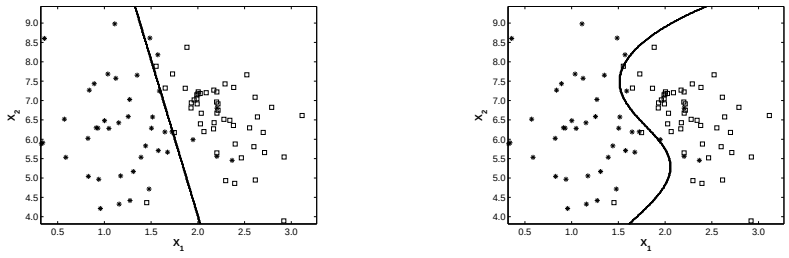
Given an input vector $\mathbf{x}$, the classifier in the dual space takes the form

$$y(\mathbf{x}) = \text{sign}\left[\sum_{n=1}^{N} \alpha_n y_n K(\mathbf{x}, \mathbf{x}_n) + b\right], \tag{2.28}$$

where $K(\cdot, \cdot)$ is a kernel function. This reformulation allows us to work in the high-dimensional feature space without explicitly constructing it by using a positive definite kernel

$$K(\mathbf{x}, \mathbf{z}) = \varphi(\mathbf{x})^T \varphi(\mathbf{z}). \tag{2.29}$$

The feature space is determined by the specific kernel function used. The linear kernel, $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T\mathbf{z}$, results in a linear separation between the two classes in the original input space (Figure 2.3(a)). On the other hand, nonlinear kernels such as the radial basis function (RBF) kernel, $K(\mathbf{x}, \mathbf{z}) = \exp\big(-\|\mathbf{x} - \mathbf{z}\|^2\big)/\sigma^2$, create a nonlinear decision boundary in the input space (Figure 2.3(b)).



(a) **(b)**

Figure 2.3: Decision boundary for an SVM with a (a) linear or (b) nonlinear kernel.

Finally, some issues with respect to SVMs and LS-SVMs are shortly mentioned. First, optimization for SVMs and LS-SVMs is convex, such that a unique solution exists. This is a major advantage of these methods over MLP models. Second, when the SVM formulation is used, many of the α values will turn out to be zero. This makes SVMs sparse: the training cases for which the support value α_n is zero are not used in the classifier. The cases with a positive support value (in SVMs, $\alpha_n \geq 0$), are called support vectors are these are usually cases that are located close

to the decision boundary between both classes. For LS-SVMs, due to the 2-norm in the cost function, typically no support value will be zero and all training cases are support vectors. The third condition for optimality in equation (2.26) shows that support values can be negative in LS-SVM models. Training cases with large $|\alpha_n|$ are located close to or far from the decision boundary. Third, the formulation of the model in the primal weight space is a parametric problem with $\mathbf{w}$ being a weight vector with fixed length q. In the dual weight space, the problem is nonparametric as the support value vector $\boldsymbol{\alpha}$ has length N which depends on the size of the training data. Therefore, for some data sets it can be more convenient to solve the model in the primal weight space while for other data sets the dual weight space can be preferred. Fourth, (LS-)SVM models use a regularization parameter γ, and possibly a kernel parameter such as the kernel width σ when using the RBF kernel. Tuning of these parameters is often done using cross-validation methods. Fifth, the classifiers in equations (2.23) and (2.28) result in a crisp class prediction by taking a sign function on the latent output. This may not always be the best way to move from continuous output to a black-and-white prediction. Another threshold value instead of zero may be preferred, for example when one class is more common than the other. The same result can be obtained by adjusting the bias term b.

LS-SVM models for classification and function estimation have proven their reliability in various domains, such as ovarian cancer prediction [267], microarray data analysis on ovarian cancer [119], endometrial carcinoma analysis [118], brain tumour recognition [126], mental development of preterm newborns [21], marketing [455], electric load forecasting [147], financial time series [439], and soccer [423].

Bayesian LS-SVMs

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 [441]. Other advantages of a Bayesian approach are that hyperparameters (regularization parameter, kernel parameter) are automatically tuned and that it automatically entails regularization (cf. the description of Bayesian MLPs and references therein). The Bayesian approach to LS-SVMs in [441] also uses MacKay's evidence procedure [270]: the posterior distribution is approximated by a Gaussian around the mode that represents the most probable values $\mathbf{w_{MP}}$. Hyperparameters are also optimized rather than integrated out.

For the Bayesian LS-SVM, we work with a slightly modified cost function:

$$\min_{\mathbf{w},b,\mathbf{e}} \mathcal{J}_P(\mathbf{w},\mathbf{e}) = \frac{\mu}{2}\mathbf{w}^T\mathbf{w} + \frac{\zeta}{2}\sum_{n=1}^{N} e_n^2 \tag{2.30}$$

such that $\gamma = \zeta/\mu$. We can look at the Bayesian approach as a hierarchical method with three levels [270, 371]. On the lowest level, $\mathbf{w}$ and b of the LS-SVM classifier are of interest. Their prior distribution is set to be multivariate normal [441]:

$$p(\mathbf{w}, b|\log(\mu), \log(\zeta), \mathcal{H}) = p(\mathbf{w}|\log(\mu), \mathcal{H}) p(b|\log(\sigma_b), \mathcal{H}), \tag{2.31}$$

where the prior for $\mathbf{w}$ is Gaussian with zero mean and variance equal to μ^{-1}, and the prior for b is Gaussian with zero mean and variance equal to σ_b^2 with $\sigma_b \to \infty$ to approximate a uniform distribution. Applying Bayes' theorem results in the most probable values for these parameters, $\mathbf{w_{MP}}$ and b_{MP}. Since the prior distribution relates to the regularization term in (2.30), and the likelihood function relates to the sum of squared errors, obtaining the most probable values boils down to solving an LS-SVM model. On the second level, one is occupied with the hyperparameters ζ and μ. The most probable values are obtained using a uniform prior on $\log(\zeta)$ and $\log(\mu)$ [441]. The third level deals with model selection. When the RBF kernel is used, this level involves the update of σ. The prior $p(\mathcal{H}_j)$ over all possible models is taken to be uniform. The procedure for the Bayesian LS-SVM is roughly as follows: a) initialize μ, ζ, and σ if the RBF kernel is used, b) solve the linear optimization problem in equation (2.27) to obtain $\mathbf{w_{MP}}$ and b_{MP}, c) estimate ζ_{MP} and μ_{MP}, d) update σ, e) go to step b if necessary. The final output of a Bayesian LS-SVM is a class probability. This is obtained by integrating over the posterior distribution for $\mathbf{w}$ 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. More detailed information about Bayesian LS-SVMs can be found in [371, 441].

Input selection using ARD

For Bayesian LS-SVMs, ARD input selection is implemented by inserting a diagonal weight matrix $U = \text{diag}[u_1, \ ..., \ u_p] = \text{diag}(u)$ in the kernel function [442]. When using the RBF kernel for ARD, the kernel becomes

$$K(\mathbf{x}, \mathbf{z}) = \exp[-(\mathbf{x} - \mathbf{z})^T U (\mathbf{x} - \mathbf{z}) \sigma^{-1}] = \exp[-(\mathbf{x} - \mathbf{z})^T \overline{U} (\mathbf{x} - \mathbf{z})], \tag{2.32}$$

where $\overline{U} = \text{diag}(\overline{u}) = U/\sigma$. The weights $\overline{u}$ are then optimized in the Bayesian model. The initial choice for $\overline{u}$ is taken to be $[1, \ ..., \ 1]/\sigma$, with σ being the optimal RBF kernel parameter for the unweighted model. The input with the smallest weight can for example be dropped and ARD can be run again. This procedure results in a final ranking of the inputs based on when they are dropped. Cross-validation can then be used to determine how many of the most important units are to be chosen (and can also bring about changes in the ranking).

Input selection using rank-one updates

Recently, fast methods for forward and backward input selection using LS-SVMs were developed [310]. The LS-SVM model structure allows fast computation of model performance measures based on leave-one-out cross-validation (LOO-CV) [76, 75]. LOO-CV is a method that yields nearly unbiased results (with large variance,

however) [248]. It can be computationally intensive since a model is trained N times. Each time, one case is omitted from the training set to serve as a validation case. The N validation performance results are then averaged to obtain the LOO-CV performance estimate of a model. The validation performance v_n for the n^{th} case can be written as

$$v_n = C\big[f^{(-n)}(\mathbf{x}_n), y_n\big], \tag{2.33}$$

where $f^{(-n)}$ is the prediction of the classifier when trained on all cases except case n, and C is a function that weights the cost of the prediction. The prediction can be a crisp prediction using a sign function on the latent output of the LS-SVM classifier, or can be the latent output itself.

Also, adding (forward selection) or removing (backward selection) an input from the input set requires the re-computation of the LS-SVM model, which can become very cumbersome when many inputs are to be evaluated. However, it is shown in [310] that the LS-SVM model can be updated using simple rank-one adjustments in the kernel matrix. This method is very fast for performing greedy input selection (i.e. forward or backward), but is only available for linear kernel LS-SVMs.

2.2.4 Relevance vector machines

Relevance vector machines (RVMs) are another way of obtaining probabilistic output using an SVM-like problem formulation [400, 401, 52, 406]. However, there are fundamental differences between both methods. First of all, RVMs are by default approached from the Bayesian framework of MacKay [270]. RVMs use the following model, similar to equation (2.28):

$$y(\mathbf{x}) = \sum_{n=1}^{N} w_n K(\mathbf{x}, \mathbf{x}_n) + w_0, \tag{2.34}$$

where w_n is the weight for training sample n (this resembles the support value in SVMs), w_0 is a bias term, and $K(\cdot,\cdot)$ is a kernel function. Another fundamental difference with SVMs is that RVMs do not explicitly use a feature space in relation to a positive definite kernel such that the basis function $K(\cdot,\cdot)$ in the model can be any function.

The prior for w_n is taken to be Gaussian with zero mean and variance equal to α_n^{-1}, such that the prior distribution for the model parameter vector $\mathbf{w} = [w_1 ... w_N]^T$ equals

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{n=1}^{N} \mathcal{N}(w_n|0, \alpha_n^{-1}), \tag{2.35}$$

with $\boldsymbol{\alpha} = [\alpha_1 \ ... \ \alpha_N]^T$. The values α_n, $n = 1, ..., N$, are related to the strength of the prior for w_n: these hyperparameters are inversely related to the influence of sample n to the decision boundary. The higher α_n the smaller the variance of the prior for w_n and thus extreme values for w_n are considered unlikely. As prior for α_n, a Gamma(0,0) distribution is used, such that the prior distribution for $\boldsymbol{\alpha}$ becomes

$$p(\boldsymbol{\alpha}) = \prod_{n=1}^{N} \text{Gamma}(\alpha_n|0, 0). \tag{2.36}$$

This prior distribution induces sparsity in the model since model fitting will result in many α_n values approaching infinity. In such cases, the prior for w_n will be a Gaussian with mean and variance equal to zero such that the n^{th} training case has zero weight and is not used in the model formulation in equation (2.34). The other training cases are the relevance vectors. These are similar to the support vectors in SVMs, except that the cases used in RVMs are usually not close to the decision boundary (cf. SVMs) but are prototypical examples of the classes. RVMs typically result in fewer support vectors. Model fitting to find the most probable weights $\mathbf{w_{MP}}$ is done by minimizing a negative regularized log-likelihood function similar to that mentioned for binary MLPs or to that of a L_2-regularized logistic regression model. A major disadvantage of RVMs is the nonconvex optimization with many local minima solutions. SVMs do not have this drawback.

2.2.5 Kernel logistic regression

Kernel logistic regression (KLR) [338, 221, 481] is a method that is closely related to SVMs since in nature they only differ with respect to their loss function [481]. Whereas SVMs are based on the hinge loss (the loss is zero for cases lying at least on the margin at the correct side of the decision boundary and increases linearly otherwise), KLR uses a negative log-likelihood type of loss (cf. cross-entropy). As noted in [481], KLR has the advantages to directly result in a probability estimate and to be readily extended to multi-class KLR (MKLR). However, due to the change in loss function, no support vector will be zero and sparseness is lost (as in standard LS-SVMs). Recently, it has been shown that MKLR can be solved in an LS-SVM framework [220]. This method starts from a regularized version of MLR which is solved by optimizing the L_2-penalized negative log-likelihood function using iteratively regularized re-weighted least squares (IRRLS). 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 using an iteratively re-weighted LS-SVM method (irLS-SVM). We used the RBF kernel for MKLR to automatically obtain a nonlinear separation in the input space.

The LS-SVM based version of MKLR starts from an L_2-regularized version of MLR which is solved by optimizing the penalized negative log-likelihood (PNLL)

function using iteratively regularized re-weighted least squares (IRRLS). A penalty is added to the negative log-likelihood function for MLR in order to keep the parameters small:

$$\ell(\theta) = \text{NLL} + \frac{\gamma}{2}\sum_{k=1}^{K-1}(\beta_k^*)^T\beta_k^*, \tag{2.37}$$

where $\beta_k^* = [\alpha_k;\ \beta_k]^T$ contains the intercept. To account for the intercept, we will define $\mathbf{x}_n^* = [1;\ \mathbf{x}_p]^T$. Let $B = [\beta_1^*;\ ...;\ \beta_{K-1}^*]^T \in \mathbb{R}^{(K-1)(p+1)}$ be the vector containing all model parameters. Parameter optimization is performed iteratively by updating the estimated B:

$$B^{(r)} = B^{(r-1)} + s^{(r)}, \tag{2.38}$$

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 method approximates the log-likelihood function by a parabola based on $B^{(r-1)}$ and finds it maximum to yield $B^{(r)}$. 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 $\varphi(\mathbf{x}_n^*)$. To represent the input space data, define $A^T \in \mathbb{R}^{(k-1)(p+1)\times(k-1)N}$ as

$$\left[\begin{array}{ccc|ccc|c|ccc} \mathbf{x}_1^* & 0 & 0 & \mathbf{x}_2^* & 0 & 0 & & \mathbf{x}_N^* & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 & \cdots & 0 & \ddots & 0 \\ 0 & 0 & \mathbf{x}_1^* & 0 & 0 & \mathbf{x}_2^* & & 0 & 0 & \mathbf{x}_N^* \end{array}\right]. \tag{2.39}$$

To represent the feature space, define $\Phi^T \in \mathbb{R}^{(k-1)q\times(k-1)N}$ as

$$\left[\begin{array}{ccc|ccc|c|ccc} \varphi(\mathbf{x}_1^*) & 0 & 0 & \varphi(\mathbf{x}_2^*) & 0 & 0 & & \varphi(\mathbf{x}_N^*) & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 & \cdots & 0 & \ddots & 0 \\ 0 & 0 & \varphi(\mathbf{x}_1^*) & 0 & 0 & \varphi(\mathbf{x}_2^*) & & 0 & 0 & \varphi(\mathbf{x}_N^*) \end{array}\right]. \tag{2.40}$$

We also define the following vectors:

$$r_n = [I_{(y_n=1)},\ ...,\ I_{(y_n=K-1)}]^T, \tag{2.41}$$

$$\mathbf{R} = [r_1^T,\ ...,\ r_N^T]^T, \tag{2.42}$$

$$\mathbf{P}^{(r)} = [p_{1,1}^{(r)},\ ...,\ p_{K-1,1}^{(r)},\ ...,\ p_{1,N}^{(r)},\ ...\ p_{K-1,N}^{(r)}] \in \mathbb{R}^{(K-1)N}, \tag{2.43}$$

and the n^{th} block ($\in \mathbb{R}^{(K-1)\times(K-1)}$) of the blockdiagonal weight matrix $\mathbf{W}^{(r)} = \text{blockdiag}(\mathbf{W}_1^{(r)},\ ...,\ \mathbf{W}_N^{(r)})$ as

$$\begin{bmatrix} t_n^{1,1} & t_n^{1,2} & \cdots & t_n^{1,K-1} \\ t_n^{2,1} & t_n^{2,2} & & t_n^{2,K-1} \\ \vdots & & \ddots & \\ t_n^{K-1,1} & t_n^{K-1,2} & & t_n^{K-1,K-1} \end{bmatrix}, \tag{2.44}$$

with

$$t_n^{k,l} = \begin{cases} p_{k,n}^{(r)}(1 - p_{k,n}^{(r)}), & \text{if } k = l \\ -p_{k,n}^{(r)} p_{l,n}^{(r)}, & \text{if } k \neq l \end{cases} \tag{2.45}$$

By using a positive definite kernel, IRRLS can be applied to the feature space such that one ends up with an iteratively re-weighted LS-SVM (irLS-SVM) algorithm used to obtain $\pi_k(\mathbf{x}^*)$ as

$$\pi_k(\mathbf{x}^*) = \frac{\exp\left[(\beta_k^*)^T \varphi(\mathbf{x}^*)\right]}{\sum_{l=1}^{K} \exp\left[(\beta_l^*)^T \varphi(\mathbf{x}^*)\right]}, \tag{2.46}$$

where $(\beta_K^*)^T \varphi(\mathbf{x}^*) = 0$ since the K^{th} class is the reference class. In irLS-SVM, one iteratively minimizes the cost function

$$\begin{aligned} \min_{s^{(r)}, \mathbf{e}^{(r)}} \ & \frac{1}{2}(\mathbf{e}^{(r)})^T \mathbf{W}^{(r)} \mathbf{e}^{(r)} + \frac{\gamma}{2}(s^{(r)} + \mathbf{B}^{(r-1)})^T (s^{(r)} + \mathbf{B}^{(r-1)}) \\ & \text{such that } (\mathbf{W}^{(r)})^{-1}(\mathbf{R} - \mathbf{P}^{(r)}) = \Phi s^{(r)} + \mathbf{e}^{(r)}. \end{aligned} \tag{2.47}$$

2.2.6 Software

In this thesis, logistic regression models are developed using SAS version 9.1 (SAS Institute, Cary, USA), unless stated otherwise. MLP models are developed using the NETLAB toolbox for Matlab [292], LS-SVM models using the LS-SVMLAB toolbox for Matlab [318, 371], and KLR models using Matlab code from the developer [220]. Matlab version 7.1 is used (The Mathworks Inc., Natick, USA).

2.3 Evaluating model performance

The ultimate goal of model development is to achieve a model that does, as good as possible, what it is supposed to do. Therefore, model performance needs to be evaluated rigorously. Model evaluation is a crucial element of classifier development and needs more scrutiny than it often gets [332, 3, 15, 179]. For example, the accuracy (or, inversely, the misclassification error) represents the proportion of cases that are correctly classified. This is a widely used yet suboptimal measure of performance [60, 332, 3]. If classes are unbalanced, this measure is misleading. Suppose one of two classes accounts for 90% of the cases. Then, if we always predict that class by default, our accuracy will be 0.90 without any modeling. Classifiers that are able to predict the smaller class quite well too, may end up with an accuracy below 0.90 even though they are better models. The problem with the accuracy is that misclassification costs are assumed equal: predicting class 0 for a class 1 case and the reverse are seen as equally 'bad'. This is rarely the case in medical research. An overview of possible problems in the practice of performance evaluation is presented in [3].

As already indicated in the previous chapter when discussing CDS systems, diagnostic models for medical decision making (or other domains) need prospective evaluation of their performance. Moreover, they are preferably also compared with the performance of human experts, and their added value for clinical practice should be investigated. Also, it is useful to test a model in many different settings because this allows the assessment of the range of applicability of the model. One may argue that a model that was developed using data from Belgian clinical centers cannot be applied in other continents. Yet, one can try to quantify possible performance differences between centers whose data where used for model development and other centers. Performance differences between different types of centers can also be quantified. In case of ovarian tumor diagnosis, one can for example compare university hospitals to tertiary referral centers.

2.3.1 Performance measures

The basics: sensitivity and specificity

Basic quantities in model evaluation are the sensitivity and specificity. When any kind of predictive model arrives at crisp predictions, these quantities can be computed. A binary prediction model is very simple as it directly gives a crisp class prediction (e.g. ascites could be used to predict malignancy of an ovarian tumor with the presence of this condition leading to the prediction of malignancy while its absence would predict benignity). The sensitivity is computed as the proportion of cases from the class of interest (usually this is said to be class 1) that are correctly predicted as class 1 members: it is the number of true positives (TP) divided by the sum of the number of true positives and false negatives (FN). True positives are

correctly predicted class 1 cases, false negatives are class 1 cases that are erroneously predicted to belong to the other class (class 2). The specificity, then, is computed as the proportion of cases from class 2 (often also called class 0 or class -1) that are correctly predicted: it is the number of true negatives (TN) divided by the sum of the number of true negatives and false positives (FP). Prediction models leading to continuous predictions such as the probability of belonging to class 1 can be summarized by sensitivity and specificity by defining a threshold value on the continuous prediction. For example, a straightforward yet often suboptimal threshold on probabilities can be 0.50. Cases for which the predicted probability for class 1 is at least 0.50, are predicted to belong to class 1.

Using TP, TN, FP, and FN, other simple performance measures can be derived. The positive predictive value (PPV), for example, is equal to TP/(TP+FP). This represents the probability of belonging to class 1 when this class is predicted. The negative predictive value (NPV) is the analogue for class 2: TN/(TN+FN). Other very interesting summarizing quantities are the positive and negative likelihood ratios (LR+ and LR−) [205, 206]. The LR+ is computed as sensitivity/(1 − specificity). It represents the increase in the pre-test odds of class 1 when class 1 is predicted. The LR− is computed as (1 − sensitivity)/specificity and represents the decrease in the pre-test odds of class 2 when class 2 is the predicted class. Finally, the misclassification rate described above equals (FP+FN)/(TP+TN+FP+FN).

Misclassification rate, the ROC curve, and other measures

Many measures for evaluation performance have been suggested in the literature for continuous predictions. The misclassification rate is already described, and is not considered a good measure in most situations. In general, dichotomization of continuous measurements is often seen as suboptimal practice [84]. Also, many possible threshold values exist, and it is often not clear in advance what the best one would be. Of course, one can choose a threshold that performs well with respect to sensitivity and specificity, which is a common practice, but three remarks are in place. First, at least in medical applications, the best threshold value may depend on the clinical center or the patient population. Second, the threshold value is often derived using the training data, but then again overfitting may be an issue. Ideally, a new set of data should be used. Third, the desired balance between sensitivity and specificity levels may differ between clinicians.

Many measures exist that try to quantify the level of discrimination between both classes achieved by the continuous predictions, without dichotomization. The most popular is the area under the receiver operating characteristic (ROC) curve (AUC) [182], which is one of several measures that can be derived from ROC curves [375, 60, 320, 250]. An ROC curve is constructed by plotting sensitivity (y-axis) versus 1 minus specificity (x-axis) by varying the threshold over its entire range. First, the threshold is lower than the lowest continuous prediction such that all cases are predicted to be of class 1. This results in maximum sensitivity and 1 minus

specificity (both are 1, upper right corner of the plot). At the other extreme, all cases are predicted to be of class 2, yielding zero sensitivity and 1 minus specificity (lower left corner of the plot. If the predictions discriminate well between both classes, meaning that class 2 cases usually have a lower value than class 1 cases, there will be thresholds with high sensitivity and low 1 minus specificity (towards upper left corner of the plot). Thus, one hopes for a model that lifts up the curve as much as possible towards the upper left part. ROC curves can be constructed for any continuous value applied to a set of data belonging to either of two classes: it can be probabilistic or non-probabilistic continuous output from a classification algorithm, or the value of any kind of clinical measurement.

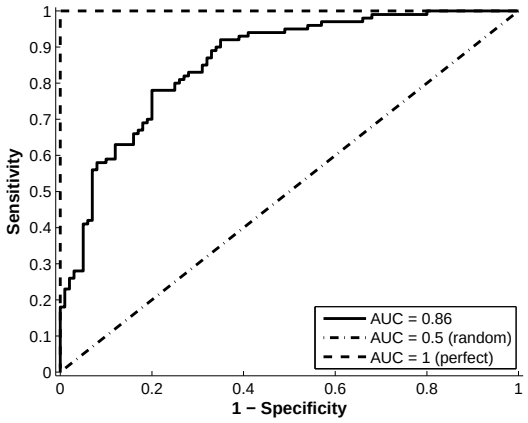


Figure 2.4: Examples of ROC curves.

The AUC computes the area under the ROC curve. For a model that gives random prediction values, the true underlying AUC is 0.5: it connects the $(0, 0)$ with the $(1, 1)$ coordinates explained above. A perfect test results in an AUC of 1: it goes from $(0, 0)$ over $(0, 1)$ to $(1, 1)$. See Figure 2.4 for exemplary ROC curves. It is often interpreted as the probability that a random (class 1, class 2)-pair of cases will be correctly discriminated by the model: the predictive value v for the class 1 case is larger than the predictive value v for the class 2 case. There are various approaches to construct the ROC curve and its accompanying AUC based on a specific data set, see [250] for a review. There are empirical nonparametric methods [182], nonparametric methods dealing with curve smoothing to avoid the bumpy empirical curve [482], parametric methods [278], and bootstrap-based techniques [307]. The empirical nonparametric method is frequently used. It approximates the ROC curve by connecting the (sensitivity,1-specificity) results by a step function and computing the AUC using the trapezoidal rule. This nonparametric estimate of the AUC is similar to the Mann-Whitney U-statistic divided by the product of the number of class 1 and class 2 cases (N_1 and N_2, respectively). If we define

$$U_{n_1,n_2} = \begin{cases} 1 & \text{if } v_{n_1} > v_{n_2} \\ 0.5 & \text{if } v_{n_1} = v_{n_2} \\ 0 & \text{if } v_{n_1} < v_{n_2} \end{cases}, \tag{2.48}$$

the AUC can be written as

$$\text{AUC} = \frac{U}{N_1 N_2} = \frac{1}{N_1 N_2} \sum_{n_1}^{N_1} \sum_{n_2}^{N_2} U_{n_1,n_2}. \tag{2.49}$$

This formula reflects the interpretation of the AUC mentioned previously.

Several approaches to compute confidence intervals on the AUC have been proposed. The method to derive the variance of the AUC as described in [120] is constructed within an empirical nonparametric framework. The obtained estimate of the standard deviation can then be used to compute a 95% confidence interval. This produces symmetric confidence intervals which may go beyond the upper limit of 1. Therefore, bootstrap-based confidence intervals can be constructed, for example using the bias-corrected bootstrap method [307]. The statistical comparison of AUCs of classifiers applied to the same data has been described by [183] and improved by [120].

Other measures based on the ROC curve can be derived. The most well-known is the partial AUC (pAUC). Typically, one is not interested in the rightmost part of the ROC curve. In this part, 1 minus specificity is high such that many (possibly most) cases from class 2 are predicted to be of class 1. Usually, one is interested in models that achieve both good sensitivity and specificity. The pAUC computes the area under that part of the curve where 1 minus specificity is within a desired range. For example, if we want a test that has a specificity level of at least 0.80, we look at that part of the ROC curve where 1 minus specificity is at most 0.20 and compute the area underneath this part of the curve. Again, both parametric [274] and nonparametric methods exist [479]. Confidence intervals can again be computed using the bias-corrected bootstrap method. When the desired range for 1 minus specificity is zero, one is interested in the sensitivity level at a fixed level of specificity [321]. This is another possible measure, which we call SensSp where Sp denotes the percentage specificity one is interested in, for example Sens75.

Next to measures dealing with the discriminatory ability of a model, one can also look at measures that quantify how well the predictions approach the truth. In case of probabilistic classifiers, suitable performance measures are the cross-entropy for binary classification problems and the entropy for multi-class classification problems. These error measures can be used to compare models to see which model gives class probabilities that best approach the truth. The Brier score is another error measure

[65].

Other curves than the ROC curve exist for the evaluation of classifiers: the cost distribution [2], the cost curve [136], and the decision curve [456]. These can be seen as variants or extensions of the ROC curve. To summarize, decision curve analysis is based on the net benefit of the model. The net benefit is the difference of the proportion TPs and the proportion FPs where these terms are weighted by the relative harm of FN and FP results. The perceived harm depends on the threshold probability one uses to assume class 1 membership (with class 1 typically indicating the presence of a condition). The decision curve plots the net benefit of the model by the threshold probability. This allows one to see if the net benefit is positive for one's own threshold probability. If so, it means that the model does more good than harm and can be used.

Multi-class extensions of ROC analysis

Extensions of the ROC curve and the AUC for multi-class problems have been proposed [290, 135, 181]. Nonparametric extensions of binary ROC curve analysis to three-class ROC analysis are discussed in [290, 135]. The M-index proposed by [181] is a more ad hoc measure that can be applied to any multi-class problem with more than two classes. More details are given in Chapter 4.

2.3.2 Evaluating classifiers or algorithms

The important distinction between a classifier and an algorithm is stressed in [129]. A classifier is a specific model formulation that uses an input vector $\mathbf{x}$ to make class predictions such as $P(Y = k|\mathbf{x})$. An algorithm is a tool that, given N input vectors $\mathbf{x}_n$ and corresponding class information (in case of supervised learning), produces a classifier. In medical research, one is interested in having a specific formula to predict the probability that a patient has a disease of interest. Thus, they need a classifier that they can apply to any patient of interest. However, for example in machine learning, one is interested in finding the best algorithm. One is not interested in a particular classifier but in the algorithm that is able to produce the best classifiers. If one is interested in classifiers, an often applied procedure is to divide available data into a training data set and a test data set (data splitting). The former is used to develop the classifiers, the latter used only for evaluating the classifiers. Possibly, a validation data set is used as well, for example to tune parameters that are needed to construct the classifier. Cross-validation and bootstrapping using the training data are alternative methods to tune parameters for which no separate validation set is needed [367, 368].

When the interest lies in the algorithm, typically multiple training data sets are constructed in order to construct many classifiers. Each time, the remaining part of the available data serves as test set. The multiple test set results can then be combined, for example by averaging over them, in order to evaluate the

algorithm. Common strategies used to construct the training data sets are multiple independent splits of the data into a training and a test part (often referred to as repeated data splitting, repeated hold-out, repeated subsampling, or resampling) [129], cross-validation [367, 368, 451], and bootstrapping [140, 142].

In the statistical literature, the methods for evaluating a classifier or an algorithm (data splitting, repeated data splitting, cross-validation, bootstrapping) outlined in the previous paragraph, are considered ways of performing an internal validation of a model [184, 213]. Internal validation concerns the evaluation of a particular model on the data set that was used for developing the classifier.

Two remarks should be made on the use of the abovementioned evaluation strategies. First, if one uses a single data set to repeatedly train and test a classifier (in order to evaluate an algorithm), there are difficulties with the computation of the variability of the average test set performance [129, 294, 45]. For repeated data splitting, the different training and test sets overlap, whereas for cross-validation the test sets are independent but the training sets are not. Second, various strategies have different bias and variance characteristics. For cross-validation, this depends on the number of folds. For bootstrapping and repeated data splitting, this depends on the number of samples or splits. LOO-CV (cf. supra), for example, has very low bias but can have very high variance [140, 207, 241]. Bootstrap methods such as the 0.632 or 0.632+ methods [140, 142] are better than cross-validation with respect to variance but are sometimes reported to suffer from high bias [241, 288, 172]. For cross-validation, stratification with respect to outcome can improve both bias and variance, and repeating the analysis can also reduce variance [241]. Repeated data splitting using 75% of the data for training and 25% for testing turned out to be a well-performing method when compared to repeated 5-fold cross-validation and the 0.632+ bootstrap [172]. To close on this point, one can say that, to evaluate an algorithm, a method with both low bias and variance is preferred. When comparing several algorithms, however, one may be willing to sacrifice some bias for low variance [241].

2.3.3 Model comparison

Often, researchers want to compare classifiers or algorithms. In applied research, this is an important step given the knowledge that no algorithm will be the best for all data or for all classification problems (cf. no free lunch theorems [467]. Traditionally, such comparisons are often done using statistical significance testing [347, 129, 122]. When applying two classifiers on the same test set, statistical tests exist to compare the resulting AUC values [183, 120]. Also, there is the McNemar test [359] to compare misclassification rates, sensitivities, or specificities. When comparing algorithms, methods for comparison based on repeated data splitting or cross-validation have been developed [12, 294, 54, 171]. Statistical tests for the comparison of more than two models have also been suggested [323]. In

[122], statistical tests are described for comparing two or more algorithms over multiple data sets. Wilcoxon's signed ranks test and the Friedman test [359] are two nonparametric statistical tests advocated in [122].

Given the difficulties and flaws associated with statistical hypothesis testing, we prefer to compare methods using confidence intervals and effect sizes. In a statistical test, the null hypothesis of no effect (e.g. algorithm A and B have equal performance) is contrasted with the alternative hypothesis (algorithm A and B perform differently). If the test provides sufficient evidence against the null hypothesis, this hypothesis is rejected in favor of the alternative hypothesis. Otherwise, the null hypothesis is not rejected (which does not mean that it is accepted). The evidence used to decide whether the null hypothesis should be rejected is usually given by a p-value. Rejection of the null hypothesis usually happens when $p < 0.05$. This number is labeled the Type-I error, referring to the probability of finding a significant difference when the null hypothesis is actually true. This practice of null hypothesis significance testing (NHST) is widely used in the applied sciences yet it has been extensively criticized by statistical experts [47, 74, 346, 85, 86, 14, 167, 364, 403], most often in the medical, psychological, and economical literature. NHST has fundamental flaws such that blind reliance on p-values for the interpretation of one's results can be dangerous [200]. As an alternative to p-values, the use of confidence intervals and effect sizes (with confidence intervals on them) has been proposed. Effect size measures try to quantify the effect under study and is more interesting than a p-value, which in fact is a statement about one's collected data rather than one's hypotheses that are tested using the data. Effect sizes can be for example differences in AUC values with bootstrap-based confidence intervals, and difference in proportions (e.g. for misclassification rate, sensitivity, specificity) with confidence information.

Finally, when many classifiers or algorithms should be compared, statistical testing can become cumbersome and yield unwieldy results. One can perform an overall test to check for any difference between the models, but usually also pairwise checks comparing each pair of models are desired in order to locate specific differences. Then, the problem of multiple comparisons arises: if we perform each test at the same Type-I error, the overall Type-I error becomes much larger such that the probability of a false positive result (a significant test when there is no true underlying difference) may grow rapidly. Methods to adjust for multiple comparisons exist, but these have drawbacks: many methods are suboptimal, the correction often depends on the number of tests performed, and they are trying to keep the Type-I error low at the expense of power (i.e. the probability of finding a significant difference when the null hypothesis is indeed false) [350, 339, 322, 151]. Power is often considered more important: what good is a statistical test with a good Type-I error but with a low probability of finding a true underlying effect? As an alternative to the hassle of the $\frac{1}{2}M(M-1)$ pairwise comparisons when comparing M models, one can use ranking methods [62, 321].

Pepe and colleagues [321], in the context of analyzing genes in microarray research, suggest to rank the genes with respect to the performance measure of choice and to provide a bootstrap based index of variability of the ranking. This method can also be used to rank classifiers. They suggest to compute $Pm(\kappa)$, which refers to the probability of a model to be ranked among the best κ classifiers. This is easily done using bootstrapping. The bootstraps can also be used, for example, to construct confidence intervals on the classifiers' rank. Brazdil and Soares [62] developed ranking methods in case algorithms are compared on several data sets. Their methods can be adjusted for single data set comparisons. They suggest two particular ranking methods: average ranks (AR) and success rate ratios (SRR). The AR method is straightforward and, in the case of a single data set, can be applied using bootstrapping. The SSR method is applied by computing the performance ratio (PR) of each pair of algorithms/classifiers and on each of the D data sets under scrutiny. The D performance ratios for each pair of algorithms/classifiers are averaged, and an overall mean PR (OMPR) is computed for each algorithm/classifier by averaging all mean PRs in which a specific algorithm/classifier is involved.

2.3.4 Calibration

Having a probabilistic classification model with good discriminatory power and/or low error rates is very important, but does not tell the whole story. The class probabilities must be accurate. The rationale behind probabilistic models is that they give uncertainty information rather than crisp (black-and-white) class predictions. This uncertainty information is important in many medical decision making situations. Obviously, this assumes that the uncertainty information is reliable. This can be investigated with calibration plots. One way of constructing such plots starts with grouping the cases in a data set according to intervals with respect to the predicted probability of class k (e.g. $[0 - 0.1]$, $]0.1 - 0.2]$, and so on). Next, the observed proportion of class k cases and the 95% confidence interval is plotted per group. An example can be found in [264]. This method can result in a calibration test in the spirit of the Hosmer-Lemeshow goodness-of-fit test in logistic regression [134, 378]. In this thesis, calibration plots are constructed using nonparametric regression based on GAM theory [187] to estimate the true relationship between the predicted probability of class k and the true probability to belong to this class. Ideally, this relationship should follow a straight line where predicted probabilities equal the true probabilities.

2.3.5 Prospective validation

Prospective validation of mathematical models concerns checking model performance on data that were collected after model development [213]. We will define prospective internal validation as a prospective validation on data from centers whose data were also used for model development. Prospective external validation

is defined as a prospective validation on data from new centers. Such external validation is different from independent validation as defined in [213]. Independent validation refers to a validation performed by independent investigators that is usually taking place at a new center. Prospective validation as performed in later chapter of this thesis was always coordinated by an investigator who was also involved in model development. For example, on of the aims of the IOTA study group is to develop and test mathematical models based on standardized data collection. Differences in data collection strategies is likely to cause performance degradation when testing models in new centers. A disadvantage of standardized data collection is that emerging mathematical models cannot be applied without a short training course in which the data collection principles are explained. This, however, is a small effort that is worth the likely gain in model performance.

2.4 Conclusion

This chapter described algorithms for supervised classification. Of course, other algorithms exist that are not described in this text. At a given point, choices need to be made with respect to the algorithms one will use. In this text, the basic method is logistic regression. The advanced algorithms are based on neural networks and support vector machines. For medical decision making problems, we consider it crucial that the algorithms are able to yield probabilistic output.

It is clear that evaluation of classifiers is important, and that many methods for this purpose have been advocated in the literature. For us, the main evaluation methods involve ROC curves and calibration curves.

Chapter 3

Ovarian tumor and PUL diagnosis: background, aims, and data

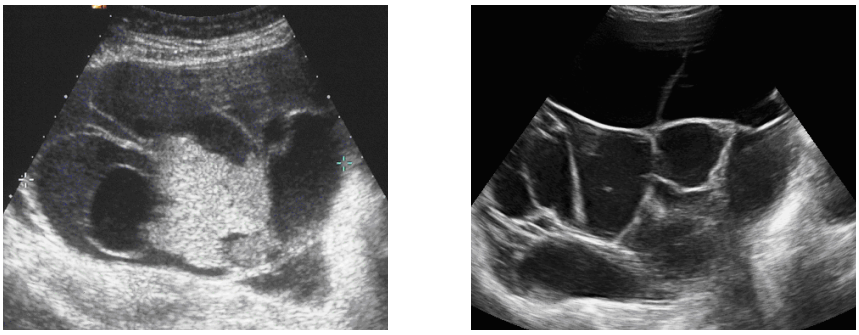
Abstract – This chapter has two parts. The first part presents background information and relevant previous research regarding ovarian tumors and pregnancies of unknown location. The second part describes the data that were available for continuing the research described in the first part.

3.1 Previous research

3.1.1 Ovarian tumors

Background

The term tumor refers to an abnormal growth of tissue that can be either benign or malignant. The term cancer is reserved for malignant tumors. Figure 3.1 presents ultrasound images of a benign and a malignant ovarian tumor. Ovarian tumors are located in the ovaries, the female reproductive glands on both sides of the uterus in which the egg cells (ova) are produced. Ovaries consist of three different tissue types: epithelial, stromal, and germ cells. The epithelial cells form the surface of the ovary. Most ovarian tumors originate in the epithelium and about 90% of the ovarian cancers are epithelial. Stromal cells produce the female reproductive hormones estrogen and progesterone. Germ cells are located on the inside of the ovary and they produce the ova which are transported to the fallopian tubes. The gonadal stroma and the germ cells rarely are the origin of ovarian tumors. In our study, we use another classification of ovarian tumors. We distinguish between four main categories: benign, primary invasive, borderline, and metastatic invasive. The latter three categories are ovarian malignancies. Primary invasive cancers have their first origin in the ovary, while metastatic ovarian cancer represents a cancer that has started elsewhere (e.g., breast, colon, stomach, and pancreas) but has spread to the ovary. Borderline ovarian cancers are primary invasive cancers of 'low malignant potential', representing less aggressive cancers that are less life-threatening. Ovarian cancers can be staged according to the criteria recommended by the International



(a) **(b)**

Figure 3.1: Two examples of ovarian tumors as seen on ultrasound examination: (a) a malignant tumor representing a multilocular cyst with solid components, and (b) a benign tumor representing a multilocular cyst.

Federation of Gynaecology and Obstetrics (FIGO) [189]. Four stages are discerned: in stage I the tumor is still confined to the ovaries, in stage II the tumor involves other organs in the pelvis, in stage III the cancer has spread either to the lining of the abdomen or to the lymph nodes, and in stage IV the cancer has spread to organs outside the peritoneal cavity (distant metastasis). With respect to our four categories of ovarian tumors, the staging can be defined for primary invasive and borderline malignant tumors.

The American Cancer Society [19, 210] estimates that 22,430 new cases of ovarian cancer will have emerged in the United States in 2007 (ranking eighth of all cancers). An estimated 15,280 women will have died of ovarian cancer during 2007 (ranking fifth). In 2004, 14,716 deaths due to ovarian cancer were reported (ranking fifth). The lifetime risk of developing cancer in the United States is about 1 in 67 (1.5%); the lifetime risk of developing ovarian cancer and dying from it is about 1 in 95 (1.1%). In England, 5,408 new cases of cancers of the ovary and other unspecified female genital organs (5,293 ovarian cancers alone) were registered in 2004 [309]. In Belgium, there were 889 new cases of ovarian cancer in 2003, ranking this cancer fifth [44]. The cumulative risk of developing ovarian cancer between 0 and 74 years was estimated to be 1.1%. For information on incidence and mortality rates per 100,000 people at risk, we can mention the data from the Globocan 2002 study [152] since identical age-adjustments for the crude rates are used for different countries. The Globocan 2002 study provides estimates of age-adjusted incidence (AI) and mortality (AM) rates per 100,000 people at risk for various cancers and countries in 2002. Age-adjustment is necessary when comparing regions possibly having different age distributions. The world standard population structure is used to compute age-adjusted rates. For the USA, the UK, and Belgium, AI for cancer of the ovary and other unspecified female genital organs was, respectively, 10.6, 13.4 and 12.0. The

AM rates were 6.1, 8.0, and 7.4. Ovarian cancer is a lethal cancer, and it is the most lethal cancer of the female reproductive system.

Optimal treatment of ovarian tumors obviously depends on the type of tumor. Benign tumors need a different approach than malignant tumors. Vergote and colleagues [454], using a multi-center database of more than 1500 patients with a primary invasive epithelial cancer of Stage I, discovered that tumor rupture during surgery was an independent factor influencing disease-free survival after laparoscopic removal of the tumor. Therefore, laparoscopic removal should be restricted to benign masses only. This finding illustrates the need for early preoperative discrimination between benign and malignant tumors.

The American Cancer Society [19] gives the following list of risk factors for ovarian cancer: age, the use of estrogen alone as postmenopausal hormone therapy, body weight, personal or family history of ovarian or breast cancer, and inherited mutations in BRCA1 and BRCA2 genes [280, 468]. Factors that reduce the risk are pregnancy, long-term use of oral contraceptives, tubal ligation (female sterilization through sealing the fallopian tubes) and hysterectomy (surgical removal of uterus). Recently, Baker and colleagues [30] observed that the use of black tea and decaffeinated (but not regular) coffee may reduce the risk. A widely used marker for ovarian cancer is CA-125 (cancer antigen 125), a glycoprotein produced by the antibody OC 125. Raised levels of CA-125 are seen as a cause of concern, even though it is also raised in some benign conditions such as endometriosis while it may be low for some types of ovarian cancer [404]. Nevertheless, it is often used as a tool for diagnosis.

General aims of the research

The aims of our research involve early preoperative discrimination between different ovarian tumor types. Four main aims are distinguished: prospective evaluation of old classifiers, development of new classifiers, prospective evaluation of all classifiers, comparison with subjective assessment of the tumor by an expert ultrasound operator, and an investigation of the role of CA-125 in ovarian tumor diagnosis. Several models to classify ovarian tumors have already been proposed in the literature. These classifiers suffer from a few important drawbacks. They were usually derived from small single center data sets without standardized data collection. This probably reduces performance when the classifiers are applied to new data from other centers. Also, typically a limited set of measurements were considered as possible predictors of malignancy. Finally, a good large and sound prospective evaluation of the classifiers using a large and multi-center data set is lacking. Such a data set should also be used to develop new, mathematically simple or advanced, classifiers. These aims are pursued under the wings of the IOTA study group. This consortium focuses on the collection of large multi-centric data sets based on standardized data collection (see section 3.2.1).

Existing mathematical models for ovarian tumors

In this section, we shortly give an overview of methods to classify ovarian tumors that have been proposed in the literature. We start off with a morphologic classification system [170]. They categorized ovarian tumors based on their morphology into five groups: unilocular, unilocular-solid, multilocular, multilocular-solid, and solid tumors. Unilocular or multilocular refers to whether the cyst cavity is divided into several parts separated by one or more septa (strands of tissue) or not. When the cyst cavity is almost completely filled with solid parts, the tumor is categorized as solid, when there are solid parts that do not fill up most of the cavity, the tumor is categorized as unilocular-solid or multilocular-solid. The definition of each type of tumor was slightly modified in [384]. This system isolates many benign tumors by checking whether the tumor has some solid parts or not. It does not reliably isolate malignant tumors, however.

A whole range of simple morphologic scoring systems have been proposed [169, 203, 348, 124, 256, 399, 153, 405, 9]. The most well-known and widely used scoring system is the risk of malignancy index (RMI) [203]. The RMI is calculated as

$$U * M * \text{CA125}, \tag{3.1}$$

where U is the ultrasound score (0, 1, or 3), M is 1 for premenopausal patients and 3 for postmenopausal patients, and CA-125 is expressed in U/ml. The ultrasound score was 0 if none of the following features were present: multilocular cyst, presence of solid areas, evidence of metastases, presence of ascites, and bilateral lesions. The score was 1 when one of these features was present, and 3 when more than one feature was present. For the RMI, a threshold of 200 is often used to indicate malignancy. The RMI was found to have good performance and is the standard in clinical practice. An Italian prospective multi-center study [153] compared five such scoring systems. The obtained AUCs were between 0.75 and 0.84. The RMI obtained an AUC of 0.88 on a set of 191 patients, and of 0.87 on a set of 160 patients from the University Hospitals Leuven (Belgium) [386, 267].

The main statistical technique for predicting a dichotomous outcome, logistic regression, has also been applied multiple times for differentiating benign from malignant ovarian tumors [8, 212, 267, 282, 329, 376, 386, 396]. Three such logistic regression models were prospectively compared on a group of 100 women from King's College London [26]. AUCs varied from 0.70 to 0.93. Another prospective comparison of two models on 79 women from the university hospitals in Trondheim and Malmö yielded AUCs of 0.84 and 0.87 [408]. More advanced and complex techniques such as artificial neural networks and support vector machines have also been used. Several MLPs have been trained [396, 49, 377, 82], and also LS-SVMs and RVMs have been used [267, 268]. Disadvantages of such methods are that one cannot describe in a paper how the model can be applied by others. For MLPs this is still feasible [396] provided the network is not too large. But in general, the model

can only be applied using appropriate software that can handle the electronically stored model parameters.

Mol and colleagues [287] set up a prospective evaluation of a large number of methods, albeit on a small data set collected at only one center (Velthoven, the Netherlands). The obtained AUCs were at most 0.90. Hitherto, ovarian tumor model development and prospective validation suffers from several drawbacks. The models are often developed on small to medium sized data sets (typically less than 200 and rarely higher than 400) that were collected at only one center. Usually, only a limited number of measurements were taken for each tumor and these measurements suffered from a lack of standardization [384]. In this sense, model development and prospective evaluations thereof can still be improved.

It is an important issue to investigate how mathematical models relate to the subjective impression of the ultrasound operator. Timmerman and colleagues [391] investigated the accuracy of the subjective assessment of ovarian tumors by having 300 patients assessed by six operators with different levels of experience. Results indicated that experience matters, with the two most experienced operators performing best. These operators had higher sensitivity and similar to higher specificity levels than the four less experienced operators. On two thirds of all masses all six operators were correct. On 90% of the masses at least 4 out of 6 operators were correct. These results suggested that subjective impression of ovarian masses is good to very good but that experience matters. In that sense, mathematical models may be a valuable add-on for clinicians, definitely if they are less experienced. In any case mathematical models alone do not perform better than an operator's subjective impression alone. Comparing subjective impression with mathematical models is not straightforward since subjective impression leads to a black and white prediction such that only sensitivity and specificity can be computed. Mathematical models result in numerical output (e.g. probability estimates) for which an ROC curve can be drawn. To compute sensitivity and specificity, a threshold on the model's output needs to be determined but a good threshold is hard to define.

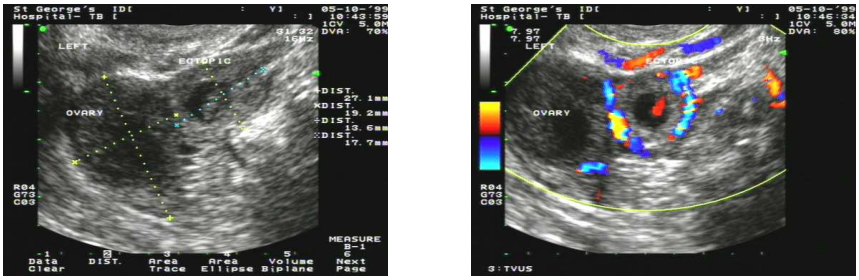
A prospective evaluation of various models on 257 patients from the University Hospitals K.U.Leuven indicated that the RMI performed similar to the logistic regression models or artificial neural networks, and that the expert's subjective assessment was superior to mathematical models.

3.1.2 Pregnancies of unknown location

Background

An ectopic pregnancy is a pregnancy occurring outside the endometrial cavity. Over 95% of such pregnancies are tubal (mostly ampullary, but also isthmic, fimbrial, or sometimes interstitial), the remaining are of ovarian, abdominal, or cervical origin [56]. See Figure 3.2 for an example of a tubal ectopic pregnancy. Over the last

decades, the incidence of ectopic pregnancy has increased [79, 164, 77, 218] with some reports of stagnation or decrease during the past few years based on research in the USA, Australia, UK, France, and Scandinavia [112, 77, 211, 218, 55, 201, 478, 31, 33, 257]. In the UK, over 10000 ectopic pregnancies were diagnosed each year during the 1991-2002 period [257]. Between 1991 and 1993, the incidence of ectopic pregnancy was 9.6 per 1000 pregnancies. Between 1994 and 2002, the incidence was stable at approximately 11 to 11.5 per 1000 pregnancies [257]. The recent decrease in could possibly be explained by better (e.g. non-invasive) treatment of primary ectopic pregnancy, improved female fertility [32, 31], and decreasing prevalence of chlamydial infections [143] and pelvic inflammatory disease [218]. The previously increasing trend has no clear explanation [313] yet could possibly be due in part to increased (earlier) diagnosis of ectopic pregnancy [23, 311], but also to the increase in ectopic pregnancy risk factor prevalence [272].



(a) **(b)**

Figure 3.2: An example of a tubal ectopic pregnancy as seen on ultrasound examination: (a) the pregnancy is located in the Fallopian tube next to the ovary, and (b) cardiac activity is observed.

Until recent decades, concern about a possible ectopic pregnancy only rose in pregnant women with severe abdominal pain, irregular vaginal blood loss and even collapse during the first pregnancy trimester. With the introduction of serum chorionic gonadotrophin (hCG) measurements, transvaginal (as opposed to transabdominal) ultrasonography (TVS), and also the uptake of less invasive surgical techniques (e.g. laparoscopy), the prevalence of ruptured ectopic pregnancies has decreased [313, 70, 249, 262, 231] and mortality due to rupture has declined [77, 277, 231]. In the US, the case-fatality rate fell 90% from 1970 to 1989 [164]. Currently, more than 90% of ectopic pregnancies can be seen on TVS [165, 398, 69, 166, 402, 103, 110], such that many early detected ectopic pregnancies can be safely treated conservatively [369, 313, 261, 227, 232, 356]. The advent of serum hCG measurements, TVS and non-invasive surgical techniques resulted in earlier ultrasound evaluation of pregnancies and in better diagnostic performance of experts, issues that have been associated with rupture and ectopic pregnancy

mortality [24, 306]. All of this may also reduce the high cost associated with ectopic pregnancies [113, 460].

Yet, in the UK, ectopic pregnancy is the fourth leading cause of maternal deaths that are directly related to pregnancy [257]. Between 1985 and 1999 in Michigan (USA), 6% of directly and indirectly (as an aggravation of an existing condition) pregnancy-related deaths were due to ectopic pregnancy [22]. Khan and colleagues [224] found that around 5% of direct and indirect deaths were due to ectopic pregnancy in developed countries. Moreover, ectopic pregnancy is the single most important cause of maternal pregnancy-related death during the first trimester of pregnancy [164]. For example, between 2000 and 2002, 11 out of 15 first trimester deaths in the UK were due to ectopic pregnancy [257]. Between 1985 and 2002, a result of 71 out of 104 (68%) was obtained [257]. Because the pregnancy is located outside the uterus, there is still a serious risk of rupture [211] and consequent hemorrhage that can be life threatening. Risk factors for rupture appear to be previous ectopic pregnancy and high hCG levels [251]. An early and accurate diagnosis of ectopic pregnancy is therefore very important. Early detection also affects effective non-invasive treatment of ectopic pregnancies.

The earlier ultrasound evaluation of pregnancies, however, also increased the number of women with a pregnancy that could not be visualized. These pregnancies of unknown location (PULs) can be defined as a situation in which there is a positive pregnancy test with no signs of intra- or extra-uterine pregnancy on TVS, and remnants of miscarriage are absent [102]. Eight to 31% of women presenting at an Early Pregnancy Unit fulfil these criteria [71, 174, 35]. Such pregnancies can evolve in four ways. The majority represent spontaneous resolution of pregnancy (failing PUL), which could be either ectopic or intra-uterine, but is never seen on TVS. Another large group of PULs are early intra-uterine pregnancies (IUPs). A third, smaller group of PULs represent ectopic pregnancies. The fourth possibility is very rare, and these can be described as persisting PULs. There is some evidence that such PULs behave biochemically like ectopic pregnancies and almost certainly represent missed ectopic pregnancies [104]. Condous and colleagues [104] found 54% failing PULs, 36% IUPs, 8% ectopic PULs, and 2% persisting PULs in a sample of 388 pregnancies. Thus, the majority of PULs do not represent ectopic pregnancies.

General aims of the research

An expectant, wait-and-see approach has been demonstrated to be safe for women with a PUL [71, 174, 35]. This approach can be a serious psychological and financial burden, however, and the dangerous and unpredictable nature of ectopic pregnancies further emphasizes the need for early prediction of this subgroup of PULs such that the well-being of this group of women can be guaranteed. The prediction of the outcome in a PUL population is a great challenge in early pregnancy today. Research into this, with a special focus on the prediction of ectopic pregnancy, is the aim of the present work on PULs. This is a multi-class classification problem.

Previous research

Historically, threshold predictions about the presence or absence of an ectopic pregnancy were investigated for pregnant women presenting with lower abdominal pain or vaginal bleeding [214, 215, 69, 37, 40, 38, 39, 219, 245]. Now, more women present at earlier gestations, with no symptoms or signs indicative of ectopic pregnancy, and often with a PUL. Prediction of ectopic pregnancy in case of asymptomatic PULs is therefore becoming increasingly important. Old predictive criteria are not necessarily sufficient for the new group of patients and have to be revised.

The classic predictive rules focused on the hCG discriminatory zone [215] and on the hCG evolution [214]. In pregnant women with abdominal pain or vaginal bleeding, the absence of an intra-uterine on ultrasound sac would predict ectopic pregnancy. However, this rule can only work if the hCG has reached the level - the discriminatory zone - on which an eventual intra-uterine sac should be visible on ultrasound [215]. The discriminatory zone was put at 1000 IU/L [69] and 1500 IU/L [37, 40]. Such procedures ignore the elevated risk of ectopic pregnancy in symptomatic women with lower hCG levels. In symptomatic women, lower hCG levels increase the risk of ectopic pregnancy [245]. Therefore, other strategies take TVS as the most important diagnostic test [117, 168, 245].

When hCG levels are low and an intra-uterine pregnancy cannot be reliably visualized on TVS, information on hCG evolution can help. A serum hCG level increase of at least 66% over 48 hours would indicate a normal intra-uterine pregnancy while otherwise an ectopic pregnancy should be considered [214]. This threshold value was later adjusted to 53% [39]. To distinguish between spontaneous abortion and ectopic pregnancy, it is suggested that a hCG decrease of less than 21% after two days or 60% after one week should raise suspicion for ectopic pregnancy [38].

Serum progesterone appears to be able to distinguish between abnormal (ectopic and abnormal IUP pregnancies) and normal pregnancies in patients with hCG levels below 1000 IU/L [115]. A meta-analysis indicated that, irrespective of hCG levels, serum progesterone is useful to identify patients at risk but that it is not specific enough to diagnose ectopic pregnancy [286], because it cannot discriminate between ectopic pregnancy and spontaneous abortion [275]. Finally, it is also described that, for women with an empty uterus and hCG levels up to 1000 IU/L, the thickness of the endometrium may have predictive value. Women with a thin endometrium (less than 7mm) were more likely to have an ectopic pregnancy [116].

Applying these ectopic pregnancy criteria to PULs resulted in low values for traditional criteria such as sensitivity, specificity, PPV, and NPV [105, 416, 99]. This demonstrates the need to revise the criteria for PULs. Based on PUL populations, Banerjee and colleagues discovered that a serum progesterone level below 20 nmol/L

was predictive of failing PUL [34]. Another study also reported on the usefulness of serum progesterone to distinguish between the different PUL outcomes [105], but also pointed at the hCG ratio as an important index. This is the ratio of the 48-hour hCG level divided by the initial hCG level. Note that Kadar and Barnhart and their colleagues [214, 38] already looked at 48-hour hCG evolutions. The hCG ratio appeared highly fruitful for distinguishing between all three PUL outcomes. It was observed that the hCG ratio outperforms individual hCG levels for the prediction of failing PULs, with a ratio below 0.87 suggesting failing PUL [92]. A prospective analysis suggested that the hCG ratio is superior to the serum progesterone level for detecting failing PULs, both with respect to AUC and thresholds suggested in the literature [107]. An hCG ratio above 1.66 increases the risk of an IUP [109]. Thus, a ratio between 0.87 and 1.66 raises concern for an ectopic pregnancy. It can be concluded that the hCG ratio threshold is superior to the progesterone threshold for diagnosing PULs as it is better able to distinguish between an ectopic pregnancy and an IUP.

In the end, the clinician remains responsible for the management of PULs, with mathematical models only serving as an additional aid. Therefore, it is important to check the accuracy of, and inter-observer variability in, the subjective assessment of a PUL. Therefore, one study had several experts with varying experience look at the biochemical profile of women with a PUL [101]. The findings tell us that experience does not lead to more accurate predictions. Also, subjective assessment may help in judging the viability of a PUL (whether it is a failing PUL or an immediately viable IUP), but is not good for judging the location of the pregnancy (i.e. intra- or extra-uterine). This can be interpreted as a cry for help in the prediction of ectopic pregnancy in the PUL population, for example using mathematical modeling.

Logistic regression analysis has been applied to predict failing PULs. For example, Banerjee and colleagues [35] developed a model using vaginal bleeding (yes/no), endometrial thickness (mm), serum progesterone level (nmol/L), and serum hCG level (IU/L). In a prospective test set of 104 women form the same hospital, this model obtained an AUC of 0.93 [34]. When using 420 women with a PUL from St. George's Hospital, an AUC of 0.91 was obtained [107], while hCG ratio by itself had an AUC of 0.98. The major challenge, however, is the prediction of all three outcomes simultaneously, with specific attention to ectopic pregnancy. Condous and colleagues [104] tried to predict all three outcomes simultaneously for the first time, using multi-class logistic regression. Their final model used only the hCG ratio, achieving test set AUC values above 0.95 for predicting failing PULs or IUPs, and 0.89 for predicting ectopic pregnancies. Possibly other measurements than those considered in this study can improve the prediction.

3.2 Available data

3.2.1 Ovarian tumors: the IOTA group

As explained above, existing work was often based on small data sets that were recruited at few centers without standardization of data collection. With respect to this latter issue, it turned out that several ultrasound-based indices were interpreted differently by different clinicians [384]. This lead to the foundation of the International Ovarian Tumor Analysis (IOTA) group. The activities of the IOTA group are coordinated by professor Dirk Timmerman (University Hospitals Katholieke Universiteit Leuven, Belgium). Members of the IOTA steering committee are Dirk Timmerman, Lil Valentin (Malmö University Hospital, Lund University, Sweden), Tom Bourne (St. George's Hospital Medical School, University of London, UK), William P. Collins (King's College London, UK), Sabine Van Huffel (Katholieke Universiteit Leuven, Belgium), and Ignace Vergote (University Hospitals Katholieke Universiteit Leuven, Belgium). The IOTA steering committee discussed the issues of standardization of interpretation and formulation of procedures to derive tumor features. Their consensus opinion was made public in [395].

The first goal of the IOTA group was to collect a data set of at least 1000 patients with a persistent ovarian mass at multiple centers that all used a unified protocol for data collection (phase I of the IOTA study). This data set was to be used for building mathematical models for ovarian tumors prior to surgery, to evaluate existing models, and to evaluate subjective assessment of expert operators. phase II of the study had the aim of prospectively testing the built models on new data from centers that contributed to the phase I data and from other centers. After closure of phase I data collection but before the start of phase II, three main centers from phase I had continued to collect data on patients with ovarian masses. This effort was labeled phase Ib, and the data served as a first prospective internal validation of built models.

IOTA phase I data set

The IOTA phase I multi-center observational study comprised nine European centers: University Hospitals Katholieke Universiteit Leuven Leuven (Belgium), Malmö University Hospital at Lund University (Sweden), Università Cattolica del Sacro Cuore in Rome (Italy), Dipartimento di Scienze Cliniche 'Luigi Sacco' at Università degli Studi di Milano (Italy), Ospedale San Gerardo at Università degli Studi di Milano in Monza (Italy), Napoli (Italy), Università degli Studi di Napoli (Italy), Hôpital Boucicaut in Paris (France), Centre Medical des Pyramides in Maurepas (France), and King's College Hospital in London (UK). Patients presenting with at least one persistent ovarian mass who where investigated by a principal investigator at one of the nine centers were eligible for inclusion. If more than one mass was found, data from the most complex mass were included. Patients were excluded from the study if they were pregnant, refused transvaginal

sonography, underwent surgery more than 120 days after sonography, disagreement in the final histology, or incomplete submission of data. The unit of analysis was the patient rather than the tumor as the patient is the central element of clinical practice. This issue was also not uniform in previous attempts to classify ovarian tumors.

Data collection took place between 1999 and 2002, and was done using a secure web-based system [4]. Each patient automatically received a unique identification number. The patient's name was not recorded to ensure her privacy. Data security was obtained by encrypting all data communication between data entry and the central database with a 56-bit secure socket layer (SSL) certificate. To ensure data integrity, JavaScript checks, XML-encoded rule checks, and manual checks by three experts were performed. Data collected involved demographic information, personal and family history of breast and ovarian cancer, the presence of pain during examination, and around 40 morphologic and blood flow measurements to describe each tumor. More details on the measurements are given in the Appendix. The investigator also provided a subjective impression of the tumor, once by indicating on a five-point scale how certain (s)he was of the nature of the tumor (certainly benign - probably benign - uncertain - probably malignant - certainly malignant) and once by giving a black-or-white prediction (benign - malignant). Except for CA-125 levels, the measurement of which was encouraged but not obligatory, and structural missings (i.e. missing values because nothing could be observed), the IOTA phase I data set is complete. Examples of structural missings are the diameters of the largest solid tumor component when no solid parts exist or color Doppler blood flow indices when no intratumoral blood flow velocity waveforms were detected. When structural missings involved physical measurements of magnitude or amount, they were imputed with zero. Structural missings involving the color Doppler indices pulsatility index (PI), resistance index (RI), peak-systolic velocity (PSV), and time-averaged maximum velocity (TAMXV) were imputed with 2, 1, 3 cm/sec, and 1 cm/sec, respectively.

The data set consists of 1066 patients with a mean age of 47 years. Eight hundred of these are benign (75%) and 266 are malignant (25%). Of the malignant tumors, 169 were primary invasive, 55 were borderline, and 42 were metastatic. The number of women contributed by each center and the distribution of outcome per center is shown in Table 3.1. For model building, this data set is split into a training part and a test part. The training part consists of 754 patients (71%), the test part of the remaining 312 patients. This split is stratified for center and outcome. Descriptive statistics are presented in Tables 3.2 and 3.3. The descriptive statistics are computed without imputation of structural missings.

Multicollinearity analysis

When building a logistic regression model, it is interesting to investigate beforehand in what way the variables are interrelated. When two or more variables are

Table 3.1: The distribution of the outcome in the nine centers participating in IOTA phase I.

	Number of	Outcome			
Center (code)	Patients	Benign	Primary invasive	Borderline	Metastatic
Malmö (MSW)	315, 29.6%	247, 78.4%	40	17	11
Leuven (LBE)	263, 24.7%	170, 64.6%	62	14	17
Roma (RIT)	126, 11.8%	81, 64.3%	23	12	10
Milano (MIT)	87, 8.2%	79, 90.8%	6	1	1
Paris (BFR)	80, 7.5%	71, 88.8%	7	2	0
Maurepas (MFR)	64, 6.0%	57, 89.1%	6	1	0
London (KUK)	54, 5.1%	38, 70.4%	10	4	2
Monza (OIT)	46, 4.3%	29, 63.0%	12	4	1
Napoli (NIT)	31, 2.9%	28, 90.3%	3	0	0

Table 3.2: Descriptive statistics for continuous/ordinal variables (IOTA phase I data).

	Benign	Malignant		Prim. Inv.	Borderline	Metastatic
Variable	N (Me)	N (Me)	AUC*	N (Me)	N (Me)	N (Me)
age	800 (42)	266 (56)	0.708	169 (58)	55 (52)	42 (57.5)
postmenoy	229 (10)	145 (12)	0.571	102 (11.5)	17 (12)	25 (18)
parity	800 (1)	266 (2)	0.615	169 (2)	55 (2)	42 (1)
lesmax	800 (63)	266 (100.5)	0.693	169 (101)	55 (108)	42 (82)
lesvol	800 (73.4)	266 (302.8)	0.689	169 (301.4)	55 (374.6)	42 (163.4)
fluid	140 (12)	144 (24)	0.714	107 (25)	15 (14)	22 (16.5)
septum	343 (2.1)	143 (4)	0.730	91 (4.8)	36 (3.9)	16 (4.4)
papheight	156 (7)	121 (14)	0.776	77 (16)	35 (13)	9 (14)
papmax	156 (10)	121 (21)	0.771	77 (22)	35 (18)	9 (16)
papvol	156 (0.19)	121 (2)	0.777	77 (2.5)	35 (1.7)	9 (1.9)
ratiopaples	156 (0.003)	121 (0.006)	0.623	77 (0.006)	35 (0.009)	9 (0.006)
papnr	156 (1)	121 (4)	0.743	77 (4)	35 (4)	9 (4)
nrlocules	800 (1)	266 (3)	0.601	169 (3)	55 (4)	42 (2)
solidmax	309 (21)	244 (50)	0.791	160 (59.5)	43 (25)	41 (58)
solidvol	309 (1.6)	244 (34.4)	0.788	160 (47.3)	43 (3.5)	41 (56.9)
ratiosolles	309 (0.03)	244 (0.24)	0.700	160 (0.26)	43 (0.02)	41 (1)
colscore	800 (2)	266 (3)	0.782	169 (3)	55 (3)	42 (3)
PI	506 (0.95)	246 (0.74)	0.296	157 (0.7)	48 (0.74)	41 (0.76)
RI	506 (0.59)	246 (0.50)	0.307	157 (0.5)	48 (0.48)	41 (0.53)
PSV	506 (11.4)	246 (24.3)	0.753	157 (26)	48 (19.8)	41 (23.3)
TAMX	498 (6.9)	241 (17)	0.785	153 (18.1)	48 (13.8)	40 (17.1)
CA-125	567 (17)	242 (167)	0.831	152 (317)	52 (25)	38 (76)

Me = Median

* The AUC is computed using the training set patients only. An AUC > 0.5 indicates a positive relation to malignancy, an AUC < 0.5 indicates a negative relationship.

interrelated, multicollinearity is said to exist. In practical applications, there will always be some associations between some variables. However, if there is strong

Table 3.3: Descriptive statistics for binary variables (IOTA phase I data).

Variable	Benign N (%)	Malignant N (%)	Odds ratio* (95% CI)	Prim. Inv. N (%)	Borderline N (%)	Metastatic N (%)
pershistovca	800 (0.8)	266 (3.0)	3.6 (1.1 – 12.0)	169 (0.6)	55 (3.6)	42 (11.9)
pershistbrca	800 (2.9)	266 (5.6)	2.7 (1.2 – 5.7)	169 (6.5)	55 (0.0)	42 (9.5)
famhistovca	800 (2.5)	266 (4.9)	1.9 (0.8 – 4.6)	169 (5.9)	55 (3.6)	42 (2.4)
famhistbrca	800 (10.8)	266 (12.4)	1.2 (0.7 – 2.1)	169 (14.8)	55 (7.3)	42 (9.5)
nullipara	800 (41.9)	266 (25.2)	0.4 (0.3 – 0.6)	169 (25.4)	55 (27.3)	42 (21.4)
hysterectomy	800 (6.3)	266 (10.5)	1.9 (1.1 – 3.3)	169 (8.9)	55 (14.6)	42 (11.9)
menoyn	800 (32.9)	266 (63.5)	3.5 (2.5 – 4.9)	169 (69.2)	55 (47.3)	42 (61.9)
hormtherapy	800 (23.5)	266 (17.7)	0.7 (0.5 – 1.0)	169 (69.2)	55 (47.3)	42 (61.9)
pmb	229 (14.9)	145 (17.9)	1.4 (0.7 – 2.7)	102 (16.7)	17 (35.3)	25 (12.0)
bilateral	800 (16.6)	266 (31.2)	2.4 (1.7 – 3.5)	169 (37.3)	55 (12.7)	42 (31.0)
pain	800 (28.8)	266 (19.6)	0.5 (0.3 – 0.7)	169 (15.4)	55 (25.5)	42 (28.6)
ascites	800 (2.9)	266 (42.1)	20.9 (12.0 – 36.3)	169 (52.7)	55 (9.1)	42 (42.9)
incomplseptum	800 (9.3)	266 (4.1)	0.4 (0.2 – 0.8)	169 (4.1)	55 (3.6)	42 (4.8)
papillation	800 (19.5)	266 (45.5)	3.6 (2.5 – 5.1)	169 (45.6)	55 (63.6)	42 (21.4)
papflow	156 (33.3)	121 (84.3)	13.6 (6.5 – 28.6)	77 (89.6)	35 (71.4)	9 (88.9)
papsmooth	156 (49.4)	121 (82.6)	5.3 (2.7 – 10.4)	77 (81.8)	35 (82.9)	9 (88.9)
wallreg	800 (32.8)	266 (81.6)	9.8 (6.5 – 14.9)	169 (85.8)	55 (72.7)	42 (76.2)
shadows	800 (13.0)	266 (1.5)	0.1 (0.02 – 0.3)	169 (0.6)	55 (1.8)	42 (4.8)
venous	800 (9.0)	266 (3.4)	0.3 (0.2 – 0.8)	169 (3.6)	55 (5.5)	42 (0.0)
locularity						
locul1	800 (38.9)	266 (0.8)	0.01 (0.001 – 0.06)	169 (0.0)	55 (3.6)	42 (0.0)
locul2	800 (11.0)	266 (16.5)	1.8 (1.1 – 2.9)	169 (16.6)	55 (25.5)	42 (4.8)
locul3	800 (22.0)	266 (7.5)	0.3 (0.2 – 0.5)	169 (5.3)	55 (18.2)	42 (2.4)
locul4	800 (21.0)	266 (43.6)	3.1 (2.2 – 4.4)	169 (45.0)	55 (47.3)	42 (33.3)
locul5	800 (6.5)	266 (31.6)	6.4 (4.0 – 10.0)	169 (33.1)	55 (5.5)	42 (59.5)
echogenicity						
echo1	800 (37.9)	266 (40.2)	1.2 (0.8 – 1.7)	169 (43.2)	55 (34.6)	42 (35.7)
echo2	800 (18.6)	266 (22.6)	1.2 (0.8 – 1.8)	169 (20.7)	55 (32.7)	42 (16.7)
echo3	800 (24.0)	266 (12.4)	0.4 (0.3 – 0.7)	169 (9.5)	55 (23.6)	42 (9.5)
echo4	800 (1.0)	266 (0.8)	0.8 (0.2 – 4.1)	169 (0.6)	55 (0.0)	42 (2.4)
echo5	800 (14.3)	266 (6.8)	0.4 (0.2 – 0.7)	169 (7.7)	55 (5.5)	42 (4.8)

* The odds ratio is computed using the training set patients only.

multicollinearity, the analysis may become unstable. In this case, it is good practice to change the set of variables for example by performing principal component analysis to project the set of variables into a smaller set of new variables or by simply removing variables from the input selection analysis. Principal component analysis is an interesting tool to tackle strong multicollinearity, yet for our applications it is not desirable. The smaller set of new variables is created from the original set of variables, making it necessary to always measure all the original variables. Clinicians want to make a good prediction for which a small set of measurements is needed.

Multicollinearity can be checked by computing intercorrelations. However, this only detects bivariate associations, whereas often more than two variables are

interrelated. Therefore, an analysis using variance inflation factors (VIF) is useful as well. For logistic regression, some people are concerned when VIFs exceed 2.5 [10]. We applied VIF analysis to a set of 51 variables. The categorical variables locularity and echogenicity were both entered using five dummy variables: locul1 (unilocular tumor), locul2 (unilocular-solid tumor), locul3 (multilocular tumor), locul4 (multilocular-solid tumor), and locul5 (entirely solid tumor); echo1 (anechoic), echo2 (low-level echogenic), echo3 (ground glass), echo4 (hemorrhagic), and echo5 (mixed echogenic). Due to high skewness, some variables were log-transformed. In a few steps, we left out the following variables: log(TAMX), log(PSV), RI, log(CA-125), log(lesvol), solidvol, ratiosolles, papmax, papillation, postmenoy, menoyn, locul2, locul4, and echo1. The VIF for log(CA-125) was small but since CA-125 levels were not recorded for several patients, we decided to drop this variable such that all training cases can be used to determine VIFs. Moreover, it was not our aim to include CA-125 in mathematical models for ovarian tumors. The initial and final VIFs are presented in Table 3.4.

The average of the initial VIFs (including all 51 variables) was 9.6, which is high. The average of the final VIFs (including a selection of 36 variables) was 2.0. The Doppler indices TAMX, PSV, PI, and RI were highly correlated (Pearson correlations were between 0.72 and 0.99), such that it is sufficient to retain one of these four variables. Obviously, maximum diameter variables (of largest papillary structure, lesion, and largest solid component) were strongly related to corresponding volume variables (Pearson correlations were between 0.75 and 0.91), such that it is sufficient to retain one of both. Equally obvious is the strong association between number of years postmenopause (postmenoy) and menopausal status (menoyn) on the one hand and age on the other (Pearson correlations were 0.75 and 0.81). Retaining age is most logical. Another natural association is that between the percentage solid tissue in the lesion (ratiosolles) and locul5, with a Pearson correlation of 0.88. Since ratiosolles is associated to other variables as well, its VIF is high and thus this variable was dropped. Finally, some variables are binary versions of others. For example, papillation (presence of papillary structures) is a binary version of papnr (number of papillary structures). This explains the high VIF for papillation. Nullipara and parity is another such example, yet these variables do not have an elevated VIF.

IOTA phase Ib data set

This data set consists of 507 patients collected at the three main centers from IOTA phase I: University Hospitals Katholieke Universiteit Leuven, Malmö University Hospital at Lund University, and Università Cattolica del Sacro Cuore in Rome. The mean age was 47 years. Data collection started in 2002 and ended in 2005. The number of women contributed by each center and the distribution of outcome per center is shown in Table 3.5. Descriptive statistics are presented in Tables 3.6 and 3.7.

Table 3.4: Variance inflation factors for the IOTA phase I training data, sorted by the initial VIF.

Variable	Initial VIF	Final VIF	Variable	Initial VIF	Final VIF
log(TAMX)	56.2		papsmooth	3.7	3.5
locul4	51.3		menoyn	3.5	
log(PSV)	45.3		postmenoy	3.5	
locul1	41.1	2.8	papflow	3.3	3.1
locul5	32.8	2.5	nullipara	2.7	2.6
locul3	32.0	2.0	septum	2.7	1.9
locul2	25.4		parity	2.5	2.4
PI	17.7	2.8	wallregularity	2.4	1.9
papmax	15.2		ascites	2.4	2.1
RI	13.5		fluid	2.2	2.0
solidmax	11.7	3.0	ratiopaples	2.2	1.6
echo1	11.0		log(CA-125)	2.1	
ratiosolles	9.4		venous	1.6	1.4
papillation	8.2		shadows	1.5	1.5
echo2	8.0	1.3	hysterectomy	1.5	1.1
lesmax	7.5	2.0	echo4	1.3	
echo3	7.3	1.4	pmb	1.2	1.2
log(lesvol)	7.2		pain	1.2	1.2
papheight	6.8	4.3	pershistovca	1.2	1.1
age	6.0	1.7	incomplseptum	1.2	1.1
echo5	5.4	1.5	hormtherapy	1.2	1.1
papnr	4.7	3.6	famovca	1.2	1.1
solidvol	4.7		bilateral	1.1	1.1
papvol	4.6	2.4	fambrca	1.1	1.1
colscore	4.4	2.8	pershistbrca	1.1	1.1
nrlocules	3.9	2.3	*average VIF*	9.6	2.0

Table 3.5: The distribution of the outcome in the three centers participating in IOTA phase Ib.

	Number of	Outcome			
Center (code)	Patients	Benign	Primary invasive	Borderline	Metastatic
Leuven (LBE)	287, 56.6%	202, 70.4%	62	12	11
Malmö (MSW)	124, 24.5%	98, 79.0%	17	6	3
Roma (RIT)	96, 18.9%	64, 66.7%	24	2	6

IOTA phase II data set

Data collection for the IOTA phase II data set was initiated in 2005 and ended in November 2007. Less variables were recorded as compared to IOTA phase I. Some uninteresting variables were not considered any more. Data collection was done using a software program designed by Astraia Software GMBH (München, Germany, www.astraia.de). Collected data could be sent automatically to a data

Table 3.6: Descriptive statistics for selected continuous/ordinal variables (IOTA phase Ib data).

Variable	Benign N, Me	Malignant N, Me	AUC*	Prim. Inv. N, Me	Borderline N, Me	Metastatic N, Me
age	364 (41.5)	143 (55)	0.739	103 (56)	20 (48)	20 (56)
lesmax	364 (62)	143 (100)	0.694	103 (100)	20 (115.5)	20 (98.5)
papheight	70 (7.5)	69 (12)	0.707	52 (12)	12 (13)	5 (9)
papnr	70 (1)	69 (4)	0.724	52 (4)	12 (4)	5 (1)
solidmax	144 (25)	138 (58.5)	0.739	100 (62)	18 (37.5)	20 (68)
colscore	364 (2)	143 (3)	0.784	103 (3)	20 (3)	20 (3)
PI	200 (0.88)	125 (0.71)	0.373	92 (0.73)	17 (0.71)	16 (0.72)
RI	200 (0.56)	125 (0.51)	0.388	92 (0.52)	17 (0.49)	16 (0.51)
PSV	200 (11.7)	125 (22)	0.743	92 (22.8)	17 (19)	16 (24.7)
TAMX	200 (7)	125 (15.1)	0.763	92 (15.6)	17 (13.6)	16 (15.5)
CA-125	217 (18)	128 (264)	0.862	95 (321)	20 (55)	20 (242)

Me = Median

* An AUC > 0.5 (> 0.5) indicates a positive (negative) relation to malignancy.

Table 3.7: Descriptive statistics for selected binary variables (IOTA phase Ib data).

Variable	Benign N (%)	Malignant N (%)	Odds ratio (95% CI)	Prim. Inv. N (%)	Borderline N (%)	Metastatic N (%)
pershistovca	364 (0.8)	143 (2.8)	3.5 (0.8 – 15.7)	103 (2.9)	20 (5.0)	20 (0.0)
pershistbrca	364 (3.3)	143 (2.8)	0.8 (0.3 – 2.7)	103 (2.9)	20 (5.0)	20 (0.0)
menoyn	364 (29.1)	143 (67.8)	5.1 (3.4 – 7.8)	103 (71.8)	20 (40.0)	20 (75.0)
hormtherapy	364 (21.2)	143 (14.0)	0.6 (0.4 – 1.0)	103 (13.6)	20 (15.0)	20 (15.0)
bilateral	364 (14.0)	143 (32.9)	3.0 (1.9 – 4.7)	103 (33.0)	20 (15.0)	20 (50.0)
pain	364 (27.2)	143 (23.1)	0.8 (0.5 – 1.3)	103 (26.2)	20 (20.0)	20 (10.0)
ascites	364 (1.1)	143 (37.8)	54.6 (19.3 – 155)	103 (40.8)	20 (20.0)	20 (40.0)
papflow	70 (27.1)	69 (88.4)	20.5 (8.3 – 50.6)	52 (92.3)	12 (75.0)	5 (80.0)
wallreg	364 (40.7)	143 (75.5)	4.5 (2.9 – 7.0)	103 (79.6)	20 (75.0)	20 (55.0)
shadows	364 (17.9)	143 (1.4)	0.1 (0.02 – 0.3)	103 (1.9)	20 (0.0)	20 (0.0)
locularity						
locul1	364 (39.0)	143 (1.4)	0.02 (0.005 – 0.09)	103 (1.0)	20 (5.0)	20 (0.0)
locul2	364 (12.6)	143 (16.8)	1.4 (0.8 – 2.4)	103 (15.5)	20 (35.0)	20 (5.0)
locul3	364 (20.3)	143 (2.1)	0.1 (0.03 – 0.3)	103 (1.9)	20 (5.0)	20 (0.0)
locul4	364 (15.7)	143 (41.3)	3.8 (2.4 – 5.9)	103 (42.7)	20 (50.0)	20 (25.0)
locul5	364 (11.3)	143 (38.5)	4.9 (3.1 – 7.9)	103 (38.8)	20 (5.0)	20 (70.0)
echogenicity						
echo1	364 (31.6)	143 (21.0)	0.6 (0.4 – 0.9)	103 (25.2)	20 (10.0)	20 (10.0)
echo2	364 (10.7)	143 (21.0)	2.2 (1.3 – 3.7)	103 (17.5)	20 (60.0)	20 (0.0)
echo3	364 (28.3)	143 (4.9)	0.1 (0.06 – 0.3)	103 (5.8)	20 (5.0)	20 (0.0)
echo4	364 (0.3)	143 (1.4)	5.1 (0.5 – 57.2)	103 (1.9)	20 (0.0)	20 (0.0)
echo5	364 (17.9)	143 (13.3)	0.7 (0.4 – 1.2)	103 (10.7)	20 (20.0)	20 (20.0)

server maintained at ESAT (Katholieke Universiteit Leuven). Participating centers

could, if desired, also send their data by e-mail or using CD-ROMs. The phase II centers can be divided into two groups: the centers that participated in phase I, and new centers. There are data from twenty centers in eight countries (Italy, Belgium, Sweden, UK, Poland, Czech Republic, China, and Canada).

Because data cleaning is still in progress at this very moment, a preliminary data set was used in this thesis. This data set contains patients for which the data were complete and correct as of January 8, 2008. The data set contains 1916 patients from 18 centers. The mean age was 46 years. There were 1394 benign (72.8%), 365 primary invasive (19.0%), 101 borderline (5.3%), and 56 metastatic invasive (2.9%) tumors. Seven centers also participated in IOTA phase I (993 patients, 51.8%). The remaining 11 centers were newly recruited (923, 48.2%). The complete list of centers, together with the number of contributed patients and distribution of tumor type, can be found in Table 3.8. Because this data set is only preliminary, no tables with descriptive statistics are shown.

Table 3.8: The distribution of the outcome in the centers participating in IOTA phase II.

	Number of	Outcome			
Center (code)	Patients	Benign	Primary invasive	Borderline	Metastatic
Italy	957, 49.9%	719, 75.1%	167	40	31
Monza (OIT)*	246, 12.8%	196, 79.7%	31	15	4
Roma (RIT)*	210, 11.0%	119, 56.7%	68	12	11
Sardinia (SIT)	154, 8.0%	134, 87.0%	13	3	4
Bologna (BIT)	132, 6.9%	122, 92.4%	5	4	1
Milano2 (CIT)	71, 3.7%	33, 46.5%	27	4	7
Napoli (NIT)*	63, 3.3%	50, 79.4%	9	2	2
Milano (MIT)*	45, 2.3%	41, 91.1%	3	0	1
Udine (UDI)	17, 0.9%	10, 58.8%	6	0	1
Milano3 (VIT)	11, 0.6%	11, 100%	0	0	0
Napoli2 (GIT)	8, 0.4%	3, 37.5%	5	0	0
Belgium	415, 21.7%	304, 73.3%	71	27	13
Leuven (LBE)*	228, 11.9%	143, 62.7%	51	22	12
Genk (GBE)	187, 9.8%	161, 86.1%	20	5	1
Sweden	174, 9.1%	148, 81.0%	25	7	1
Malmö (MSW)*	136, 7.1%	109, 80.2%	21	6	0
Lund (LSW)	38, 2.0%	32, 84.2%	4	1	1
Lublin, Poland (LPO)	149, 7.8%	97, 65.1%	45	3	4
Prague, Czech Rep. (PCR)	90, 4.7%	39, 43.3%	35	15	1
Beijing, China (BCH)	66, 3.4%	55, 83.3%	8	0	3
London, UK (KUK)*	65, 3.4%	39, 60.0%	14	9	3

* Centers that participated in IOTA phase I.

3.2.2 PULs: from St. George's to IPULA

The data set for the classification of PULs consisted of women presenting with a PUL at the Early Pregnancy Unit of St. George's Hospital (University of London) in London (UK) between June 2001 and October 2004. All women underwent TVS. Also, peripheral blood was taken to measure serum hCG and progesterone levels at presentation and 48 hours later. Criteria for exclusion from the study population were: any evidence of an intra-uterine sac at the first ultrasound scan, an adnexal mass thought to be an EP at the initial ultrasound scan, an endometrial thickness of >15 mm on TVS with the presence of heterogeneous irregular tissues within the uterus thought to be an incomplete miscarriage, or clinical instability or signs of intra-abdominal bleeding or a hemoperitoneum according to the scan images. All women were followed up with TVS, hCG measurements, and/or laparoscopy until the final clinical status of the pregnancy was established (the outcome). Possible outcomes were failing PUL (FP), normal intra-uterine pregnancy (IUP), ectopic pregnancy (EP), and, in rare occasions, persisting PUL. Persisting PULs refer to women in whom no pregnancy was seen at any time. These cases were excluded from the analysis because the final outcome was unknown.

The diagnosis of a failing PUL was defined as a serum progesterone level below 20 nmol/L at presentation [34, 92], with a subsequent fall in serum hCG levels to 5 IU/L, and the location of these pregnancies remaining unknown. The diagnosis of an intra-uterine pregnancy (IUP) was made on TVS when a gestational sac was visualized within the endometrial cavity [459]. An EP was diagnosed using TVS and/or laparoscopy with confirmatory histology of chorionic villi. At St. George's, the use of ultrasound as the primary way to diagnose EPs is accurate in 93.2% of the cases [103]. The diagnosis of EP was based on the positive visualization at TVS examination of an adnexal mass. Ultrasound diagnosis of an EP was made if one of the following grey-scale appearances was present: an inhomogeneous area adjacent to the ovary and moving separate to this, known as the blob sign [103], a mass with a hyperechoic ring around the gestational sac, referred to as the bagel sign [103], or a gestational sac with a fetal pole with or without cardiac activity [103].

In total, 14 variables were recorded. First of all, the hCG ratio and the average hCG (IU/L) were computed. 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 patient's age was recorded, as well as the gestational age (gestation), thickness of the endometrium (endthick), character of the midline echo (midline; intact versus disrupted), and presence of free fluid in the pouch of Douglas (freefluid). Finally, the following presenting complaints were documented: 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), presence of lower abdominal pain (pain), history of ectopic pregnancies (prevect; yes or no) or miscarriages (prevmisc; yes or no), unsure date of last menstrual period (unsuredates; yes or no), or maternal

anxiety (anxiety; yes or no). The two latter presenting complaints were omitted from further analysis based on a priori discussions with the clinicians.

Table 3.9: Descriptive statistics for the PUL data.

	Failing	IUP	Ectopic	AUC†		
	PUL (n=460)	(n=330)	(n=66)	IUP vs	IUP vs	EP vs
Cont. variable	Me (IQR)	Me (IQR)	Me (IQR)	Failing	EP	Failing
hCG, 0hr (IU/L)	155 (456)	467 (558)	531 (904)	0.699	0.474	0.716
hCG, 48hr (IU/L)	75.5 (208)	1065.5 (1233)	575 (936)	0.918	0.637	0.861
hCGratio	0.41 (0.31)	2.20 (0.60)	1.17 (0.55)	0.982	0.908	0.906
hCGmean	124.5 (330)	754.8 (885)	535.3 (1136)	0.830	0.575	0.784
Prog., 0hr (nmol/L)	4 (7.5)	68 (36)	29 (27)	0.971	0.803	0.892
Prog., 48hr (nmol/L)	3 (3)	63 (39)	23.5 (25)	0.984	0.818	0.939
progmean	3.5 (5)	66.8 (35)	26 (24)	0.980	0.816	0.918
endthick (mm)	8.3 (5.9)	13.5 (5.7)	10.3 (8.2)	0.795	0.651	0.631
age (years)	31 (9)	29 (10)	30 (6)	0.403	0.407	0.475
gestation (days)	44 (14.5)	32 (8)	41 (11)	0.255	0.283	0.462
Bin./ord. variable	N_{Yes} (%)	N_{Yes} (%)	N_{Yes} (%)		OR‡	
vaginal bleeding						
none	109 (23.7)	251 (76.1)	23 (34.8)	10.2	7.4	1.4
without clots	248 (53.9)	76 (23.0)	38 (57.6)	0.2	0.2	1.3
with clots	103 (22.4)	3 (0.9)	5 (7.6)	0.04	0.1	0.4
pain	216 (47.0)	161 (48.8)	25 (37.9)	1.6	2.0	0.8
prevect	10 (2.2)	21 (6.4)	8 (12.1)	4.0	0.5	7.3
prevmisc	2 (0.4)	7 (2.1)	0 (0.0)	6.3	-	-
midline	95 (20.7)	15 (4.6)	11 (16.7)	0.1	0.1	1.4
freefluid	49 (10.7)	93 (28.2)	10 (15.2)	3.9	1.6	2.4
anxiety	10 (2.2)	16 (4.8)	1 (1.5)	3.6	-	-
unsuredates	13 (2.8)	61 (18.5)	3 (4.5)	6.5	2.2	3.0

Me = Median; N_{Yes}: N for which the condition is present

† The AUCs are computed using the training set patients only. An AUC > 0.5 indicates a positive relationship, an AUC < 0.5 a negative relationship.

‡ The ORs are computed using the training set patients only.

In total, 1003 PULs were included in the data set. Of these, 872 had complete data (87%). The clinicians confirmed that the missing data were 'missing completely at random' (MCAR) [265]. This means that the complete cases can be seen as a random (i.e. not selective) subsample of the complete data set. Therefore, we decided to continue with the complete cases. This can be dangerous if missing values are caused by other observed or unobserved measurements. Sixteen of the complete cases were persisting PULs and were discarded. Most of the remaining 856 PULs were failing PULs (460, 53.7%), 330 were IUPs (38.6%), and 66 were ectopic pregnancies (7.7%). Descriptive statistics for the 856 PULs are presented in Table 3.9. The data were chronologically split up in a training and a test set. The first 508 cases (59%) were put in the training set, the remaining 348 cases in the test

set. The training set contained 42 ectopic pregnancies while the test set contained 24.

In the spirit of the IOTA study group, George Condous (Nepean Hospital, University of Sydney, Australia), who worked extensively on PULs while affiliated with St. George's Hospital, and Tom Bourne are working at this moment to establish the International Pregnancy of Unknown Location Analysis (IPULA) group, which aims at collecting a large data set of PULs in different centers from Belgium, Italy, the UK, Australia, Singapore, Canada, and the USA.

Chapter 4

Issues in probabilistic multi-class classification

Abstract – In this chapter, we will discuss two issues related to probabilistic multi-class classification. Many methods to obtain multi-class probabilities exist, some of which are based on the combination of binary classifiers while others use an 'all-at-once' procedure. The first part of this chapter compares many methods using LS-SVM based algorithms. The comparison is performed on the data for ovarian tumors and pregnancies of unknown location. With respect to model evaluation, approaches to extend binary AUC metrics to multi-class alternatives have been described in the literature. However, a disadvantage of AUC metrics is that they do only look at discriminatory ability. The magnitude of predicted probabilities is ignored. Therefore, the second part of this chapter describes two binary AUC metrics that incorporate probability information, and suggests extensions to the multi-class case.

4.1 Comparison of methods for multi-class probabilities

4.1.1 Introduction

Multi-class classification is less straightforward than binary classification, in particular for margin-based classifiers such as support vector machines (SVMs). Methods for multi-class tasks are based on the combination of binary classifiers or on 'all-at-once' classification. Binary classifiers are typically constructed by contrasting all pairs of classes (pairwise or 1-vs-1) [160], by contrasting each class with all other classes combined (1-vs-All) [337], or by other techniques such as minimal output coding (MOC) [374] or error-correcting output coding (ECOC) [130, 11]. In ECOC, each class k is given a codeword $c_k \in \{-1, +1\}^L$ of length L. Then, for each position in the codeword, a binary classifier is constructed in which the classes with value -1 in the codeword are contrasted with the classes with value $+1$. ECOC uses longer codewords than MOC, where a codeword of length L is used for problems with up to

2^L classes. In the 1-vs-1 approach final class predictions are often made using max-wins voting [160, 137], but the use of a directed acyclic graph has also been suggested [325]. For 1-vs-All classification, the final decision is typically made using winner-takes-all, where the class whose classifier (contrasting it with all other classes) gives the best result becomes the predicted class. Using ECOC, class decisions are often based on Hamming decoding or loss based decoding.

Often, however, one is interested in class probabilities. These are important because they give information about the uncertainty of class membership as opposed to black-and-white class predictions. In medical decision making, uncertainty information can influence the optimal treatment of patients. In this section, we compare several methods to obtain multi-class probability estimates using the data sets on pregnancies of unknown location and ovarian tumors. For both conditions, early probabilistic predictions are needed for optimizing patient care and its financial implications. All-at-once, 1-vs-1, and 1-vs-All methods were used. Some methods that we did not use in this study can be found in [246, 440, 380, 476].

4.1.2 Obtaining multi-class probability estimates

Creating probabilistic output from binary (LS-)SVMs

We implemented four techniques to get probabilistic outputs from the binary LS-SVM classifiers (with RBF kernel): a Bayesian LS-SVM [441] with and without sparseness induction, and a transformation of the LS-SVM latent output using either an improvement on Platt's method [324, 260] or isotonic regression [477]. For non-Bayesian LS-SVMs, the hyperparameters γ and σ have to be tuned using the training data. To do this, we applied the leave-one-out method suggested in [75] with the predicted residual sum-of-squares (PRESS) statistic. The techniques are abbreviated using this `style`.

Bayesian LS-SVM (`bay`; `bays`). Bayesian LS-SVMs are an approach to LS-SVMs that results in binary class probabilities. This method is explained in 2.2.3. Since for LS-SVMs $\alpha_n = \gamma e_n$, support values can be negative for easy cases. Sparseness can be imposed (`bays`) by repeatedly pruning training data with negative support values until none are left [267].

LS-SVM + Platt (`plt`). Based on (LS-)SVM output, Platt [324] estimates the class probability using

$$P(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp(Af + B)}, \tag{4.1}$$

where the LS-SVM latent output $f = \sum_{n=1}^{N} \alpha_n y_n K(\mathbf{x}, \mathbf{x}_n) + b$. Training data f values were used to find A and B using maximum likelihood estimation. These f

values were derived using 5-fold cross-validation (5CV) on the training data in order to obtain less biased values. We used an improved implementation of Platt's method [260]. Figure 4.1 presents an example of Platt's method. The bold line represents the 'true' relationship between f and the probability for class 1 (as obtained using nonparametric regression). The smooth thin line is the approximation based on Platt's method.

LS-SVM + isotonic regression (`iso`). Zadrozny and Elkan [477] propose the use of isotonic regression to obtain probabilities based on (LS-)SVM output (f). This nonparametric regression technique maps f to probabilities using a monotonically increasing function. The isotonic regression method used is that of pooled adjacent violators [29]. The training cases are ranked with respect to f. The corresponding class membership (coded as 0 versus 1) is taken as the initial estimated probability of belonging to class 1. If these initial probabilities are entirely separated in the ranking (i.e. they are isotonic), they are taken as the final estimates. If the initial estimates are not isotonic for two cases, both probabilities are averaged to serve as the new estimates. This is repeated until the estimated probabilities are isotonic. Training data f values used for the isotonic regression were again based on 5CV. When a new f value fell between two training data f values with different estimated probabilities, we estimated the probability for the new output value using simple linear interpolation. Figure 4.1 presents an example of isotonic regression, and shows the different operation of Platt's method and isotonic regression. The step function is the approximation of the 'true' relationship based on isotonic regression, while the smooth thin line is the approximation based on Platt's method.

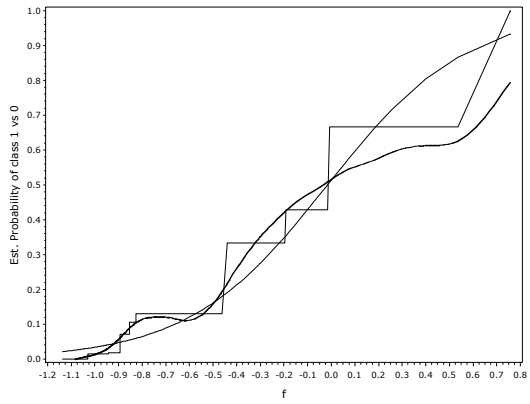


Figure 4.1: Visualization of Platt's method and isotonic regression (bold line: true relationship, smooth line: Platt's approximation, step function: isotonic regression).

Methods to combine binary probabilities

Let us denote the probability of a case to belong to the k^{th} of K classes as $\pi_k = P(Y = k|\mathbf{x})$ (estimated by $\hat{\pi}_k$), and the probability of a case to belong to the k^{th} class conditional on membership of class k or l as $\pi_{kl} = P(Y = k|Y \in \{k, l\}, \mathbf{x})$, (estimated by $\hat{\pi}_{kl}$). To combine binary probabilities into multi-class probabilities, ten methods were used that will be shortly described. The methods are abbreviated using this `style`.

Refregier and Vallet (`rvall`; `rvone`) [336]. Taking two classes k and l, one starts from the relation

$$\frac{\hat{\pi}_{kl}}{\hat{\pi}_{lk}} \approx \frac{\pi_k}{\pi_l}, \tag{4.2}$$

to estimate the multi-class probabilities using any $K - 1$ binary classifiers. Because $\sum_{k=1}^{K} \pi_k = 1$ the multi-class probabilities are estimated by solving a linear system. A problem is that one typically cannot choose which $K - 1$ binary classifiers to use. We implemented this method once by averaging the multi-class probabilities for all subsets of $K - 1$ binary classifiers (`rvall`) and once by randomly selecting one such subset (`rvall`).

Ensembles of nested dichotomies (`endall`; `endone`) [157]. This method obtains multi-class probabilities by walking through a tree of nested binary classifiers. At the top level, all classes are divided into two subgroups for which a binary classifier is trained. Subgroups that consist of more than one class are again divided into two groups until all classes form a subgroup of their own. The tree then consists of K leaves. The multi-class probability $\hat{\pi}_k$ is obtained by multiplying all binary probabilities involving class k. Figure 4.2 presents a tree of nested classifiers for a four-class problem. The estimated probability of class 4 would be obtained by multiplying $\hat{\pi}_{1,3,4}$ (first binary classifier) with $\hat{\pi}_{1,4|1,3,4}$ (second binary classifier) and $\hat{\pi}_{4|1}$ (third binary classifier). For any multi-class problem, multiple trees can be constructed that are typically equally plausible a priori. We implemented this method once by averaging the multi-class probabilities of all trees (`endall`) and once by randomly selecting one tree (`endone`). Of course, the number of different trees for K classes, $T(K)$, grows exponentially with K:

$$T(K) = (2K - 3)T(K - 1). \tag{4.3}$$

This makes `endall` impractical for large classes: $T(4) = 15$, but $T(5) = 105$. Experiments in [157] show, however, that an average of 10 to 20 randomly selected trees is sufficient.

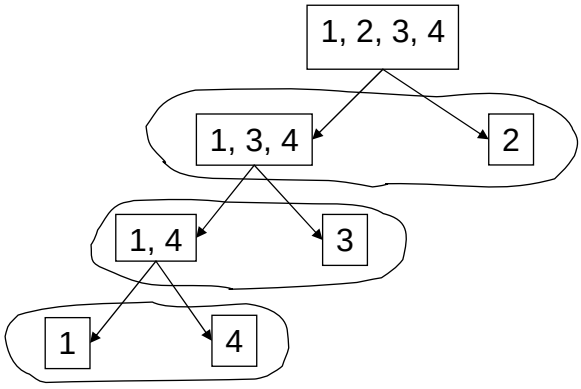


Figure 4.2: Example of a tree of nested binary classifiers for a 4-class problem.

Price *et al.* (`pkpd`) **[328].** By stating that

$$\sum_{l=1,l\neq k}^{K} (\pi_k + \pi_l) - (K-2)\pi_k = 1, \tag{4.4}$$

and using the following approximation

$$\hat{\pi}_{kl} \approx \pi_{kl} = \frac{\pi_k}{(\pi_k + \pi_l)}, \tag{4.5}$$

one ends up with

$$\pi_k \approx \frac{1}{\sum_{l=1,k\neq l}^{K} (1/\hat{\pi}_{kl}) - (K-2)}. \tag{4.6}$$

The result is normalized to ensure that $\sum_{k=1}^{K} \hat{\pi}_k = 1$.

1-versus-All normalization (`1van`)**.** This method simply normalizes all obtained 1-versus-All probabilities such that they sum to 1.

Pairwise coupling (`pc-ht; pc-wu1; pc-wu2`)**.** The pairwise coupling method was introduced by Hastie and Tibshirani [185]. They estimate π_k by maximizing the negative weighted Kullback-Leibler distance ℓ_{KL} between π_{kl} and $\hat{\pi}_{kl}$. After setting the derivative of ℓ_{KL} with respect to π_k to zero, they aim to estimate the multi-class probabilities π_k such that

$$\sum_{l=1,k\neq l}^{K} n_{kl}\pi_{kl} = \sum_{l=1,k\neq l}^{K} n_{kl}\hat{\pi}_{kl} \text{ with } \sum_{k=1}^{K} \pi_k = 1,\ \pi_k > 0\,, \tag{4.7}$$

where n_{kl} represents the number of training cases belonging to class k or l. An iterative procedure is used to solve this system per input vector $\mathbf{x}$. This method is called `pc-ht`.

Wu and colleagues [472] derived two other pairwise coupling methods, based on the assumption that the multi-class data are balanced such that weighting is not necessary. They suggest to solve

$$\pi_k = \sum_{l=1,k\neq l}^{K} \frac{\pi_k + \pi_l}{K-1}\hat{\pi}_{kl},\ \forall k,\ \text{with}\ \sum_{k=1}^{K}\pi_k = 1,\ \pi_k \geq 0\,, \tag{4.8}$$

which can be done by a linear system. We call this method `pc-wu1`. These authors also suggest a second approach

$$\min_{\pi_1,\ldots,\pi_K} \sum_{k=1}^{K}\sum_{l=1,k\neq l}^{K} \left(\hat{\pi}_{lk}\pi_k - \hat{\pi}_{kl}\pi_l\right)^2 \text{ with } \sum_{k=1}^{K}\pi_k = 1,\ \pi_k \geq 0\,, \tag{4.9}$$

which is again solved by a linear system. Note that the expression in equation 4.9 is similar to the basis of Refregier and Vallet's method [336]. We call this method `pc-wu2`.

Huang *et al.* (`gbt1va`) [196]. These authors proposed methods based on the generalized Bradley-Terry model [196]. The Bradley-Terry model [61] estimates the 'quality' of objects based on paired preference data. It is often used in sports, where the 'quality of an object' relates to the skill of a player, and the paired preference data consist of information on how many times players have beaten each other. In the context of multi-class classification, we talk of estimating π_k and $\hat{\pi}_{kl}$ can be seen as the probability that class k beats class l. Hastie and Tibshirani's pairwise coupling method [185] is formally similar to a Bradley-Terry model, but note that the different $\hat{\pi}_{kl}$ values are not independent. In the generalized Bradley-Terry model the players (i.e. classes) are repeatedly divided into two teams playing a series of games. The model is again used to estimate the players' individual skill (i.e. π_k) based on the outcome of the games. We implemented the method where teams are created using the 1-versus-All strategy.

Duan *et al.*'s softmax methods (`sm1va`; `sm1v1`)

Duan and colleagues [137] propose to apply softmax functions to the latent output of 1-versus-All or 1-versus-1 (LS-)SVMs to directly estimate multi-class probabilities. For the 1-versus-All strategy, the multi-class probabilities are computed as

$$\hat{\pi}_k = \exp(a_k f_k + b_k) \Bigg/ \left(\sum_{l=1}^{K}\exp(a_l f_l + b_l)\right), \tag{4.10}$$

where f_k is the LS-SVM latent output for the 1-versus-All classifier for class k, and a_k and b_k ($i = 1, \ldots, k$) are parameters to be tuned by minimizing a negative log-likelihood function on the training data. For the 1-versus-1 strategy, the probabilities are computed analogously using the LS-SVM outputs for all 1-versus-1 classifiers [137]. Regularized versions were also suggested in [137], but did not perform better in their own evaluation. The latent output f for the 1-versus-All or 1-versus-1 LS-SVM models were obtained by 5CV on the training data.

Multi-class Kernel Logistic Regression (`mklr`)

Finally, an all-at-once method was used, being the weighted LS-SVM based version of multi-class kernel logistic regression with RBF kernel [220], that was explained in section 2.2.5. The hyperparameters γ and σ were tuned using ten-fold cross-validation.

4.1.3 Experimental Setup

Data

In this comparison study, we used the IOTA phase I and the PUL data sets. Because the results of the analyses in the present section can have consequences for classifier development, we only used the training parts of the data sets. Therefore, the first data set contains 508 women with a PUL. A PUL can turn out to be any of three outcomes ($K = 3$): a failing PUL (283 cases, 56%), a normal intra-uterine pregnancy (IUP) (183, 36%), or an ectopic pregnancy (42, 8%). The last class the most difficult one, because it is small does not seem to be well separated from both other groups. The inputs used to predict the outcome ($p = 5$) are the ratio of the hCG levels at presentation and 48 hours later, the logarithm of the average hCG level, the logarithm of the average progesterone level, vaginal bleeding, and the mother's age. The second data set contains 754 patients with at least one ovarian tumor (in case of bilateral tumors only the most complex tumor is included), collected by the IOTA study group. The type of ovarian tumor is one of the following ($K = 4$): benign (563 cases, 75%), primary invasive (121, 16%), borderline (40, 5%), and metastatic (30, 4%). Again, the classes are clearly unbalanced. This time, the borderline class will probably be the most difficult one since it is small and contains 'borderline malignant' tumors. The inputs used to predict the outcome ($p = 9$) are the woman's age, the maximal diameter of the lesion and of the largest solid component, and the presence of i) a personal history of ovarian cancer, ii) blood flow within the papillary projections, iii) irregular internal cyst walls, iv) ascites, v) an entirely solid tumor, vi) bilateral tumors.

Performance Evaluation

The performance of all methods was evaluated using stratified 5CV [241]. The data were split up in 5 folds of equal size with equal target class distributions. Each

fold served once as test set to test the methods after they were applied to the other four folds that served as the training set. This resulted in five test set results for each method. This cross-validation procedure was repeated 20 times [241] such that we obtained 100 test set results. All inputs were standardized to zero mean and unit variance separately in each of the 100 training sets. The inputs in the accompanying test sets were transformed with the same formulas used for their training set colleagues. The test set results were summarized by the AUC and by the average entropy error (log loss). Even though multi-class AUC measures exist, we preferred to consider a separate AUC for each class. The average entropy is computed as

$$\frac{1}{N}\sum_{n=1}^{N} -log(\hat{\pi}_y), \tag{4.11}$$

where $\hat{\pi}_y$) represents the estimated probability for a case's true class. The higher the estimated class probability of each case's true class, the lower the average entropy. The AUC and entropy results of the 100 test set evaluations were summarized by their median and interquartile range (IQR).

Due to the complex experimental setup, we adopted a heuristic strategy to compare methods. Based on [62], we used average ranks (AR). When comparing the ten methods to combine binary class probabilities, their performance was ranked for each of the algorithms to obtain binary probabilities (`bay`, `bays`, `plt`, `iso`) and the resulting ranks were averaged. The opposite was done when comparing the four algorithms for obtaining binary probabilities. This method summarized all obtained results well.

4.1.4 Results and Discussion

Because the first data set has three classes and the second has four, we have AUC information on seven classes. To reduce the output, we averaged the K AUCs for each data set. Lack of space precludes showing the full AUC results. The results for the PUL and IOTA data are shown in Table 4.1. Even though differences between methods were often not huge, the best probability combination methods were `endall` and `pc-wu1`. `pc-wu2` and `gbt1va` also performed very well. `gbt1va`'s AR of 5.0 on the IOTA data set is caused by degraded performance for the two small classes when applying `gbt1va` to `bay`. Otherwise, this method performed excellent. The methods `pkpd`, `endone`, and `rvone` were outperformed. Differences in average AUC were reduced by the 'easier' classes. Bigger differences arose for the difficult classes ectopic pregnancy (PUL), borderline (IOTA), and metastatic (IOTA). Similar conclusions could be drawn when looking at these classes only: `endall` performed best, followed by `pc-wu1`, `pc-wu2`, and `gbt1va`.

The entropy results are shown in Table 4.2. This leads to a similar observation: `endall`, `gbt1va`, and `pc-wu1` performed best among the probability combination methods. `pkpd`, `endone`, `rvone`, and `1van` performed worst.

With respect to both softmax methods [137], we can see that their performance did not differ substantially, yet `sm1va` always had the advantage over `sm1v1`. Also, their performance was similar to that of the best probability combination methods. When looking at AUC performance, however, these methods seemed to perform less than the probability combination methods applied to `bays`.

Interestingly, `mklr` had excellent performance throughout. It was similar to `endall` applied to `bays`, the best approach involving a probability combination method.

Table 4.1: Average AUC for the different methods shown as the median over all 100 test set results after 20 stratified 5-fold CV runs (IQR between brackets).

PUL	`bays`	`bay`	`plt`	`iso`	AR	Other methods	
`endall`	.966 (.015)	.959 (.024)	.961 (.020)	.961 (.024)	1.5	`sm1va`	.963 (.024)
`gbt1va`	.966 (.016)	.960 (.025)	.960 (.023)	.960 (.025)	2.1	`sm1v1`	.960 (.025)
`pc-wu1`	.964 (.015)	.946 (.038)	.960 (.023)	.959 (.027)	4.6	`mklr`	.964 (.018)
`pc-wu2`	.964 (.015)	.946 (.040)	.960 (.024)	.958 (.026)	5.1		
`pc-ht`	.964 (.014)	.945 (.037)	.959 (.024)	.959 (.025)	5.9		
`1van`	.963 (.020)	.952 (.032)	.958 (.026)	.958 (.030)	6.0		
`endone`	.962 (.019)	.948 (.035)	.960 (.027)	.953 (.031)	6.3		
`rvall`	.966 (.013)	.946 (.037)	.954 (.031)	.951 (.030)	6.5		
`pkpd`	.964 (.016)	.945 (.053)	.958 (.025)	.956 (.028)	7.1		
`rvone`	.962 (.016)	.929 (.048)	.948 (.031)	.942 (.037)	9.9		
AR	1.0	3.9	2.3	2.9			
IOTA	`bays`	`bay`	`plt`	`iso`	AR	Other methods	
`endall`	.881 (.029)	.859 (.037)	.869 (.027)	.868 (.029)	1.0	`sm1va`	.862 (.032)
`pc-wu1`	.878 (.029)	.856 (.024)	.862 (.031)	.858 (.031)	2.5	`sm1v1`	.861 (.034)
`pc-wu2`	.878 (.027)	.853 (.026)	.857 (.031)	.854 (.032)	4.1	`mklr`	.883 (.027)
`rvall`	.877 (.029)	.857 (.026)	.857 (.034)	.851 (.032)	4.6		
`gbt1va`	.878 (.028)	.761 (.067)	.858 (.035)	.855 (.032)	5.0		
`1van`	.875 (.029)	.840 (.035)	.859 (.039)	.847 (.034)	5.8		
`pc-ht`	.865 (.039)	.852 (.037)	.850 (.048)	.853 (.026)	6.5		
`pkpd`	.874 (.029)	.847 (.030)	.849 (.036)	.845 (.033)	7.3		
`endone`	.873 (.032)	.835 (.038)	.837 (.055)	.831 (.031)	8.8		
`rvone`	.858 (.046)	.836 (.045)	.798 (.107)	.790 (.072)	9.5		
AR	1.0	3.4	2.4	3.3			

The AR results for the four algorithms `bays`, `bay`, `plt`, and `iso` reveal that a Bayesian LS-SVM with sparseness was better than one without, that `bays` was better than `plt` and `iso`, and that `bay` is worse than `plt` and `iso`. Differences between `plt` and `iso` were typically in favor of `plt`. Differences in performance between probability combination methods were often smaller when `bays` or `plt` were used.

Table 4.2: Entropy for the different methods shown as the median over all 100 test set results after 20 stratified 5-fold CV runs (IQR between brackets).

PUL	bays	bay	plt	iso	AR	Other methods	
endall	.298 (.104)	.378 (.169)	.286 (.056)	.288 (.095)	2.0	sm1va	.273 (.070)
gbt1va	.307 (.115)	.402 (.170)	.287 (.062)	.310 (.154)	3.5	sm1v1	.288 (.071)
pc-wu1	.289 (.098)	.421 (.229)	.293 (.056)	.302 (.107)	3.9	mklr	.276 (.063)
pc-ht	.291 (.092)	.430 (.225)	.292 (.058)	.299 (.085)	3.9		
pc-wu2	.293 (.100)	.427 (.260)	.292 (.054)	.312 (.115)	4.6		
rvall	.288 (.097)	.404 (.204)	.302 (.065)	.318 (.122)	4.8		
endone	.320 (.107)	.469 (.295)	.291 (.065)	.356 (.187)	7.0		
1van	.310 (.118)	.441 (.216)	.298 (.062)	.323 (.156)	7.3		
pkpd	.324 (.123)	.508 (.337)	.293 (.060)	.377 (.206)	8.6		
rvone	.313 (.129)	.561 (.419)	.308 (.067)	.412 (.224)	9.5		
AR	1.8	4.0	1.3	2.9			
IOTA	bays	bay	plt	iso	AR	Other methods	
endall	.495 (.063)	.653 (.099)	.512 (.039)	.497 (.048)	1.0	sm1va	.504 (.052)
pc-wu2	.503 (.069)	.847 (.159)	.527 (.039)	.510 (.053)	2.8	sm1v1	.510 (.054)
pc-wu1	.502 (.065)	.834 (.149)	.536 (.035)	.512 (.047)	3.3	mklr	.492 (.055)
gbt1va	.546 (.117)	.877 (.239)	.517 (.046)	.542 (.110)	5.5		
pc-ht	.530 (.074)	.873 (.134)	.545 (.045)	.518 (.042)	5.8		
rvall	.506 (.066)	.874 (.144)	.559 (.038)	.529 (.054)	5.8		
endone	.526 (.101)	1.18 (.530)	.543 (.061)	.576 (.115)	6.8		
pkpd	.526 (.082)	1.20 (.217)	.534 (.059)	.603 (.176)	7.0		
1van	.563 (.146)	1.26 (.239)	.531 (.044)	.553 (.116)	7.3		
rvone	.568 (.127)	1.41 (.664)	.612 (.182)	.764 (.331)	10.0		
AR	1.5	4.0	2.3	2.2			

Overall, on the present data sets the best methods are `endall` and `mklr`. While the latter is an all-at-once method, the former is an ensemble method for which many binary models have to be fit (e.g. 25 when $K = 4$). This makes `mklr` more attractive. When using 1-versus-1 classifiers as the basis for obtaining multi-class probabilities, `pc-wu1` (preferably using `bays`) or `sm1v1` are good methods. When 1-versus-All classifiers are used, `sm1va` or `gbt1va` (preferably using `bays`) can be used. Out of the four implemented binary classification algorithms, `bays` seems a good choice to obtain probabilistic outputs. Using isotonic regression instead of Platt's method to obtain probabilities did not result in a clear benefit.

An important drawback of this comparison study is that only two data sets were used. More data sets would allow stronger conclusions, but, given the vast number of methods compared, would also risk an overload of information. Another point of critique is that we simply averaged the class-specific AUC values. Due to the interest in the class-specific results, it was impossible to construct a multi-class AUC measure a posteriori.

4.2 Weighted multi-class AUC metrics

4.2.1 Introduction

Performance evaluation for classifiers is crucial. An important aspect of this evaluation is the model's discriminative potential. Good binary classifiers are able to separate cases from both classes well. For binary classifiers with a numerical score output rather than a black-and-white class prediction, a useful metric is the AUC [60, 250, 182, 331]. In many research domains, the AUC is well-known as a tool to evaluate performance of classifiers. An ROC curve is constructed by plotting sensitivity (true positive rate) versus 1 minus specificity (false positive rate) by varying the decision threshold over its whole range. For a given threshold, class 1 (C_1) is predicted if the score is larger than or equal to the threshold, and class 2 (C_2) is predicted otherwise. The AUC summarizes model performance over all possible thresholds. See section 2.3.1 for information on ROC curves and the AUC.

Apart from its interpretation as the probability to correctly identify the C_1 case when confronted with a randomly selected case from each class [182], the nonparametric AUC has the advantages to have no underlying assumptions regarding the distribution of the scores in both classes, and to be insensitive to varying class distributions. The AUC is superior to the often used misclassification rate for the evaluation of model performance [195, 332]. Moreover, the misclassification rate requires the use of a single threshold, which is often not desirable. Other performance measures such as the cross-entropy or Brier score, for example, evaluate the class probabilities by measuring how well they approach the true class. However, they do not evaluate discriminatory ability. A drawback of the AUC is that, despite its representation of the degree of overlap between score distributions for both classes, it largely ignores the actual scores. The scores are merely used to rank the cases. Nevertheless, these scores may convey useful information. Let us mention an example given in [471], using a sample consisting of 3 C_1 and 3 C_2 cases. Assume that model A gives scores 1.0, 0.7, and 0.6 for the C_1 cases, and 0.5, 0.4, and 0.0 for the C_2 cases. Model B gives 1.0, 0.9, 0.5, and 0.6, 0.2, 0.0, respectively. Even though model A perfectly separated the cases from both classes, it is pointed out that model B may still be preferred as its ranking appears less sensitive to variation in or perturbation of the scores [471]. To our knowledge, there are two AUC metrics that incorporate actual scores [154, 470]. In the present work, we suggest extensions of such weighted metrics to multi-class problems to go beyond traditional multi-class AUC extensions [290, 181].

Hereafter, we will assume that the model outputs the predicted probability for C_1 (binary classification), or the predicted probabilities for all K classes (multi-class classification). Methods exist to transform other scores (such as those obtained with support vector machines) to probability estimates [260, 477, 305].

4.2.2 Weighted AUC metrics

Let $\hat{\pi}_{1n}$ and $\hat{\pi}_{2n}$ represent the probability of C_1 and C_2 for members of C_1, $n = 1, \ldots, N_1$. Let $\hat{\pi}_{1m}$ and $\hat{\pi}_{2m}$ represent the probability of C_1 and C_2 for members of C_2, $m = 1, \ldots, N_2$.

Probabilistic AUC (prAUC)

Building upon the above-mentioned interpretation of the AUC, a probabilistic AUC is defined in [154] as an estimate of the score difference between a randomly selected (C_1, C_2)-pair of cases (prAUC). The unbiased estimator they use is the difference between the average probability of C_1 for C_1 cases and C_2 cases. To obtain a metric on the same scale as the AUC, this difference is linearly transformed:

$$\text{prAUC} = 0.5 + \frac{1}{2}\left(\frac{\sum_{n=1}^{N_1}\hat{\pi}_{1n}}{N_1} - \frac{\sum_{m=1}^{N_2}\hat{\pi}_{1m}}{N_2}\right). \tag{4.12}$$

Reformulating this expression to render it consistent with the expression of the standard AUC, one obtains

$$\text{prAUC} = 0.5 + \frac{1}{2N_1N_2}\sum_{n=1}^{N_1}\sum_{m=1}^{N_2} d(\hat{\pi}_{1n}, \hat{\pi}_{1m}), \tag{4.13}$$

where $d(\hat{\pi}_{1n}, \hat{\pi}_{1m}) = \hat{\pi}_{1n} - \hat{\pi}_{1m}$ [154].

Next, a probabilistic ROC curve (prROC) was derived for which the area under the curve equals prAUC. The prROC curve is obtained by constructing intervals of width d around the estimated probability of C_1, based on the idea that this probability is only approximately equal to the true probability. For a given threshold, the true positive rate is computed by checking for each C_1 case what part of the interval d around $\hat{\pi}_{1n}$ exceeds the cut-off. The sum over all N_1 cases divided by dN_1 yields the true positive rate. The false positive rate is computed analogously. The interval length d can be found numerically by running through a grid of possible values until the value is found for which the area under the prROC curve equals prAUC. Note that the solution may not be unique. In this case, the smallest d satisfying area(prROC)=prAUC is chosen. The resulting prROC curve is a smoothed curve that evolves towards the original ROC curve for increasing sample size.

Scored AUC (sAUC)

Another extension of the original AUC is the scored AUC [470] (sAUC). It is expressed as a weighted version of the original AUC, in which correctly classified pairs are weighted by $\hat{\pi}_{1n} - \hat{\pi}_{1m}$:

$$\text{sAUC} = \frac{1}{N_1 N_2} \sum_{n=1}^{N_1} \sum_{m=1}^{N_2} (\hat{\pi}_{1n} - \hat{\pi}_{1m}) U_{n_1,n_2}, \tag{4.14}$$

with U_{n_1,n_2} defined as in section 2.3.1. Equivalently, sAUC can be written as

$$\text{sAUC} = \frac{1}{N_1 N_2} \left(\sum_{m=1}^{N_2} \sum_{\hat{\pi}_{1n} > \hat{\pi}_{1m}} \hat{\pi}_{1n} - \sum_{n=1}^{N_1} \sum_{\hat{\pi}_{1m} < \hat{\pi}_{1n}} \hat{\pi}_{1m} \right). \tag{4.15}$$

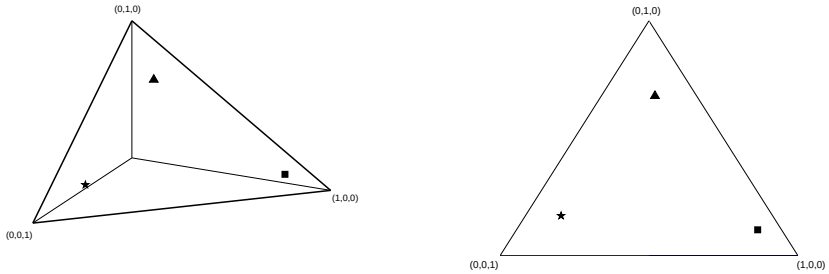
Recently, the sAUC was interpreted as the area under the scored ROC (sROC) curve [471]. The sROC curve plots a margin-based AUC measure versus τ, where τ is a number between 0 and 1 that is subtracted from the probability of C_1 cases. The sAUC reflects the stability of the standard AUC with increasing margin τ [471].

4.2.3 Weighted multi-class AUC metrics

In K-class classification, the model outputs K estimated class probabilities $\hat{\pi}_k$ for each data point, with $0 \leq \hat{\pi}_k \leq 1$, $\sum_{k=1}^{K} \hat{\pi}_k = 1$. A data point can thus be represented in K-dimensional space using its estimated class probability vector $\mathbf{p} = (\hat{\pi}_1, \ldots, \hat{\pi}_K)$. In fact, a $K-1$-dimensional space would suffice because the K estimated class probabilities sum to one. For example, in a 3-class problem, a data point i with vector $(\hat{\pi}_{1i}, \hat{\pi}_{2i}, \hat{\pi}_{3i})$ can be represented in 3-dimensional space (Figure 4.3(a)). Because the class probabilities sum to one, all data points are situated on a 2-dimensional subplane defined by the coordinates $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ (i.e. the class corners). Further, because probabilities lie between 0 and 1, all data points are located on the equilateral triangle defined by the class corners (Figure 4.3(b)). Alternatively, the estimated probabilities for one of the classes can be ignored, and the data points $(\hat{\pi}_{1i}, \hat{\pi}_{2i})$ can be plotted in two-dimensional space (as a projection of the three-dimensional space). This time, all points lie within an isosceles right triangle defined by the origin, $(1, 0)$, and $(0, 1)$ (Figure 4.3(c)).

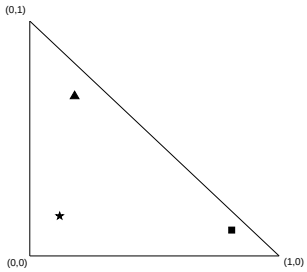
Standard multi-class AUC metrics

Mossman [290, 135] has extended the ROC curve to three-class problems. The three-class extension of the ROC curve plots the true positive rates for the three classes (TPR triple). By varying the decision function to map the probabilistic output $(\hat{\pi}_{1i}, \hat{\pi}_{2i}, \hat{\pi}_{3i})$ into a crisp class prediction, many TPR triples are obtained. Plotting these triples in three-dimensional space results in a ROC surface. The volume under the surface (VUS) can be interpreted as the probability that a randomly selected (C_1, C_2, C_3)-triplet of cases is correctly classified by the model. It is not straightforward, however, to determine what it means to correctly classify a triplet. One rule that is also used in [135] states that a triplet is correctly rated if the total length of the lines connecting each case with its true class corner in the equilateral triangle is smaller

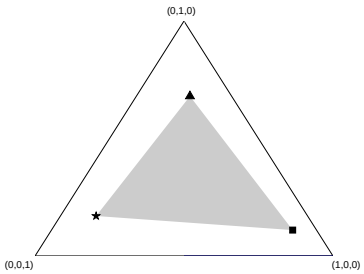


(a)

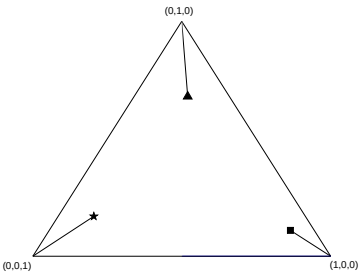
(b)



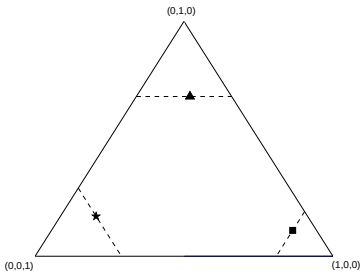
(c)



(d)



(e)



(f)

Figure 4.3: Representation of predicted probabilities for a three-class problem. The square refers to a C_1 case, the triangle to a C_2 case, and the star to a C_3 case.

than any other total length obtained by connecting the three cases to the three class corners. Let N_k denote the number of cases in class k, and let $\mathbf{p}_n$, $\mathbf{p}_m$, and $\mathbf{p}_q$ represent the class probability vectors for cases from C_1, C_2, and C_3, respectively. Then, the VUS is estimated as

$$\text{VUS} = \frac{1}{N_1 N_2 N_3} \sum_{n=1}^{N_1} \sum_{m=1}^{N_2} \sum_{q=1}^{N_3} C(\mathbf{p}_n, \mathbf{p}_m, \mathbf{p}_q), \tag{4.16}$$

where

$$C(\mathbf{p}_n, \mathbf{p}_m, \mathbf{p}_q) = \begin{cases} 1 & \text{if correctly classified} \\ 0 & \text{otherwise} \end{cases} . \tag{4.17}$$

This approach to the computation of the three-class VUS can be used for any K-class problem. The rule that is used to define correct classification is rather technical, however, and in our opinion not always optimal. Another rule is suggested further on.

Hand and Till [181] have suggested a simple extension of binary AUC values to K-class AUCs for any value of K. They compute AUC_{kr}, the AUC for each pair of classes k and r ($k \neq r$) using $\hat{\pi}_k$. Note that $\text{AUC}_{kr} \neq \text{AUC}_{rk}$, except for binary problems. Their index M is computed as

$$M = \frac{1}{K(K-1)} \sum_{k \neq r} AUC_{kr}. \tag{4.18}$$

Provost and Domingos [330] applied a similar metric based on the K 1-versus-all AUCs (AUCs for each class versus all other classes). Their index equals the average of the 1-versus-all AUCs weighted by the prevalence of a class in the data set.

Area of the triangle

An attractive measure for three-class problems, starts with the computation of the average class probability vector for each class k, $\mathbf{p}_k$:

$$\mathbf{p}_k = \frac{1}{N_k} \sum_{n_k=1}^{N_k} \mathbf{p}_{n_k}, \ k = 1, \ 2, \ 3. \tag{4.19}$$

The average probability vectors for each class can be plotted in three-dimensional space, and the area of the triangle spanned by the three points is computed and rescaled to yield a number between 0 and 1 (AOT). Since the three class

probabilities sum to one, this triangle is located within the equilateral triangle mentioned previously (see Figure 4.3(d)). The more a model tears the classes apart in terms of the predicted probabilities, the larger AOT. The area cannot be seen as the average area over all possible combinations of three data points from the three classes, since the volume does not contain information about the exact location of the data points in three-dimensional space. A combination in which the probability vector of the C_k case lies close to the C_r class corner and vice versa will yield a similar area compared to a similar combination where the C_k and C_r cases lie close to their own class corner. Thus, a high AOT is obtained when the model tears the classes apart, regardless of the class corner to which each average probability vector is pulled. In practice, however, useful models will on average pull classes toward the correct class corner.

Note that the area of the triangle is proportional to the area of the triangle when projecting it on to any possible two-dimensional space defined by two of the three classes (as in Figure 4.3(c)). It does not matter which projection is chosen. The maximum area of the original triangle within the bounds of the equilateral triangle is $\frac{\sqrt{3}}{2}$, so the AOT is obtained by dividing the area of the triangle by $\frac{\sqrt{3}}{2}$.

Average total length between probability coordinates and class corners

The abovementioned rule to determine whether K randomly chosen cases from the K classes are correctly classified, can also be exploited to incorporate probabilities. A possibility is to compute the total length of the lines connecting the K average probability vectors with their class corners (TL_0, cf. Figure 4.3(e)). If a classifier is perfect, TL_0 is 0. If we want the index to vary between 0 to 1 (with 1 indicating a perfect classifier), we could transform TL_0 to obtain $\text{TL} = 1 - \frac{\text{TL}_0}{z}$, with z as the maximum of TL_0. The maximum of TL_0 for a K-class problem is $K\sqrt{2}$, and is obtained when all $\mathbf{p}_k$ are located exactly on a wrong class corner.

Weighted VUS metrics

In the sense of sAUC, we could define a weighted version of the VUS (wVUS). For a three-class problem, this metric weights correctly classified triplets using the total length of the lines connecting each case with its class corner. Let l_n, l_m, and l_q represent the length of the lines for cases from C_1, C_2, and C_3, respectively. The wVUS can be written as

$$\text{wVUS} = \frac{1}{N_1 N_2 N_3} \sum_{n=1}^{N_1} \sum_{m=1}^{N_2} \sum_{q=1}^{N_3} W_{nmq} C(\mathbf{p}_n, \mathbf{p}_m, \mathbf{p}_q), \tag{4.20}$$

where

$$W_{nmq} = \left(1 - \frac{l_n + l_m + l_q}{z}\right). \tag{4.21}$$

Using the idea of the AOT, we suggest another rule to define correct classification when confronted with one case from each class. This results in an alternative multi-class AUC index (VUS_2), for which a weighted version can be derived. The easier to understand rule for correct classification is that the highest probability for C_k should be obtained for the C_k case. It is natural to think that the case with highest $\hat{\pi}_k$ will probably be the C_k case. Using the suggested rule, it is easy to see whether the classifier works well when given one case from each class. Figure 4.3(f) illustrates the new rule for a three-class problem. The dashed lines through each case represent the line of equal probability for the true class of that case. If any case crosses the line of another case, the combination of cases is not correctly classified. For a three-class problem, a weighted version (wVUS_2) weights a correctly classified set of cases with the AOT. This is justified, since correct classification now means that each probability vector is closest to its own class corner such that no two cases are switched.

It is always true that $\text{VUS}_2 \leq \text{VUS}$. If the newly suggested rule applies, each probability vector is closest to its own class corner. Thus, the sum of lines connecting the cases with their class corner is always minimal. As a consequence, there cannot be situations where the VUS_2 rule applies but the VUS rule does not.

Combination of M and weighted AUC metrics

A simple generalization of binary weighted AUC metrics [154, 470] for any K-class problem could be to combine them in the sense of Hand and Till's M index [181]:

$$M_p = \frac{1}{K(K-1)} \sum_{k \neq r} \text{prAUC}_{kr}, \tag{4.22}$$

or

$$M_s = \frac{1}{K(K-1)} \sum_{k \neq r} \text{sAUC}_{kr}. \tag{4.23}$$

4.2.4 A real world application

Experimental setup

We will now demonstrate the indices to the PUL data. The PUL data set consists of 856 pregnancies collected at St. George's Hospital (London, UK), for which 12

measurements were recorded.

We performed input selection and used six algorithms to derive a multi-class model for PULs (see Chapter 8 for a detailed description of the multi-class models): a multi-class logistic regression model (MLR; with checks for linearity in the logit and interactions), a Bayesian multi-layer perceptron (BMLP), a Bayesian perceptron (BPER; similar to a regularized MLR, but then without the checks mentioned for the MLR model), 1-versus-1 Bayesian least squares support vector machines using a linear or an RBF kernel (BLSSVMlin and BLSSVMrbf; the binary models are combined using pairwise coupling to yield multi-class probabilities), and a multi-class kernel logistic regression algorithm using an RBF kernel (MKLR). To compare the algorithms, they were applied to 100 random stratified splits of the data into a training (60%) and a test set (40%). Each algorithm was trained on the training set and was evaluated on the test set using the weighted and unweighted metrics described above. We summarized the 100 test set values for each metric by taking the mean. The performance of the six algorithms is compared by computing their average rank over the 100 test set performances (AR), and by computing the proportion of times they were ranked on top ($Pm(1)$, along the ideas of [321] as explained in section 2.3.3). More important at this moment, however, was our aim to compute correlations between the various metrics using the 100 test set results. This allowed us to check how well different metrics are interrelated. It is interesting to investigate the associations between weighted and unweighted metrics, and between different types of weighted metrics.

Next, we used the various metrics to evaluate the importance of different measurements in predicting the PUL outcome. Currently, hCG information is considered most important. But, it is interesting for clinicians to know how important it is to add progesterone information as an input to the model, and how important it is to also add clinical and demographic information as inputs. Therefore, we applied the MKLR algorithm using three different input sets: a) hCG ratio and hCG average; b) all hormonal information, being both hCG measures and progesterone average; and c) all inputs that were selected using input selection, being the three hormonal measures, the level of vaginal bleeding, age of the mother, and history of ectopic pregnancy. The three models (referred to as the hCG, hormonal, and full model) were evaluated using their performance on the 100 test sets.

Results

The correlations between metrics were computed for each algorithm, and were averaged. Table 4.3 shows the average Spearman correlations between the multi-class AUC metrics. The Pearson correlations were nearly identical. The unweighted metrics had intercorrelations between 0.90 and 0.97, with the highest correlation between the M index and VUS_2. Except for wVUS, the weighted metrics were also highly related, with intercorrelations between 0.92 and 0.99. The intercorrelations between the weighted (except wVUS) and unweighted metrics varied between

0.41 and 0.65, indicating that incorporating the predicted probabilities can give a different picture when performing model selection or evaluation. The correlations obtained for wVUS deserve special attention. This weighted metric correlated between 0.74 and 0.82 with the other weighted metrics, but correlated between 0.88 and 0.89 with the unweighted metrics. The problem with wVUS may be either the rule it uses to determine correct classification, or the weights it uses. Extra analyses showed that another measure, in which correct classification was determined according to (w)VUS_2 but weighting was according to wVUS, that uses the rule of $wVUS_2$ to determine correct classification but weights the correctly classified cases using the weights from wVUS, behaved like wVUS with respect to its correlations with other metrics. Thus, the problem of wVUS appears to be the weights it uses.

Table 4.3: Spearman correlation matrix

	M	VUS	VUS_2	AOT	TL	M_p	M_s	wVUS
M	1							
VUS	.94	1						
VUS_2	.97	.90	1					
AOT	.53	.48	.60	1				
TL	.56	.50	.65	.93	1			
M_p	.56	.51	.65	.94	.99	1		
M_s	.49	.44	.58	.95	.98	.99	1	
wVUS	.88	.88	.89	.79	.81	.82	.77	1
$wVUS_2$	.47	.41	.54	.99	.92	.93	.95	.74

We compared wVUS and $wVUS_2$ using a few real examples. We chose one of the 100 runs for the MKLR model. For all $N_1N_2N_3 = 184 * 132 * 26 = 631488$ triplets in the test set, we determined whether they were correctly classified or not according to both metrics, and computed the weight that would be assigned if correctly classified. Combinations were correctly classified for both metrics in 77.9% of the occasions, and both types of weights were correlated 0.90 (Figure 4.4(a) presents a scatter plot for 1% of these combinations). Combinations were correctly classified only for wVUS in 9.5% of the occasions, and the correlation between weight types was only 0.25.

Before checking out individual combinations of K cases, one of the best triplets is shown in Figure 4.4(b). The instances from class 1 and 2 had near perfect predicted probability for the true class and near zero probability for the other classes. Given the low prevalence of class 3, the predicted class probability for the C_3 case was very good as well. Figure 4.4(c) presents two triplets that were correctly classified for wVUS only, with one triplet (black triplet) having markedly better weight for wVUS than the other (white triplet). Even though the rule for correct classification does

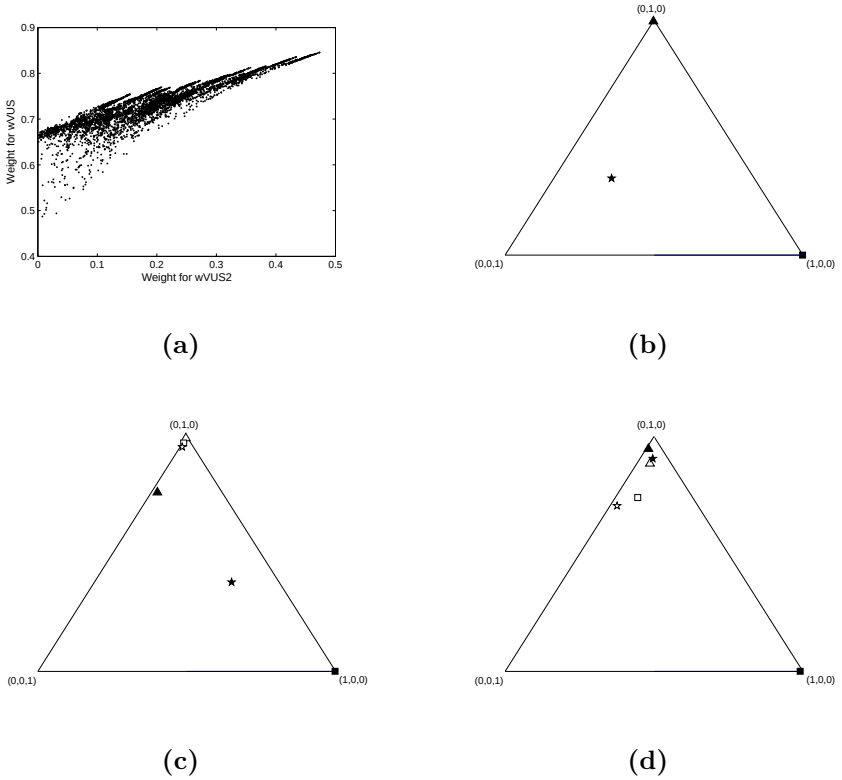


Figure 4.4: Comparison of wVUS and $wVUS_2$. The square refers to a C_1 case, the triangle to a C_2 case, and the star to a C_3 case.

not seem to be problematic as such, both triplets show that the limit for correct classification is less strict for wVUS (or VUS) than for $wVUS_2$ (or VUS_2). The black triplet does not look that bad, but there is no clear agreement as to what case should be the C_3 case. Actually, the C_1 case has a higher probability to belong to C_3 than the C_3 case. The white triplet is not good at all.

Figure 4.4(a) shows that triplets with high weight for $wVUS_2$ also have a high weight for wVUS. When $wVUS_2$ gives a low weight, however, the weight for wVUS can vary. Figure 4.4(d) shows two triplets that are correctly classified using both rules, but for which $wVUS_2$ gives low weight while wVUS gives clearly lower weight for one triplet (white) relative to the other (black). To us, $wVUS_2$ appears more correct because the three cases are not clearly separated from each other in both triplets.

Table 4.4 presents the results for five metrics, two unweighted (M and VUS_2) and three weighted ones (AOT, M_p, and $wVUS_2$). The unweighted metrics suggest that MLR is the best algorithm, followed by MKLR and BLSSVMrbf. The probability that MLR is the best of the algorithms is estimated to be 0.55. The multi-layer

and single-layer perceptron models perform worst. When looking at the weighted metrics, however, we observe a stronger separation between algorithms that perform well and those that perform poor. This time, both MLR and BLSSVMrbf perform best, while MKLR is less outstanding. Apparently, BLSSVMrbf produced better probabilities than MKLR. Finally, note that the AOT is rather small. This reflects the low prevalence of ectopic pregnancies.

Table 4.4: Average test set performance of the six algorithms: the average performance, and AR and $Pm(1)$ (small font).

	M	VUS_2	AOT	M_p	$wVUS_2$
MLR	.954	.829	.288	.801	.296
	1.9 − .55	1.8 − .65	1.5 − .57	1.6 − .45	1.5 − .55
BMLP	.943	.779	.123	.706	.125
	4.2 − .08	5.0 − .01	5.7 − .00	6.0 − .00	5.7 − .00
BPER	.941	.781	.137	.744	.140
	5.0 − .00	5.2 − .00	5.3 − .00	5.0 − .00	5.3 − .00
BLSSVMlin	.946	.808	.233	.787	.241
	4.1 − .02	3.5 − .07	3.4 − .00	3.2 − .00	3.4 − .00
BLSSVMrbf	.948	.815	.280	.802	.290
	3.5 − .07	2.9 − .16	1.7 − .43	1.5 − .55	1.6 − .45
MKLR	.952	.818	.227	.781	.233
	2.4 − .29	2.6 − .11	3.5 − .00	3.7 − .00	3.5 − .00

When using AUC metrics to decide whether hCG information is sufficient or extra information is indispensable, unweighted and weighted metrics again tell a different story. Unweighted metrics suggest that progesterone provides useful additional information, but that even more variables (i.e. age, vaginal bleedings, and history of ectopic pregnancy) may well be added. Weighted metrics, however, suggest that progesterone information is indeed necessary but that it is not crucial to add more variables. This is shown in Table 4.5 and Figure 4.5. The figure shows the performance of each metric for all three input sets, where the performance values are transformed such that they are zero when only hCG information is used and one if the full model is used. The bold lines, representing weighted metrics, put the results for the hormonal model much closer to 1 than the unweighted metrics (thin lines). From a clinical point of view, it could of course be argued that adding age, level of vaginal bleeding, and history of ectopic pregnancy should be added since these variable are very cheap to measure.

4.2.5 Discussion and conclusions

Inspired by the work on weighted AUC metrics for binary classification, weighted AUC metrics for the multi-class case were suggested. These metrics are based on the average predicted class probabilities for cases belonging to the same class (AOT, TL), on weighting only correctly classified combinations of cases from each class (wVUS,

Table 4.5: Average test set performance of MKLR models with different input sets

	M	VUS_2	AOT	M_p	wVUS_2
hCG	.930	.755	.179	.765	.187
	$3.0 - .00$	$2.9 - .00$	$2.9 - .01$	$2.9 - .01$	$2.9 - .01$
Hormonal	.947	.806	.221	.781	.229
	$1.8 - .25$	$1.8 - .27$	$1.6 - .40$	$1.6 - .44$	$1.6 - .42$
Full	.952	.818	.227	.781	.233
	$1.3 - .75$	$1.3 - .73$	$1.5 - .59$	$1.5 - .55$	$1.5 - .57$

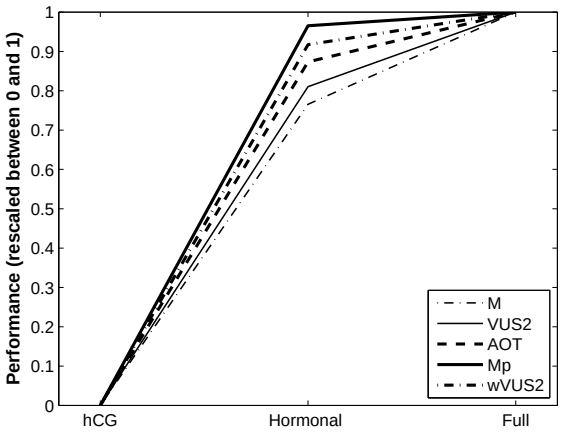


Figure 4.5: Performance of MKLR algorithms with different input sets. The performance results are re-scaled such that the hCG model has performance 0 and the full model had performance 1.

wVUS_2), or on the simple combination of binary weighted AUC metrics (M_p, M_s). We presented correlations between these metrics, obtained from the application of six algorithms to 100 random train-test splits of a data set on pregnancies of unknown location. The correlations suggest that unweighted metrics are highly correlated, and that all but one weighted metrics are highly correlated. The wVUS metric appeared to correlate less strong with other weighted metrics than with unweighted metrics. Unweighted and weighted metrics are moderately correlated. Since metrics such as TL, M_p and M_s are easy to compute, they are attractive for problems where $K >> 2$. Metrics such as AOT or wVUS_2 are interesting and visually attractive when $K = 3$.

We demonstrated the metrics by applying them for the sake of model comparison, and observed that both types of metrics did not favor the same set of models. Regarding model comparison, it is interesting to check for differences in results based on unweighted versus weighted metrics. Unweighted metrics can be considered most important, but weighted metrics are interesting if small difference are observed for

unweighted metrics. We acknowledge that another highly useful application area for weighted multi-class AUC metrics is that of model selection. An experimental study on this issue would form an important topic for further research. At this moment, we can point at experimental studies where the use of the sAUC for model selection in binary classification problems yielded promising results [471, 470]. Increasing attention is given to methods that construct classifiers that try to maximize AUC rather than minimize some error function. Weighted AUC metrics can build upon this.

Chapter 5

Diagnostic models for ovarian tumors

Abstract – This chapter presents the models that are developed to predict the nature of ovarian tumors prior to surgery. Both binary models to distinguish benign from malignant tumors, and multi-class models to distinguish between benign, primary invasive, borderline malignant, and metastatic invasive tumors are presented. The models are based on logistic regression, multi-layer perceptrons, least squares support vector machines, relevance vector machines, and kernel logistic regression. The models are developed on the training part of the IOTA phase I data set and have undergone a preliminary performance evaluation on the test part of the same data set.

5.1 Introduction

As discussed in Chapter 3, a good preoperative judgment of an ovarian mass is important because it helps to optimize the treatment of the tumor. There exists evidence that optimal treatment has beneficial consequences for the patient, for example in terms of improved prognosis. Mathematical models can be used to provide an automated diagnosis. Such models can help to clarify which tumor and patient characteristics are related to the tumor type, and they may assist humans in their judgment and treatment decisions when confronted with an ovarian mass. The performance of mathematical models should be evaluated, and the performance should be compared with the performance of expert operators based on ultrasound examination of the tumor. It is possible that models outperform ultrasound operators irrespective of their experience, or that they only outperform operators with less experience. That would make these models a promising tool for clinical practice. We hope at least to equal the performance of highly experienced operators.

There are many approaches to link measurements with a categorical outcome. Some of these methods are more simple than others. Generally, there is no method

that will be universally superior, hence it is good to apply various methods and compare their performance for the problem at hand. This also allows to judge whether a simple method performs similar to more advanced ones. If so, the simpler method is preferred. For us, the basic method for classification is logistic regression analysis, using either unregularized [193] or regularized [382, 315, 480, 448] versions. More advanced models can be of many types, but the main one in this thesis is the least square support vector machine. Support vector machines are very flexible methods, and the least squares version is an elegant variant of standard SVMs. Details on these algorithms can be found in Chapter 2. All models are built using the data set from phase I of the IOTA study group. This data set is described in section 3.2.1.

5.2 Binary diagnostic models

5.2.1 Logistic regression models

Unregularized: LR1 and LR2

The first models for binary prediction of ovarian tumors were developed by Ameye [20, 394]. Because our contribution to this work was limited, we will describe them shortly.

Before building the model, relationships among the available measurements were investigated. Several variables were dropped in order to keep intercorrelations small and to avoid issues of stability when building the model. Input selection was done using stepwise selection, checks for the assumption of linearity in the logit, and checks for interaction effects [20]. The final model, referred to as LR1, contained 12 inputs: personal history of ovarian cancer (1; pershistovca), current use of hormonal therapy (2; hormtherapy), age (3), maximum diameter of the lesion (4; lesmax), pain during examination (5; pain), ascites (6), blood flow in the papillary projections (7; papflow), entirely solid tumor (8; locul5), maximum diameter of the largest solid component (9, bounded at 50 such that diameters > 50 mm are replaced with 50; solidmax2), irregular internal cyst walls (10; wallreg), acoustic shadows (11; shadows), and color score of intratumoral blood flow (12; colscore). The probability of malignancy for LR1 is computed as $\frac{1}{1+e^{-z}}$, where

$$\begin{aligned} z &= -6.7468 + 1.5985(1) - 0.9983(2) + 0.0326(3) + 0.00841(4) \\ &\quad -0.8577(5) + 1.5513(6) + 1.1737(7) + 0.9281(8) \\ &\quad +0.0496(9) + 1.1421(10) - 2.3550(11) + 0.4916(12). \end{aligned} \tag{5.1}$$

Thus, this model independently links pershistovca, age, lesmax, ascites, papflow, locul5, solidmax2, wallreg, and colscore with an increased risk of malignancy. Hormtherapy, pain, and shadows are independently linked with a decreased risk

of malignancy. See Table 5.1 for LR1's parameters estimates and related results.

Table 5.1: Statistical results for the 12 selected inputs for LR1.

Input	Parameter Estimate (SE)	Odds ratio (95% CI)	Standardized Par. Est.	LR Chi-squared	p-value (df=1)
Intercept	−6.7468 (.7419)				
Pershistovca	1.5985 (.8032)	4.95 (1.02 − 23.9)	.1057	4.01	0.0452
Hormtherapy	−0.9983 (.3539)	0.37 (0.18 − 0.74)	−.2330	8.54	0.0035
Age	0.0326 (.0094)	1.39 (1.15 − 1.67)	.2809	12.43	0.0004
Lesmax	0.0084 (.0028)	1.09 (1.03 − 1.15)	.2193	9.12	0.0025
Pain	−0.8577 (.3383)	0.42 (0.22 − 0.82)	−.2118	6.76	0.0093
Ascites	1.5513 (.4030)	4.72 (2.14 − 10.4)	.2853	16.07	< 0.0001
Papflow	1.1737 (.3499)	3.23 (1.63 − 6.42)	.2336	11.50	0.0007
Locul5	0.9281 (.4453)	2.53 (1.06 − 6.06)	.1707	4.41	0.0357
Solidmax2	0.0496 (.0090)	1.64 (1.38 − 1.96)	.5465	31.79	< 0.0001
Wallreg	1.1421 (.3317)	3.13 (1.64 − 6.00)	.3141	12.24	0.0005
Shadows	−2.3550 (.8090)	0.10 (0.02 − 0.46)	−.3889	13.06	0.0003
Colscore	0.4916 (.1699)	1.64 (1.17 − 2.28)	.2668	8.56	0.0034

The standardized regression coefficients shown in Table 5.1 suggest that the six most important inputs are solidmax2, shadows, wallreg, ascites, age, and colscore. A second, reduced model was also developed (LR2). This contained the six variables that were first entered in the model when stepwise selection was used. These are identical to the six with the strongest standardized regression coefficients except that colscore is replaced by papflow. The parameter estimates of LR2 are of course obtained by refitting the model using only the subset of six inputs:

$$\begin{aligned} z &= -5.3718 + 0.0354(3) + 1.6159(6) + 1.1768(7) \\ &\quad +0.0697(9) + 0.9586(10) - 2.9486(11). \end{aligned} \tag{5.2}$$

Aiming to obtain 90% sensitivity and 75% specificity on new data, a threshold of 0.10 was chosen for LR1 and LR2 to arrive at a crisp prediction. Tumors with a higher probability of malignancy are then predicted to be malignant, while all other tumors are predicted to be benign.

Regularized: using L_1-regularization

The use of regularized versions of the logistic regression algorithm has often been suggested in order to prevent overfitting. Such regularized versions are particularly helpful in situations where there are much more measurements than there are data cases, as in microarray or proteomics data. Nevertheless, it can be useful to apply such an algorithm here. We used a LASSO-based algorithm [315], combining regularization with input selection using L_1-regularization (see

section 2.2.1). It is interesting to investigate how such automated regularized procedures that incorporate input selection compare with unregularized approaches using manual or stepwise input selection.

The analysis was performed using the `glmpath` package for R (www.r-project.org). This package implements an algorithm that is able to determine the entire path of the parameter estimates over the regularization parameter λ [315]. Starting from a large value for λ (i.e. large amount of regularization), the algorithm fits a series of solutions with decreasing λ. By decreasing λ, inputs can enter or leave the model as in a traditional stepwise selection procedure, and regularized solutions can be obtained for any number of inputs in the model. We need to select a final solution by tuning λ. Because this is not a core part of this thesis, we decided to tune λ using the BIC rather than an extensive cross-validation analysis.

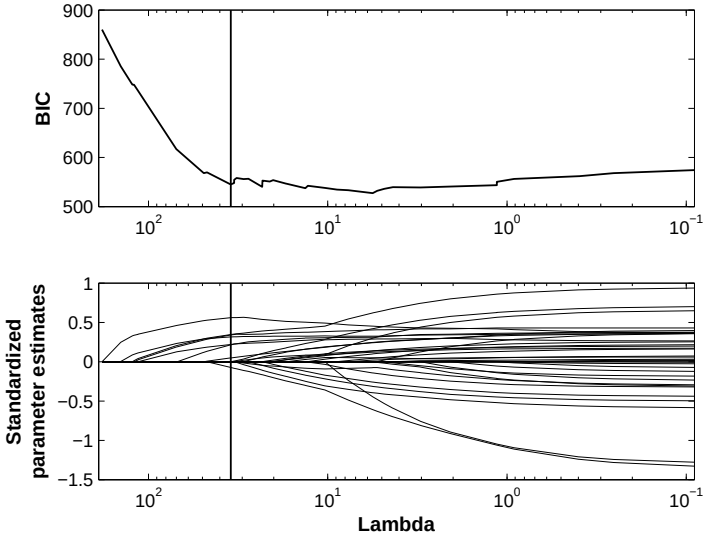


Figure 5.1: BIC and coefficient paths for the first run.

The variables that were fed to the algorithm were the 36 that survived the check for multicollinearity using variance inflation factors. This multicollinearity analysis is presented in detail in section 3.2.1. In Figure 5.1, the standardized parameter estimates of all 36 inputs (the coefficient paths) and the BIC are shown versus λ. The vertical reference line pinpoints the selected λ based on the BIC path. The model has selected the maximum diameter of the largest solid component (unbounded; solidmax), ascites, wallreg, colscore, papflow, age, amount of fluid in the Pouch of Douglas (fluid), and unilocular tumor (locul1). Then, we checked the assumption of linearity in the logit for the continuous and ordinal variables conditional on all other selected inputs. The linearity assumption was not always perfectly satisfied, but no extreme violations were detected. To be consistent with LR1, these checks resulted

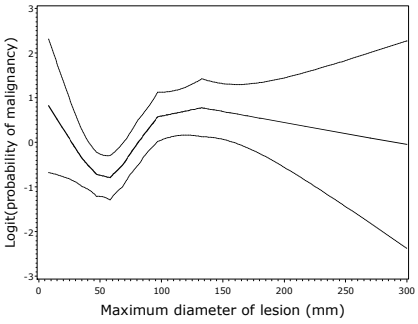


Figure 5.2: Effect of lesmax on the logit (bold line), conditional on the other selected inputs. The thin lines represent the 95% confidence band. Most maximum diameters were between 50 and 150.

in bounding the maximum diameter of the largest solid component at 50 mm (cf. solidmax2). Next, the L_1 regularization algorithm was run again, followed by checks for linearity in the logit. This time, the maximum diameter of the lesion (lesmax) was selected, but it was suggested to left-bound it at 60 mm and right-bound it at 170 mm (lesmax2) (Figure 5.2). The algorithm was run for the third time, followed by linearity checks. These did not suggest any additional transformations, such that the results of this third run of the algorithm are considered final. The BIC and coefficient paths are plotted in Figure 5.3. The plots to check for linearity in the logit are shown in Figure 5.4.

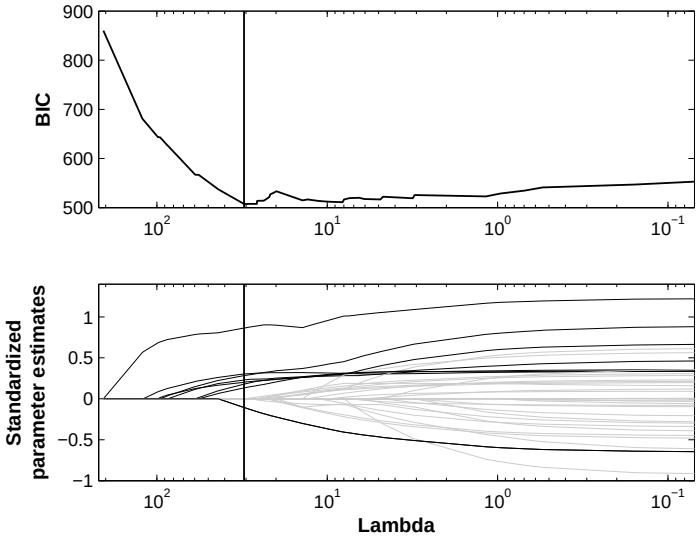


Figure 5.3: BIC and coefficient paths for the final run. The coefficient paths of non-selected inputs are shown in gray.

Using $\lambda = 30.78$, a classifier using eight inputs was selected (RegLR): solidmax2

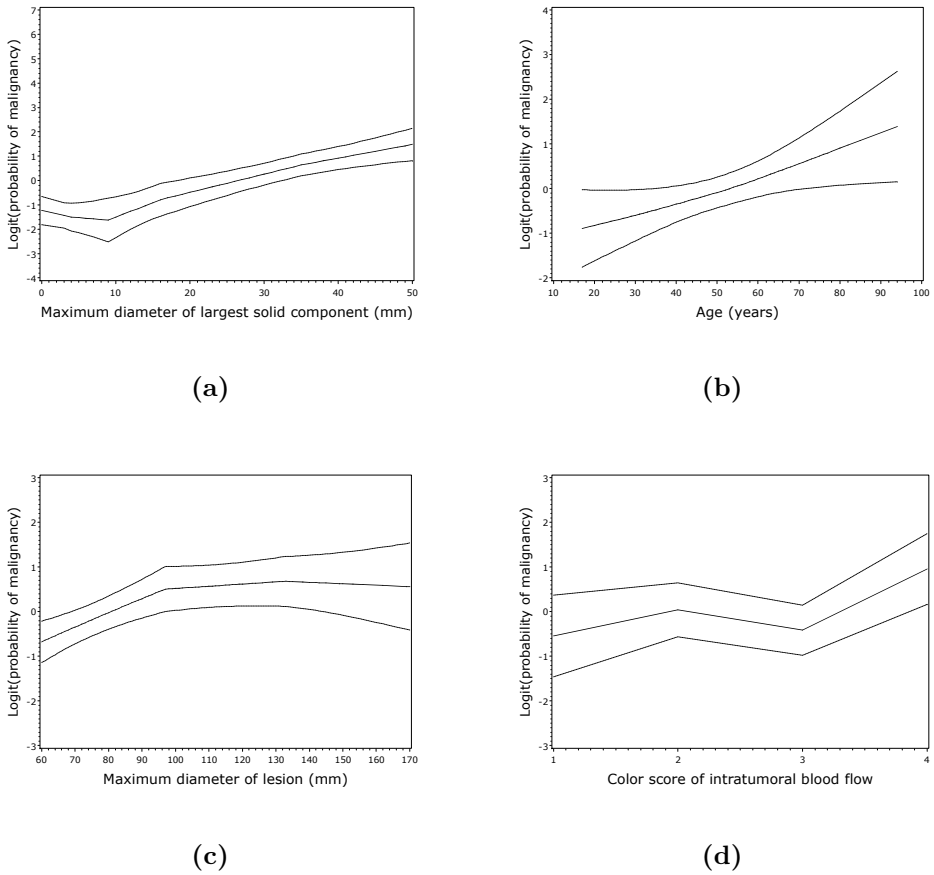


Figure 5.4: Checks for linearity in the logit for the continuous/ordinal inputs of the final L_1 regularized logistic regression model.

(1), ascites (2), wallreg (3), colscore (4), papflow (5), age (6), shadows (7), lesmax2 (8), where

$$\begin{aligned} z &= -4.0696 + 0.0433(1) + 0.9108(2) + 0.5663(3) + 0.2329(4) \\ &\quad +0.5597(5) + 0.0114(6) - 0.3540(7) + 0.0040(8). \end{aligned} \tag{5.3}$$

The parameter estimates and related results are given in Table 5.2. On the training set, a threshold probability of 0.20 to predict malignancy resulted in 91.6% sensitivity and 78.2% specificity (LR+ 4.19, LR- 0.11). Out of interest, a small bootstrap sensitivity analysis was performed by drawing 100 bootstrap samples from the training data. The algorithm was run on each sample, using BIC to select λ. For all 36 inputs, a histogram of the 100 parameter estimates can be made. The parameter estimate is zero if the input is not selected, and non-zero otherwise. The histograms are shown in Figure 5.5. The figure nicely shows that the eight inputs selected for RegLR (upper eight histograms) less often have a zero effect.

Table 5.2: Parameter estimates for the 8 inputs of the regularized logistic regression model.

Input	Parameter Estimate	Odds ratio	Unregularized estimate	Shrinkage
Intercept	−4.0696		−7.0937	
Age	0.0114	1.12	0.0321	64%
Lesmax2	0.0040	1.04	0.0119	66%
Ascites	0.9108	2.49	1.5591	42%
Papflow	0.5597	1.75	1.0030	44%
Solidmax2	0.0433	1.54	0.0593	27%
Wallreg	0.5663	1.76	1.0373	45%
Shadows	−0.3540	0.70	−2.5720	86%
Colscore	0.2329	1.26	0.4009	42%

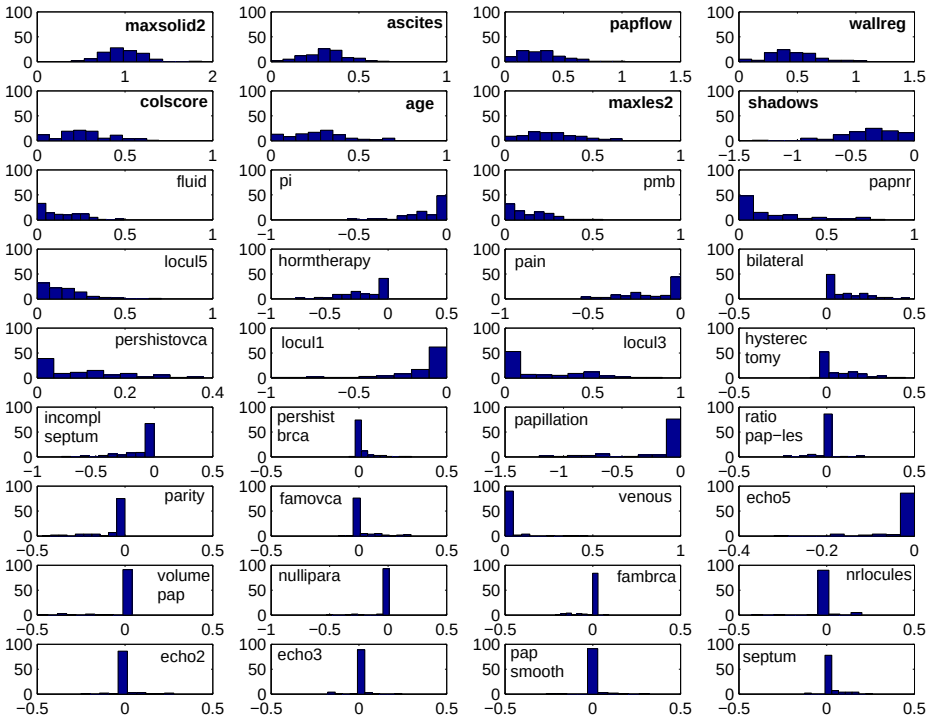


Figure 5.5: Histograms of standardized parameter estimates based on 100 bootstrap samples.

An unregularized model using objective inputs

Because LR1 includes subjective measurements such as pain and colscore, we derived a model that uses only objective inputs. Objectivity of the inputs was determined

by an experience gynecologist [20]. The two subjective inputs of LR1 were excluded and replaced by locul1. This input is selected based on its strong relationship with tumor outcome in the training set: unilocular tumors were malignant in one out of 228 cases (0.4%), whereas other tumors were malignant in 36.1% of the cases. Its inclusion does not make the application of the model more complicated as one of the other inputs in the model is locul5, which is another value of the categorical variable locularity (see Appendix). This model (ObjLR) has 11 inputs: pershistovca (1), hormtherapy (2), age (3), lesmax2 (4), ascites (5), papflow (6), locul1 (7), locul5 (8), solidmax2 (9), wallreg (10), and shadows (11). The probability of malignancy is computed as $\frac{1}{1+e^{-z}}$, where

$$\begin{aligned} z &= -5.8881 + 1.5620(1) - 0.8095(2) + 0.0325(3) + 0.0139(4) + 1.5876(5) \\ &\quad +1.2977(6) - 2.1008(7) + 0.8893(8) + 0.0501(9) + 0.8913(10) - 2.8520(11). \end{aligned} \tag{5.4}$$

On the training set, a threshold probability of 0.15 to predict malignancy resulted in 92.7% sensitivity and 80.1% specificity (LR+ 4.66, LR- 0.09). Hopes were to obtain a model with at least similar performance to LR1. Moreover, in a prospective evaluation using newly recruited clinical centers, the avoidance of subjective inputs may be advantageous.

5.2.2 Advanced models

Bayesian least squares support vector machines and relevance vector machines

In this section, the construction of mathematical models using Bayesian LS-SVMs and RVMs are described. These models were built by Lu [266, 426], and our contribution to them was limited. Bayesian LS-SVMs and RVMs can construct nonlinear models and are thus an interesting alternative or addition to logistic regression models. They are mathematically complex models that are flexible in that they are able to create a nonlinear decision boundary as opposed to for example logistic regression. Recently, an opinion paper was published, in which characteristics of support vector machines were compared with those of logistic regression [326]. In total, six models were constructed: Bayesian LS-SVM and RVM models with either a linear, radial basis function (RBF), or additive RBF kernel. A disadvantage of using nonlinear kernels such as the RBF kernel is that it is difficult to disentangle the influence of each variable on the outcome. Linear models do not have this disadvantage since linear models generally use a weighted sum of the predicting variables to predict the outcome such that we can examine the effect of each predicting variable separately. Therefore, additive methods have been developed for LS-SVMs with RBF kernel [317], and were applied in this study.

To select the predicting variables for the models, a forward selection method was used within the context of Bayesian LS-SVM models with either linear or

RBF kernel. In each step of this greedy selection process, the variable that most increased the Bayesian model evidence was added to the model. The selection process was stopped when no additional variable was able to further increase the model evidence. Because this forward selection method may select many variables, an additional backward elimination procedure may be necessary in order to obtain a more reliable set of predicting variables. Such backward elimination can be done subjectively based on knowledge of the variables (e.g. knowledge about the subjectivity and measurement accuracy of variables, knowledge about associations between variables), and by investigating which variables least decrease the model evidence.

The adopted variable selection procedure selected 12 variables: solidmax2, age, maximum diameter of the whole ovary (ovmax), ascites, papflow, shadows, wallreg, very strong intratumoral blood flow (col4), hormtherapy, a multilocular solid tumor (locul4), pershistovca, and suspicion at ultrasound examination of ovarian origin of the tumor (origin1). This subset of selected variables served as the variable set for all six models that were fitted. The bounded version of solidmax (solidmax2) was used to be consistent with LR1. In the discussion, we shortly comment on the clinically awkward observation that ovmax is used in the model.

Thresholds chosen for the LS-SVM and RVM models varied between 0.12 and 0.20 (see Table 5.4 for more information).

Bayesian multi-layer perceptrons

Another popular type of nonlinear models that can be used for classification are MLPs. By using a hidden layer to connect the input and output layers, MLPs are flexible nonlinear models that constitute 'universal approximators' [192]: a neural network with one hidden layer is able to approximate any continuous function to an arbitrary degree. Such models are able to capture any nonlinearity in the decision boundary. We used a Bayesian approach to MLPs based on the evidence procedure. Bayesian MLPs are not widely used in practice even though they have several advantages over regular MLP algorithms [293].

Input selection and neural network modeling Because it is preferable that all the input variables have a similar range for neural network modeling, we re-scaled the continuous variables into the $[-1;\ +1]$ interval. This was considered better than standardization due to the skewness of some variables. Binary variables were coded as -1 versus $+1$. Since there were over 40 input variables and previous input selection analyses were already done (cf. previous sections), we decided to start with a smaller set of 19 promising variables based on previous input selection procedures. This set consisted of age, pershistovca, hormtherapy, postmenopausal bleeding (pmb), pain, ascites, lesmax, the number of papillary projections (papnr), papflow, 5 dummy variables based on locularity (locul1-locul5), solidmax, wallreg,

shadows, colscore, and a binary variable indicating whether lesmax and ovmax are the same (lesovdif).

We decided to apply ARD on this subset to avoid needlessly complex models whose application is more effortful because more measurements need to be made. Since the solutions can vary depending on the initial values of the weights and hyperparameters, the ARD model was run 10 times. Based on the resulting α_i values that represent the inverse variance of the prior distribution of the weights fanning out from input i, the variables were ranked with respect to their relevance for the model. The median of the 10 ranks was used as the main criterion to select which input(s) to drop. The ARD analysis was repeated whenever inputs were dropped in order to account for ranking changes caused by modified input sets [293]. One disadvantage of ARD is that the endpoint of the pruning is subjective. We decided not to use a simple endpoint but to take an upper and lower limit for the number of inputs in the model. We continued with the suggested input sets of size 6 to 14 (these limits are subjectively chosen but could be extended if the results would suggest this).

Consistent priors were used throughout, meaning that separate hyperparameters were used for each layer's weights and biases. Experience has shown that this is necessary for the algorithm to work properly. Since weight decay parameters are automatically tuned in Bayesian neural network modeling, the only parameters to be tuned were the number of inputs and the number of hidden neurons (i.e., the network architecture). For this, stratified 5CV was used. The training set was divided into five parts of equal size with stratification for outcome [241]. Five MLPs were fitted using only four of the five parts. For each MLP another part was left out to serve as the validation set. The mean validation cross-entropy and AUC were recorded as model selection criteria.

Next, more attention was given to the region of optimal performance based on 5CV. This is done using repeated 5CV (R5CV): the basic 5CV method was applied 20 times, each time with a different randomly chosen stratified partition of the training set into five parts. Whereas in normal 5CV one has five validation set results to be averaged, R5CV leads to 20*5 = 100 validation set results. This strategy accounted for unwanted effects of using a single partition of the training set. Because it is more time-consuming, R5CV was only used for a limited set of architectures, as specified later.

When a network architecture was chosen, ten such models were fitted to the entire training set using re-initialization of the model parameters. Different initial values may lead to different results because parameter optimization for neural networks is not a convex problem. The network with the best training performance was chosen as the final network.

As a linear benchmark model, we also fitted a perceptron using the evidence framework (this is in fact a Bayesian logistic regression model using ARD input selection). Here, we preferred the term perceptron over logistic regression because the necessity of interaction terms, higher-order terms, or transformations to meet the linearity in the logit assumption was not investigated.

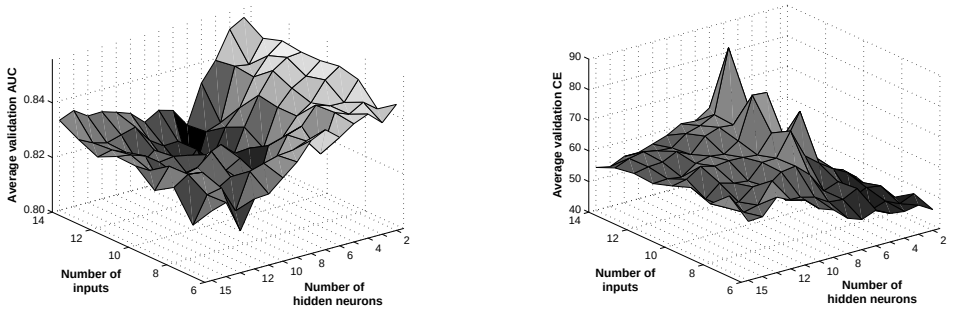
Input selection results ARD was used to obtain input sets with 6 to 14 variables. The analysis suggested the following 14 inputs: pershistovca, shadows, unilocular-solid tumor (locul2), wallreg, hormtherapy, papnr, papflow, locul5, ascites, locul4, colscore, age, lesmax, and solidmax. Further analyses dropped inputs in this order until only the last six remained. Thus, the input set of size x included the last x inputs mentioned above. Note that ARD analyses use a hidden layer of user-defined size. We chose 10 hidden neurons in order to allow enough flexibility for eventual nonlinearity to emerge (further referred to as ARD10). This means that ARD analyses will select variables that are apt for the predefined hidden layer size but not necessarily for other sizes. Therefore, the number of hidden neurons suggested by cross-validation was used to redo the ARD analyses. We investigated whether this improved model performance. The largest network trained thus consisted of all 19 inputs and 10 hidden layers, resulting in 211 weights for 754 input patterns, or 3.6 patterns for each weight.

Cross-validation to determine network architecture For the input sets with 6 to 14 inputs, 5CV was performed for architectures containing 2 to 15 hidden neurons. The results are shown in Figure 5.6 (for the input sets of size 6 to 12 only). In general, two hidden neurons were the best choice. Therefore, we zoomed in to the two hidden neuron architectures and performed R5CV. These results are given in Figure 5.7(a). The overall CV results suggested a model using 11 inputs and 2 hidden neurons (11-2 architecture).

Since CV suggested the use of only two hidden neurons, the ARD analyses were redone with 2 hidden neurons (ARD2). This time, ARD and subsequent R5CV (see Figure 5.7(b)) again suggested a model with 11 inputs but only 7 of the inputs were also used in the 11-2 network based on ARD10: wallreg, papflow, locul5, ascites, age, lesmax, solidmax, pershistovca, pain, shadows, locul1. Note the clearly superior R5CV results for models based on ARD2 as compared to ARD10 (Figure 5.7).

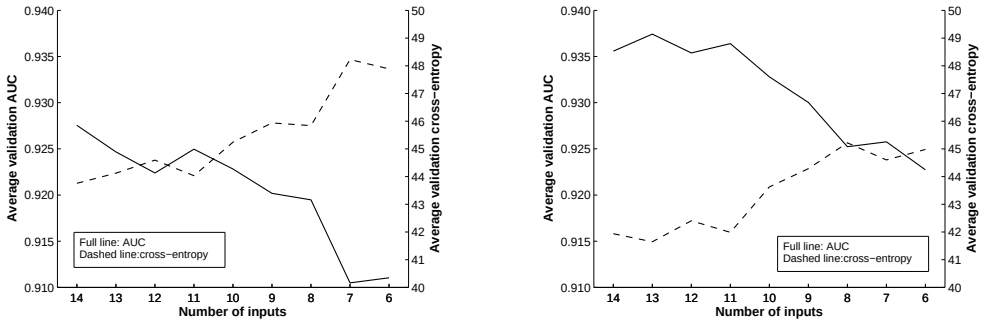
The final networks Out of the ten 11-2 neural networks based on ARD10 that were fitted using different initial values for the model parameters, the one with the lowest training set cross-entropy (BMLP11-2a) had a training set AUC of 0.943. For the model based on ARD2, the final network (BMLP11-2b) had a training set AUC of 0.952. For both models, 0.15 was chosen as the threshold to predict malignancy.

A linear benchmark: Bayesian perceptron For the Bayesian perceptron model the same strategy was followed. The ARD analyses yielded the following



(a) **(b)**

Figure 5.6: The average validation set (a) AUC and (b) cross-entropy based on five-fold cross-validation



(a) **(b)**

Figure 5.7: The average validation set AUC and cross-entropy, based on R5CV, for selecting inputs using (a) ARD10 and (b) ARD2. An MLP with two hidden neurons is used.

set of 14 inputs: lesovdif, hormtherapy, pain, colscore, pershistovca, wallreg, locul5, lesmax, papflow, ascites, locul1, age, shadows, solidmax. Further ARD analyses dropped the inputs in the order they are presented until the last six inputs remain. Choosing for the best input set was done using the R5CV method. These analyses indicated that more than 12 inputs could be interesting. We decided, however, to choose the 11 input model since that model had the same number of variables as BMLP11-2a and BMLP11-2b, and since 13 or more inputs was considered too complex. The 11 input model (BPER11) had a training set AUC of 0.940. Again, 0.15 was the selected threshold to predict malignancy.

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 p inputs to generate the probabilities, the use of a threshold value means that a decision boundary is defined in p-dimensional space to separate regions with a different crisp prediction. However, in advanced models such as LS-SVMs with RBF kernel or MLPs, the decision boundary can be highly nonlinear such that it is often not clear how the inputs are used and combined 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) [149], and was applied to some of the models described in this section [28]. This was done at Liverpool John Moores University, with limited contributions from ourselves.

OSRE automatically extracts low order rules using the decision boundary and the training data that were used to derive the model and accompanying boundary. For each training data point in p-dimensional space that is predicted to be malignant, the algorithm searches in orthogonal directions for hypercubes spanning the part of the data space for which the model makes the prediction 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 in the hypercube. There are as many rules as there are cases that are predicted to be malignant. The full set of rules is pruned using some criterion. In the application of OSRE to BMLP11-2a [28], rule pruning was performed by maximizing the positive predictive value of the (disjunctive) 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 5.3. The set of five rules for BMLP11-2a had a sensitivity of 52.6% and a specificity of 99.6%. The PPV and NPV were 97.9% and 86.3%; the LR+ and LR- were 140.4 and 0.48. 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 I data and are presented in section 5.2.4. The procedure used to extract the rules from Table 5.3 has lead to a set of disjunctive rules with near perfect specificity and PPV. Thus, these rules only help to detect cases for which one can be highly confident that they are malignant.

Table 5.3: OSRE rule set for BMLP11-2a.

Rule	Conjunctive conditions	Rule	Conjunctive conditions
1	Wallreg=1	4	Wallreg=1
	Colscore=4		Locul5=1
	$51.6 \leq$ Age ≤ 93.5		$49.7 \leq$ Age ≤ 92.3
2	Hormtherapy=0	5	Hormtherapy=0
	Wallreg=1		Papnr 3, 4, or 5
	Colscore=4		Colscore 3 or 4
	$108 \leq$ Lesmax ≤ 403		$28.7 \leq$ Solidmax ≤ 224.2
3	Papnr=4		
	Papflow=1		
	Colscore 3 or 4		
	$59 \leq$ Lesmax ≤ 401		
	$19.9 \leq$ Solidmax ≤ 227		

5.2.3 Model evaluation on the test set

The developed classifiers will now be applied to the test set and their performance will be compared. The performance of the models will also be related to the performance of the subjective assessment of the mass by a expert operator. Since one model, ObjLR, was derived differently (subjective measurements could not be selected as inputs for the model), it is not compared to all other models, but only to the other logistic regression models. Calibration curves will not be produced for the test set only. The IOTA I test set will be merged with the IOTA Ib data set before investigating calibration. This was done to reduce the output.

Evaluation and comparison of the classifiers

Table 5.4 presents the test set performance of the classifiers. The main criterion to evaluate performance was the AUC. A 95% confidence interval (CI) on the AUC was computed using the bias-corrected bootstrap method based on 1000 bootstrap samples of the test set [141]. Other criteria were the partial AUC (pAUC) for that part of the ROC curve where sensitivity is at least 75%, Sensitivity at 75% specificity (Sens75), the scored AUC (sAUC), and, using the threshold determined on the training set, sensitivity, specificity, LR+, and LR-. The performance of subjective assessment is also shown, as well as the AUC of the widely used risk of malignancy index (RMI). The table shows excellent performance of all models, with AUCs between 0.92 and 0.95. Differences in AUC were small, as can be seen in Figures 5.8(a)-5.8(d) showing the test set ROC curves (and the performance of the expert operator's subjective assessment). pAUC and Sens75 were also very good. When incorporating the actual predicted probabilities into the AUC metric (cf. sAUC), the Bayesian LS-SVM models performed best. This is an important observation, as this may indicate that small changes in predicted probabilities do

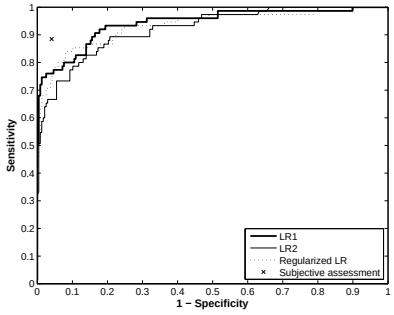
not result in big changes in the AUC, making these classifiers more robust. The subjective assessment of the expert operators had excellent performance, mainly for specificity. On the contrary, the thresholds for mathematical models were derived based on the opinion that a false negative (an undetected cancer) is worse than a false positive. Thus, thresholds were derived that resulted in larger sensitivity versus specificity. The ROC curves show that, relative to subjective assessment, mathematical models obtain lower sensitivity levels for near perfect specificity levels.

Table 5.4: Test set results for subjective assessment and mathematical models.

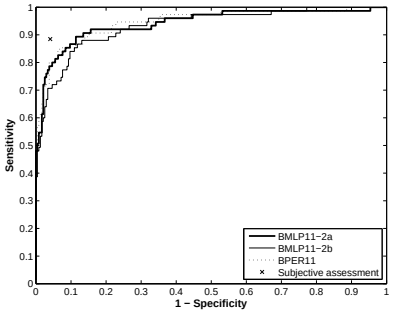
Model (threshold)	AUC (95% CI)	AUC_{CA}	pAUC	Sens75	sAUC	Sens Spec LR+ LR-
Subj. ass.						86.7 – 93.7 – 13.7 – 0.16
LR1 (0.10)	.942 (.905 – .972)	.936	.209	93.3	.605	93.3 – 75.5 – 3.81 – 0.09
LR2 (0.10)	.922 (.882 – .954)	.916	.193	89.3	.549	89.3 – 73.4 – 3.36 – 0.15
Reg. LR (0.20)	.932 (.888 – .965)	.924	.205	93.3	.447	89.3 – 77.6 – 3.99 – 0.14
BMLP11-2a (0.15)	.942 (.900 – .974)	.934	.212	92.0	.589	92.0 – 80.6 – 4.74 – 0.10
BMLP11-2b (0.15)	.933 (.892 – .964)	.929	.200	92.0	.551	92.0 – 74.7 – 3.63 – 0.11
BPER11 (0.15)	.947 (.908 – .974)	.942	.211	94.7	.569	94.7 – 75.9 – 3.94 – 0.07
BLSSVMlin (0.15)	.946 (.911 – .973)	.940	.209	96.0	.622	90.7 – 84.4 – 5.81 – 0.11
BLSSVMrbf (0.12)	.945 (.909 – .971)	.941	.208	94.7	.632	93.3 – 81.0 – 4.92 – 0.08
BLSSVMaddrbf (0.12)	.943 (.905 – .970)	.937	.205	96.0	.665	90.7 – 82.3 – 5.12 – 0.11
RVMlin (0.20)	.949 (.913 – .975)	.942	.212	96.0	.607	92.0 – 82.3 – 5.19 – 0.10
RVMrbf (0.15)	.946 (.910 – .972)	.940	.209	96.0	.606	92.0 – 80.6 – 4.74 – 0.10
RVMaddrbf (0.15)	.946 (.910 – .973)	.941	.208	96.0	.619	93.3 – 81.4 – 5.03 – 0.08
RMI		.870				
Obj. LR (0.15)	.937 (.897 – .964)	.931	.202	90.7	.597	90.7 – 76.0 – 3.77 – 0.12

AUC_{CA}: AUC using only the patients for whom CA-125 levels were determined.

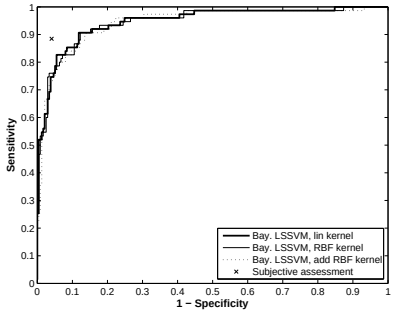
Next, models were compared using a ranking analysis based on 1000 bootstrap samples drawn from the test set. The ranking analysis and its rationale are explained in section 2.3.3. For each sample, the models were ranked based on a performance criterion. The average rank (AR) was computed, as well as the proportion of times a model was ranked first (Pm(1)) or among the 3 best models (Pm(3)) [321]. The ranking analysis was performed for the AUC and the sAUC (see Table 5.5). With respect to the AUC, three models had a large Pm(1): RVMlin (0.36), BPER11 (0.24), and RVMaddrbf (0.16). Yet, no large differences in AUC were found, with AR for the second and ninth model being rather close (4.4 vs 6.9). With respect to sAUC, differences were much larger. It is clear that the Bayesian LS-SVM models performed best, most notably the model with the additive RBF kernel. These classifiers seemed to have produced the best probabilities, and may thus be preferred. BPER11, which had one of the best AUCs, performed poor for sAUC. Thus, BPER11's AUC may be less robust. The regularized LR model appeared to have shrunk the parameter estimates to a large extent, considering its poor sAUC value and AR. The larger variation in sAUC values versus AUC values is clear from the box plots in Figure 5.9. The comparison of ObjLR with the other LR models shows that ObjLR performed



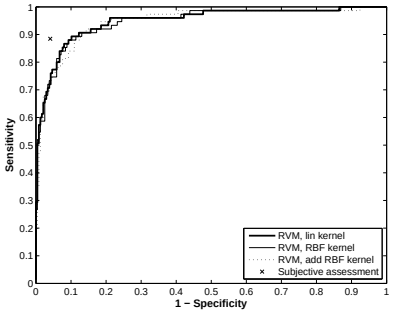
(a)



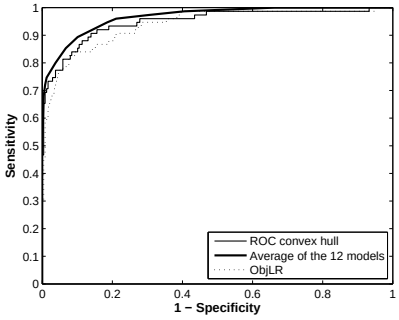
(b)



(c)



(d)



(e)

Figure 5.8: ROC curves of the developed classifiers on the test set part of the IOTA phase I database. The × marks the performance of the expert operator's subjective assessment.

less than LR1, yet the difference is small. LR2 and RegLR performed worst.

Table 5.5: Ranking of the mathematical models using bootstrap analysis. At the bottom, the additional ObjLR is compared with the other LR models.

	Ranking for AUC					Ranking for sAUC			
Model	AUC	AR	Pm(1)	Pm(3)	Model	sAUC	AR	Pm(1)	Pm(3)
RVMlin	0.9485	2.5	.362	.790	BLSSVMaddrbf	0.665	1.0	.984	1
BPER11	0.9466	4.4	.241	.446	BLSSVMrbf	0.632	2.3	.016	.947
RVMrbf	0.9463	4.5	.016	.336	BLSSVMlin	0.622	3.4	.000	.568
RVMaddrbf	0.9462	4.4	.158	.417	RVMaddrbf	0.619	4.0	.000	.390
BLSSVMlin	0.9462	4.6	.021	.282	RVMlin	0.607	5.8	.000	.007
BLSSVMrbf	0.9452	5.4	.022	.167	RVMrbf	0.606	6.1	.000	.002
BLSSVMaddrbf	0.9430	6.8	.041	.166	LR1	0.605	6.0	.000	.063
LR1	0.9423	6.8	.031	.139	BMLP11-2a	0.589	7.6	.000	.023
BMLP11-2a	0.9420	6.9	.081	.191	BPER11	0.569	9.1	.000	.000
BMLP11-2b	0.9331	9.8	.007	.025	BMLP11-2b	0.551	10.3	.000	.000
Reg. LR	0.9324	10.1	.005	.020	LR2	0.549	10.4	.000	.000
LR2	0.9217	11.7	.001	.001	Reg. LR	0.447	12.0	.000	.000
LR1	0.9423	1.3	.762		LR1	0.605	1.2	.800	
Obj. LR	0.9369	2.2	.161		Obj. LR	0.597	1.8	.200	
Reg. LR	0.9324	2.8	.061		LR2	0.549	3.0	.000	
LR2	0.9217	3.8	.012		Reg. LR	0.447	4.0	.000	

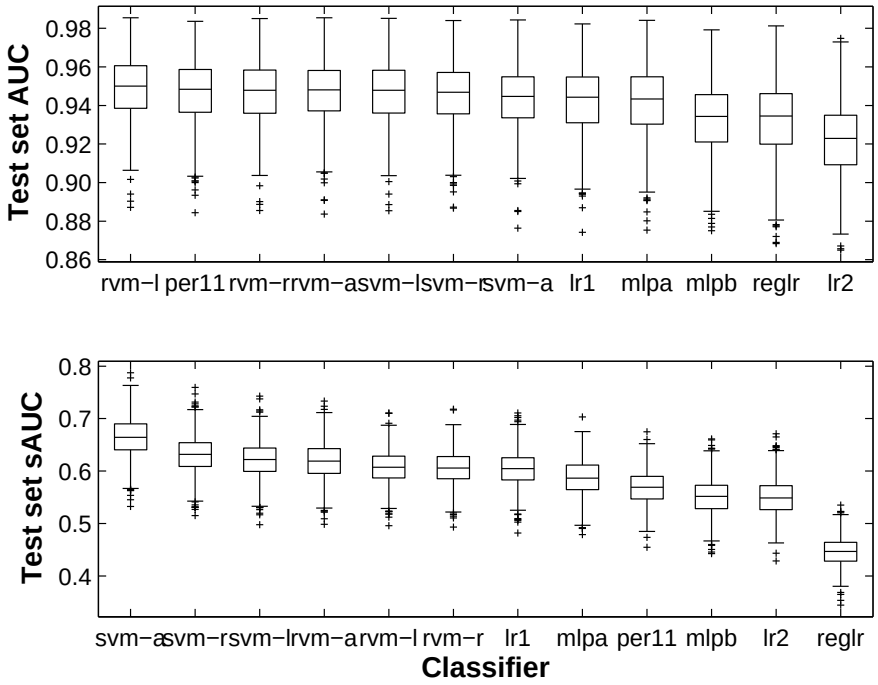


Figure 5.9: Boxplots of AUC and sAUC values using 1000 bootstrap samples.

Extra ROC analyses: combining classifiers

A lot of research is devoted to classifier ensembles and classifier fusion, aimed at constructing better hyperclassifiers by combining strengths of various individual classifiers [247, 357]. See [133] for an application to cancer diagnosis. This is beyond the scope of this work, yet we investigated two simple methods. In the first method, which is one of the simplest fusion methods, we merely compute an unweighted average of the classifiers (except ObjLR). In the second method, we compute the ROC convex hull (ROCCH) of the classifiers (except ObjLR) [331]. The ROCCH is the convex hull of all ROC points of the classifiers, it is the northwest boundary of these points. Therefore, the ROCCH is the curve with highest AUC. A hard classifier (e.g. a probabilistic model combined with a threshold) is potentially optimal if it lies on the ROCCH, which would imply that only the hard classifiers that are part of the ROCCH should be considered. The ROCCH and the ROC curve based on the unweighted average of the classifiers are shown in Figure 5.8(e). The former has an AUC of 0.966, the latter has an AUC of 0.948. The ROC points on the ROCCH are listed in Table 5.6. It shows that a threshold of 0.14 for the RVMaddrbf model yields a sensitivity and specificity of 94.7 and 81.4%, respectively. This threshold is close to the one developed for the RVMaddrbf model (i.e. 0.15). The ROCCH uses results from 6 classifiers. Based on the ROCCH, an ad hoc ensemble classifier can be derived:

- if LR2 < 0.018 then score=0 (occurred for 80 benign and 0 malignant test set cases, 80vs0)
- else if RVMaddrbf < 0.041 then score=1 (66vs1)
- else if BLSSVMaddrbf < 0.038 then score=2 (22vs1)
- else if RVMlin < 0.159 then score=3 (21vs1)
- else if RVMaddrbf < 0.140 then score=4 (8vs1)
- else if RVMlin < 0.311 or BPER11 < 0.337 then score=5 (30vs7)
- else if BPER11 < 0.460 or LR1 < 0.589 then score=6 (7vs10)
- else if LR1 < 0.681 then score=7 (2vs4)
- else if LR1 < 0.781 then score=8 (1vs12)
- else score=9 (0vs38).

Some ROCCH points are combined such that a higher score is related with a higher malignancy rate in the test set. This ad hoc hyperclassifier needs evaluation on new data as it was constructed on the test set.

Table 5.6: ROC points on the ROC convex hull.

Sensitivity	Specificity	Probability of Malignancy (classifier)
50.7	100	0.781 (LR1)
68.0	99.6	0.681 (LR1)
72.0	99.2	0.622 (LR1)
74.7	98.7	0.589 (LR1)
80.0	96.2	0.460 (BPER11)
85.3	93.2	0.337 (BPER11)
89.3	89.9	0.311 (RVMlin)
94.7	81.4	0.140 (RVMaddrbf)
96.0	78.9	0.159 (RVMlin)
97.3	69.6	0.038 (BLSSVMaddrbf)
98.7	59.5	0.041 (RVMaddrbf)
100	33.8	0.018 (LR2)

Comparing algorithms

Because the classifiers were developed by several researchers, no thorough analysis to compare the used algorithms is available. Table 5.7 presents the comparison of the neural network algorithms with the LR1 algorithm. This analysis shows that the LR1 algorithm gave the best results, and that the BMLP11-2b algorithm (where the size of the hidden layer for the ARD input selection analysis was tuned) had better results than the BMLP11-2a algorithm.

Table 5.7: Ranking of four mathematical models using 100 train-test splits.

Model	Median AUC	AR	Pm(1)
LR1	0.940	1.5	.670
BMLP11-2b	0.939	2.8	.180
BPER11	0.938	2.2	.100
BMLP11-2a	0.930	3.5	.050

5.2.4 A simple alternative: simple rules for immediate use

Mathematical models can be promising additional tools in clinical practice. However, they cannot be used at the exact moment of the ultrasound examination. The data needed by the model have to be entered in a computer in order to obtain the model's predicted probability of malignancy. Thus, it is also interesting to investigate whether a few very simple rules exist that the clinician can use immediately to help diagnosing the mass [393]. Such rules should highlight masses whose nature (benign or malignant) can be determined with high confidence.

All 1233 masses found in the 1066 patients from the IOTA phase I database were used. We looked at simple rules with both high applicability and good PPV or NPV. For rules suggesting malignancy, we aimed at very high PPV (i.e. proportion of malignant tumors among those predicted to be malignant). For rules suggesting benignity, we aimed at very high NPV (i.e. proportion of benign tumors among those predicted to be benign). Aiming at high NPV for rules predicting benignity is functionally similar to aiming at low PPV when using the rule to predict malignancy. The final selection of rules depended on PPV/NPV results, applicability, measurement cost, and correlations between rules. Five rules for predicting malignancy (M-rules) and five rules for predicting benignity (B-rules) were selected, see Tables 5.8 and 5.9.

Table 5.8: The best rules for predicting malignancy. The selected M-rules are indicated with an asterisk.

	PPV		
M-rule	N	%	LR+
1. Solid tumor with irregular internal cyst wall	81/87	93.1%	37.2
2. Ascites	112/135	83.0%	14.5
3. At least 4 papillary structures	75/99	75.8%	3.62
4. Multilocular-solid tumor, irregular walls, lesmax $\geq$100mm	77/103	74.8%	8.10
5. Very strong intratumoral blood flow (Colscore 4)	120/173	69.4%	6.17

Table 5.9: The best rules for predicting benignity. The selected B-rules are indicated with an asterisk.

	NPV		
B-rule	N	%	LR-
1. Unilocular tumor	5/377	1.3%	0.04
2. Maximum diameter of largest solid component <7mm	1/43	2.3%	0.03
3. Multilocular tumor, smooth walls, lesmax <100mm	5/133	3.8%	0.11
4. Acoustic shadows	5/120	4.2%	0.12
5. No intratumoral blood flow (Colscore 1)	16/285	5.6%	0.16

Tumors for which only benign rules applied were predicted to be benign, while tumors for which only malignant rules applied were predicted to be malignant. In any other situation, no prediction was given. Even though it could be considered to cast a prediction as soon as one type of rules applied more often than the other, we decided not to do so. The aim is to develop rules that give highly reliable predictions, and reliability of a prediction when rules of both types apply can be questioned. Using this procedure, our set of simple rules yields a prediction for 937 of all 1233 tumors (applicability of 76.0%). For these cases, the simple rules had a specificity of 93.2% (259/278) and a specificity of 90.1% (594/659). PPV and NPV

Table 5.10: Applicability of M-rules and B-rules to the 1233 tumors of the IOTA phase I data set.

		Outcome			
Situation	*N*	Benign	Primary invasive	Borderline	Metastatic
Prediction benign	613	594, 96.9%	7, 1.1%	9, 1.5%	3, 0.5%
0 M-rules, 1 B-rule	370	354, 95.7%	5, 1.4%	8, 2.2%	3, 0.8%
0 M-rules, 2 B-rules	220	217, 98.6%	2, 0.9%	1, 0.5%	0, 0.0%
0 M-rules, >2 B-rules	23	23, 100%	0, 0.0%	0, 0.0%	0, 0.0%
Prediction malignant	324	65, 20.1%	182, 56.2%	36, 11.1%	41, 12.7%
1 M-rule, 0 B-rules	169	55, 32.5%	74, 43.8%	23, 13.6%	17, 10.1%
2 M-rules, 0 B-rules	85	10, 11.8%	56, 65.9%	6, 7.1%	13, 15.3%
>2 M-rules, 0 B-rules	70	0, 0.0%	52, 74.3%	7, 10.0%	11, 15.7%
Undecided	278	232, 83.5%	23, 8.3%	18, 6.5%	5, 1.8%
Both types, undecided	49	46, 93.9%	2, 4.1%	1, 2.0%	0, 0.0%
No rules	229	186, 81.2%	21, 9.2%	17, 7.4%	5, 2.2%
Inconsistent	18	12, 66.7%	5, 27.8%	0, 0.0%	1, 5.6%
Both types, mostly B-rules	11	10, 90.9%	1, 9.1%	0, 0.0%	0, 0.0%
Both types, mostly M-rules	7	2, 28.6%	4, 57.1%	0, 0.0%	1, 14.3%

are 79.9% (259/324) and 96.9% (594/613), respectively. For 229 tumors, no rule was applicable (18.6%). For 67 tumors (5.4%), rules of both types were applicable. Table 5.10 presents an overview.

5.3 Multi-class diagnostic models

As mentioned before, the presurgical diagnosis of the tumor is used to make treatment decisions. Hitherto, many models were designed to distinguish benign from malignant tumors, both by IOTA members and others [287, 385, 447]. However, different types of malignant tumors exist. In this section, we develop classifiers to distinguish between benign, primary invasive, borderline, and metastatic invasive tumors. To our knowledge, no such models were developed before. The 36 variables surviving the multicollinearity analysis described in section 3.2.1 were used as possible inputs for the models.

5.3.1 Logistic regression models

The typical logistic regression model that is used when more than two classes are discerned, is the multi-class model (MLR) described in Chapter 2. However, this model uses one set of inputs to predict all classes at once. It is possible, or even likely, that inputs are not useful to distinguish between all classes, but only for distinguishing between one or more subsets of classes. Hence, it is interesting to compare MLR with an approach where binary (1-versus-1) logistic regression models are trained and combined using pairwise coupling (LR-PC) in order to arrive at multi-class probability estimates. This method entails the development of $\frac{K(K-1)}{2}$

binary models on the appropriate training set cases, the application of these models to all test set cases, and the combination of the binary results to obtain the K class probabilities. Pairwise coupling was performed using the first algorithm of [472], which had good performance in the comparison study presented in Chapter 4.

Because good input selection for MLR and the binary models was considered crucial, it was given ample attention. In general, the procedure was as follows. First, many promising input sets were selected. Logistic regression models using these sets were compared using 20 repeated runs of 5CV (R5CV). Per outcome class, this resulted in 100 validation AUCs that were averaged by taking their mean. Next to these AUC results, models were also compared with respect to their AIC and BIC value. The aim was to find a model with low AIC, low BIC, and high average AUCs for each outcome. The number of inputs as well as the inputs themselves also played a role in model selection. Promising input sets were selected using various methods, such as automated stepwise selection (using $p < 0.4$ as criterion to enter the model, and $p < 0.1$ as criterion to stay in the model), manual backward selection (using many inputs may give computational difficulties such that an automated procedure was not preferred), and a stepwise ranking method as described in [358].

In the stepwise ranking method, a ranking of all inputs is obtained using stepwise selection by setting the entry and stay criteria to 1. Then, information criteria such as the AIC and BIC can be computed for input sets of increasing size using the ranking. This will show how many of the highest ranked inputs yield the best AIC or BIC (the optimal AIC model usually has more inputs than the optimal BIC model). This method provides interesting input sets, but also suggests how many inputs are approximately needed for obtaining good AIC or BIC.

Multi-class logistic regression (MLR)

Using the procedure described in the previous section, we detected 30 promising input sets. The BIC criterion applied to the stepwise ranking method suggests the use of around six inputs. The AIC criterion suggests no less than 17 inputs. Therefore, we suspect that around 8-12 inputs may lead to a good diagnostic model. For each of these, AIC, BIC, and the AUCs for each class based on R5CV were retrieved. The results are shown in Table 5.11. The maximum diameter of the lesion and of the largest solid component were included using the bounding as explained earlier (lesmax2 and solidmax2). This choice was confirmed when checking linearity in the logit. Performance for borderline and metastatic tumors was considered most important. However, if performance for one of these classes was good, performance for the other class was usually relatively poor. In our opinion, a good compromise is the input set L. Checks for the linearity in the logit assumption confirmed the preference of solidmax2 and lesmax2 over solidmax and lesmax. Checks for interactions suggested that the interaction between ascites and bilateral could possibly be useful. We did not include the interaction as another input for two reasons. First, when there are tumors on both ovaries, the relative risk

for ascites is 4.1. This means that one has a 4.1 times higher risk of malignancy when ascites is present than when it is absent. When one or more tumors are found on one ovary only, the relative risk is 4.6. Second, cross-validation analysis suggested only minor improvements in performance, while it is known that interactions are often hard to replicate (due to a higher risk of overfitting).

Table 5.11: Results of the input selection analysis for the MLR model to distinguish between four types of ovarian tumors.

				Average validation AUC			
Inputs	Size	AIC	BIC	Benign	Pr. Inv.	Bord.	Metast.
All input sets contain solidmax2, wallreg, ascites, and papflow (basic inputs)							
A. Basic + shadows lesmax2	6	735.1	832.3	0.929	0.931	0.772	0.830
B. Basic + shadows locul3	6	733.9	831.1	0.930	0.935	0.729	0.843
C. A + locul3	7	723.7	834.7	0.935	0.935	0.780	0.835
D. B + age	7	720.5	831.5	0.934	0.938	0.739	0.850
E. C + age	8	711.8	836.7	0.939	0.936	0.780	0.844
F. Basic + age lesmax2 persh. locul5	8	734.0	858.9	0.932	0.916	0.773	0.877
G. Basic + lesmax2 locul3 bil. hormth.	8	735.8	860.7	0.933	0.929	0.793	0.839
H. E + pershistovca	9	705.8	844.5	0.941	0.930	0.783	0.851
I. F + locul3	9	722.8	861.6	0.937	0.920	0.780	0.877
J. F + shadows	9	714.9	853.7	0.935	0.925	0.765	0.858
K. F + colscore	9	726.1	864.8	0.934	0.919	0.765	0.870
L. F + bilateral	9	726.1	864.9	0.934	0.918	0.780	0.871
M. F + hormtherapy	9	731.8	870.6	0.935	0.916	0.775	0.879
N. E + bilateral	9	704.8	843.6	0.939	0.938	0.786	0.838
O. G + age	9	722.3	861.0	0.939	0.930	0.793	0.854
P. Basic + age lesmax2 persh. locul3 bil.	9	717.8	856.6	0.937	0.924	0.798	0.852
Q. Basic + age lesmax2 persh. locul3 hormth.	9	723.4	862.2	0.939	0.922	0.791	0.865
R. H + bilateral	10	699.0	851.6	0.940	0.931	0.788	0.842
S. I + shadows	10	703.9	856.5	0.940	0.929	0.770	0.863
T. I + bilateral	10	714.6	867.2	0.938	0.923	0.787	0.872
U. I + colscore	10	715.7	868.3	0.939	0.923	0.773	0.871
V. I + hormtherapy	10	720.1	872.7	0.939	0.920	0.778	0.880
W. J + hormtherapy	10	713.6	866.2	0.939	0.925	0.767	0.860
X. J + bilateral	10	707.9	860.6	0.936	0.926	0.773	0.851
Y. J + colscore	10	714.3	867.0	0.936	0.925	0.758	0.853
Z. O + pershistovca	10	716.0	868.6	0.940	0.923	0.792	0.857
AA. G + locul5 pershistovca	10	724.5	877.2	0.936	0.920	0.782	0.869
AB. O + shadows	10	702.9	855.5	0.943	0.938	0.785	0.841
AC. R + volumepap	11	698.3	864.8	0.938	0.931	0.765	0.842
AD. AC + locul5	12	697.1	877.5	0.938	0.931	0.756	0.855

The final MLR model thus has nine inputs: pershistovca (1), bilateral tumors (2; bilateral), age (3), lesmax2 (4), ascites (5), papflow (6), locul5 (7), solidmax (8), and wallreg (9). The obtained classifier has this form:

$$\begin{cases} \log\left(\frac{\hat{\pi}_{pr}}{\hat{\pi}_b}\right) &= -8.99 + 0.76(1) + 1.13(2) + 0.042(3) + 0.012(4) + 1.87(5) \\ & \quad +1.16(6) + 0.84(7) + 0.068(8) + 2.05(9) = z_1 \\ \log\left(\frac{\hat{\pi}_{bo}}{\hat{\pi}_b}\right) &= -6.61 + 2.12(1) - 0.41(2) + 0.021(3) + 0.022(4) + 0.54(5) \\ & \quad +2.03(6) - 0.07(7) + 0.023(8) + 0.15(9) = z_2 \\ \log\left(\frac{\hat{\pi}_{me}}{\hat{\pi}_b}\right) &= -10.84 + 3.62(1) + 0.86(2) + 0.045(3) + 0.011(4) + 1.24(5) \\ & \quad +0.81(6) + 1.96(7) + 0.067(8) + 2.38(9) = z_3, \end{cases} \tag{5.5}$$

where $\hat{\pi}_b$, $\hat{\pi}_{pr}$, $\hat{\pi}_{bo}$, and $\hat{\pi}_{me}$ represent $\hat{\pi}_{benign}(\mathbf{x})$, $\hat{\pi}_{primaryinvasive}(\mathbf{x})$, $\hat{\pi}_{borderline}(\mathbf{x})$, and $\hat{\pi}_{metastatic}(\mathbf{x})$. The multi-class probabilities can be computed using the formula in given in section 2.2.1. Thus, three parameter estimates are produced for each input. Of all these 27 parameter estimates, 10 are not statistically significant at the 0.05 level. Even though statistical significance tests are not our main criterion for gauging the existence of 'effects', this suggests that some of these estimates are actually random variations around the true value zero. Thus, some level of overfitting may be induced that can be avoided by the use of LR-PC, where 1-vs-1 binary models are built, each with separate input selection.

Combination of pairwise models using pairwise coupling (LR-PC)

For LR-PC, we need to develop all six 1-versus-1 models. For each of these models, we performed an input selection analysis similar to that for the MLR model. We will not present all details here. In total, 75 input sets were tested. The selection of the final input set for each binary model was also influenced by the total number of inputs used: one prefers a well-performing model that needs a limited set of measurements. The six selected input sets contain 10 different inputs. The checks for linearity in the logit again supported the use of lesmax2 and solidmax2. No interactions were potentially useful. The following inputs were selected:

- benign vs primary invasive: age (1), ascites (2), papflow (3), solidmax2 (4), wallreg (5);
- benign vs borderline: papflow (3), lesmax2 (6), locul1 (7);
- benign vs metastatic: age (1), ascites (2), solidmax2 (4), wallreg (5), locul5 (8);
- primary invasive vs borderline: ascites (2), solidmax2 (4), bilateral (9);
- primary invasive vs metastatic: ascites (2), locul5 (8), pershistovca (10);
- borderline vs metastatic: ascites (2), solidmax2 (4), locul5 (8).

The six 1-versus-1 models are the following:

$$\begin{aligned}
\log\left(\frac{\hat{\pi}_{pr}}{\hat{\pi}_{b}}\right) &= -7.57 + 0.042(1) + 1.94(2) + 1.24(3) + 0.079(4) + 1.77(5),\\
\log\left(\frac{\hat{\pi}_{bo}}{\hat{\pi}_{b}}\right) &= -4.99 + 2.22(3) + 0.024(6) - 2.35(7),\\
\log\left(\frac{\hat{\pi}_{me}}{\hat{\pi}_{b}}\right) &= -9.58 + 0.041(1) + 1.55(2) + 0.070(4) + 2.95(5) + 1.97(8),\\
\log\left(\frac{\hat{\pi}_{bo}}{\hat{\pi}_{pr}}\right) &= 1.17 - 1.38(2) - 0.046(4) - 1.39(9),\\
\log\left(\frac{\hat{\pi}_{me}}{\hat{\pi}_{pr}}\right) &= -1.84 - 0.73(2) + 1.41(8) + 3.33(10),\\
\log\left(\frac{\hat{\pi}_{me}}{\hat{\pi}_{bo}}\right) &= -2.39 + 1.43(2) + 2.43(8) + 0.032(4). \qquad (5.6)
\end{aligned}$$

This analysis suggests that patients with a primary invasive tumor can be distinguished from patients with a benign tumor because the former are older, have a larger solid component (if any), have more often ascites, and have more often a tumor with blood flow within the papillary structures and irregular internal cyst walls. Interestingly, patients with a borderline tumor are distinguished from patients with a benign tumor because the former have less often a unilocular tumor, but have a larger tumor and have more often a tumor with blood flow within the papillary structures. Patients with a metastatic tumor are distinguished from patients with a benign tumor because the former are older, have a larger solid component (if any), have more often ascites, have more often an entirely solid tumor, and have more often a tumor with irregular internal cyst walls. Distinctions between patients with different types of malignant tumors are made using the observations that (i) patients with a borderline tumor have a tumor with smaller solid component (if any), have less often tumors on both ovaries, and have less often ascites than patients with a primary invasive tumor; (ii) patients with a metastatic tumor have more often a history of ovarian cancer, less often ascites, but more often an entirely solid tumor than patients with a primary invasive tumor; (iii) patients with a metastatic tumor have a tumor with larger solid component (if any), have more often ascites, and more often an entirely solid tumor than patients with a borderline tumor.

The application of the LR-PC starts with applying all six models to the data of a patient. Then, the six binary probability estimates are combined using the pairwise coupling algorithm to yield a predicted probability for all four tumor types.

5.3.2 Advanced models

We compared the logistic regression models to three advanced mathematical models: 1-versus-1 Bayesian least squares support vector machines combined using pairwise

coupling (BLSSVM-PC), 1-versus-1 kernel logistic regression models combined using pairwise coupling (KLR-PC), and multi-class kernel logistic regression (MKLR). Input selection was performed using two methods within the framework of Bayesian LS-SVM models. The basic input selection method was automatic relevance determination (ARD) [442], a ready extension of the Bayesian model formulation, whose performance was compared to a recently advocated method using rank-one updates for LS-SVMs [310] (see section 2.2.3 for details on these methods).

Comparison of input selection methods

We applied three input selection methods to all six 1-versus-1 classification problems on the training data: ARD using linear kernel (ARD-lin), ARD using RBF kernel (ARD-RBF), and the rank-one update method (R1U). Both ARD methods were applied to all 36 initial inputs, and the 16 with lowest weight were deleted. In the second and third ARD run, the worst 8 and 4 inputs were dropped respectively. The final run ranked the remaining 8 inputs. For R1U, the LOO-AUC (leave-one-out AUC) was used in each step to determine which input to add. In LOO cross-validation, each of the N patients serves once as a validation set. The LOO-AUC was the AUC on the N validation scores. The value of the hyperparameter γ was re-tuned in each step, by selecting from a grid the value that resulted in the best LOO-AUC performance.

The results of the three methods were compared as follows. Per binary problem, we considered 12 Bayesian LS-SVM models: we considered a Bayesian LS-SVM with linear kernel using inputs from either ARD-lin, ARD-RBF, or R1U, and a Bayesian LS-SVM with RBF kernel using inputs from ARD-RBF. For each of these four, we used the 3, 5, and 8 best inputs as determined by the respective input selection method, thus resulting in 12 models. The performance of these 12 models was determined using 20 runs of 5CV (R5CV) on the training set, yielding 100 validation AUCs per model. These AUCs were summarized by their average. The results for all binary problems are presented in Figure 5.10. Taken together, these plots clearly show that the R1U method gave the best results, even when compared to the ARD-RBF inputs. This is nicely demonstrated in the summary plot (Figure 5.11), where we combined, by averaging, the results for the binary problems.

Selecting the final input set for each binary problem

We need good input sets for all binary problems while keeping the overall number of inputs small. Therefore, we selected 16 promising inputs based on the initial R1U rankings, and ran R1U again to rank the 16 inputs per binary problem. Then, R5CV was used to determine which highest ranked inputs were to be chosen. This extensive analysis resulted in final input sets for each binary problem. Three to five inputs were selected per binary problem, resulting in 11 different inputs in total:

- benign vs primary invasive: age (a), ascites (b), shadows (c), solidmax (d), wallreg (e);

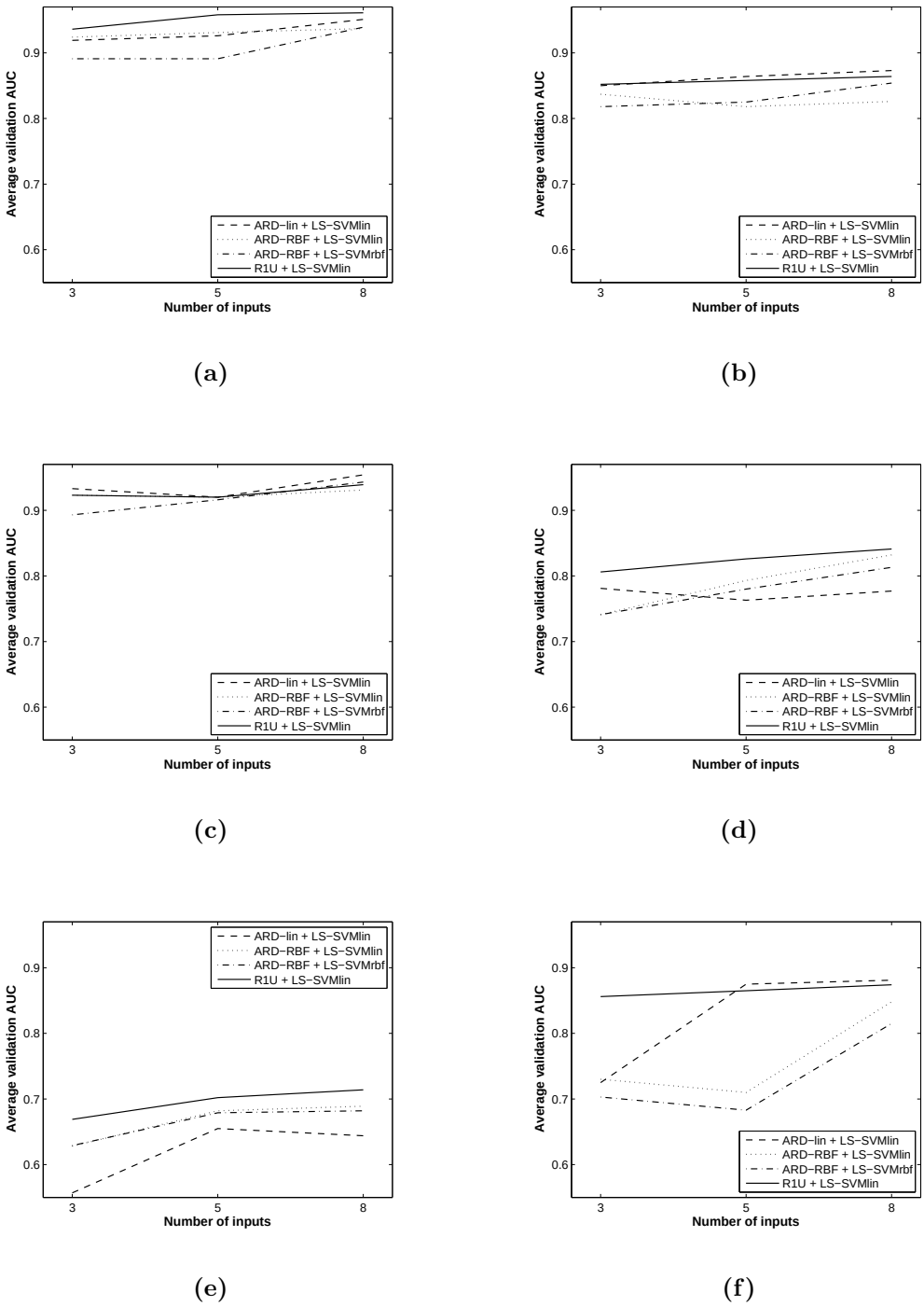


Figure 5.10: Comparison of the ARD and R1U methods for input selection for: (a) benign vs primary invasive, (b) benign vs borderline, (c) benign vs metastatic, (d) primary invasive vs borderline, (e) primary invasive vs metastatic, and (f) borderline vs metastatic.

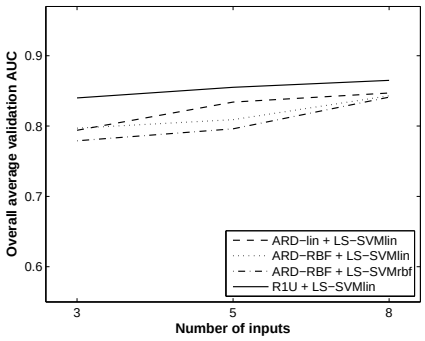


Figure 5.11: Comparison of the ARD and R1U methods for input selection averaged over all binary problems.

- benign vs borderline: age (a), pershistovca (f), lesmax (g), papflow (h), papnr (i);
- benign vs metastatic: ascites (b), solidmax (d), wallreg (e), locul5 (j);
- primary invasive vs borderline: age (a), ascites (b), solidmax (d), wallreg (e), bilateral (k);
- primary invasive vs metastatic: ascites (b), pershistovca (f), locul5 (j);
- borderline vs metastatic: ascites (b), solidmax (d), locul5 (j).

There is considerable overlap with the inputs selected for the LR-PC model, yet a few changes are noted. For example, the R1U inputs include shadows and papnr but omit locul1. Therefore, we decided to use the R1U inputs to create a second model based on the combination of binary logistic regression models (LR-PC2) to see how both input sets compare. These are the resulting six 1-versus-1 binary models:

$$\begin{aligned}
\log\left(\frac{\hat{\pi}_{pr}}{\hat{\pi}_{b}}\right) &= -7.01 + 0.037(a) + 1.65(b) - 5.43(c) + 0.086(d) + 1.96(e),\\
\log\left(\frac{\hat{\pi}_{bo}}{\hat{\pi}_{b}}\right) &= -6.71 + 0.017(a) + 2.79(f) + 0.026(g) + 1.93(h) + 0.36(i),\\
\log\left(\frac{\hat{\pi}_{me}}{\hat{\pi}_{b}}\right) &= -7.40 + 1.69(b) + 0.066(d) + 2.86(e) + 2.22(j),\\
\log\left(\frac{\hat{\pi}_{bo}}{\hat{\pi}_{pr}}\right) &= 3.01 - 0.027(a) - 1.32(b) - 0.046(d) - 0.46(e) - 1.38(k),\\
\log\left(\frac{\hat{\pi}_{me}}{\hat{\pi}_{pr}}\right) &= -1.84 - 0.73(b) + 3.33(f) + 1.41(j),\\
\log\left(\frac{\hat{\pi}_{me}}{\hat{\pi}_{bo}}\right) &= -2.39 + 1.43(b) + 0.032(d) + 2.43(j). \qquad (5.7)
\end{aligned}$$

Notice that the input selection for the advanced models used lesmax and solidmax (the original, unbounded versions), while we used the bounded versions when using the R1U inputs in LR-PC2. Differences with LR-PC are that LR-PC2 uses the observations that patients with a primary invasive tumor are less likely to have acoustic shadows than patients with a benign tumor, that patients with a borderline tumor are older, more often have a history of ovarian cancer, and have a tumor with more papillary structures than patients with a benign tumor, and that patients with a borderline tumor are younger and have less often a tumor with irregular internal cyst walls than patients with a primary invasive tumor.

A technical note is that the use of shadows in the binary model contrasting primary invasive tumors with benign tumors is problematic. The reason is that no patient with a primary invasive tumor in the training set has acoustic shadows. In logistic regression, such a zero count may lead to 'quasi-complete separation' of the data points such that the maximum likelihood estimate of the parameter β for shadows does not exist [80]. The conditions of complete and quasi-complete separation indicate that there is no overlap between two classes in the space of the model inputs, leading to computational difficulties [7, 80]. This is reflected in an inflated standard error of the parameter estimate. A solution for such situations is the use of hidden logistic regression, a method for which maximum likelihood estimates always exist [341]. This method makes a distinction between the observed response Y (e.g. 1 or 0) and the true response T, with a strong stochastic relationship between both. The relationship is expressed as follows: if $T = 1$ then $P(Y = 1) = \delta_1$, while if $T = 0$ then $P(Y = 1) = \delta_0$, with δ_1 large and δ_0 small. It follows that $P(Y = 1) = \delta_1$ if $y = 1$ is observed, while $P(Y = 1) = \delta_0$ if $y = 0$ is observed [341]. Thus, $P(Y = 1) = Y\delta_1 + (1 - Y)\delta_0$. $P(Y = 1)$ is called the pseudo-outcome $\tilde{Y}$, with $\tilde{y}_n = y_n\delta_1 + (1 - y_n)\delta_0$. Hidden logistic regression fits a logistic regression model to the pseudo-outcome. The likelihood of the model parameters θ given the pseudo-outcome is seen as an estimated likelihood (because the true outcome is unobservable), and the estimate of the model parameters $\hat{\theta}$ using maximum likelihood methodology is called the maximum estimated likelihood (MEL) estimate. As in [341], we assume $\delta_0 = \frac{\overline{\pi}\delta}{1+\delta}$ and $\delta_1 = \frac{1+\overline{\pi}\delta}{1+\delta}$, with $\delta = 0.01$ and $\overline{\pi}$ the bounded proportion of cases for which $y = 1$ is observed. $\overline{\pi}$ is bounded because it is fixed at δ or $1 - \delta$ if the true proportion is outside these limits.

The selected R1U inputs are used for BLSSVM-PC and KLR-PC, which are based on 1-versus-1 models. MKLR, an all-at-once method, is based on a single model. Therefore, MKLR used all 11 inputs that were selected by R1U for any of the binary problems. Thus, KLR-PC only uses inputs where they are deemed necessary, while MKLR uses all inputs to separate the four classes.

5.3.3 Model evaluation on the test set

This section presents the test set performance of the multi-class classifiers. The different classifiers will be compared with each other. Also, the underlying algorithms are compared using repeated stratified data-splitting.

Evaluation and comparison of the classifiers

Table 5.12 presents the test set performance of all six classifiers. The main criteria to evaluate performance were the AUCs per class, and the multi-class AUC metrics such as the VUS_2 or M. A 95% confidence interval (CI) on the M-index was computed using the bias-corrected bootstrap method based on 1000 bootstrap samples of the test set [141]. Another criterion was M_s, one of the weighted multi-class AUC metrics from Chapter 4. The table shows excellent performance of all models to identify benign and primary invasive tumors, with AUCs around 0.93 for both types of tumor. Performance for the smaller classes was also very good, with AUCs between 0.80 and 0.85 for borderline tumors and AUCs around 0.91 for metastatic tumors. Thus, the models are even helpful to detect tumors with elevated risk of being borderline. Differences in AUC were typically rather small, as can be seen in Figure 5.12 showing the test set ROC curves. Only for the borderline tumors, some differences in AUC levels were observed. However, since few tumors are borderline, the confidence intervals are wide. Prospective testing will be helpful in this regard. The multi-class AUC metrics pointed at LR-PC2 as the best multi-class classifier, suggesting that R1U is a valuable input selection tool. Yet, differences were again small, as suggested by the confidence intervals. This classifier also performed well when considering weighted multi-class AUC.

Table 5.12: Test set performance results of the multi-class classifiers.

	Binary AUCs				Multi-class AUC metrics		
Model	Benign	Pr. Inv.	Borderline	Metastatic	VUS_2	M (95% CI)	M_s
LR-PC	0.929	0.933	0.834	0.908	0.221	0.842 (.787 – .887)	0.283
LR-PC2	0.931	0.934	0.849	0.908	0.258	0.850 (.795 – .898)	0.298
MLR	0.928	0.933	0.798	0.912	0.186	0.826 (.771 – .878)	0.300
BLSSVM-PC	0.934	0.938	0.838	0.906	0.214	0.841 (.788 – .888)	0.291
KLR-PC	0.934	0.936	0.850	0.896	0.215	0.841 (.791 – .889)	0.267
MKLR	0.936	0.933	0.826	0.915	0.170	0.825 (.778 – .873)	0.279

The ranking analysis is shown in Table 5.13. Despite the small differences in M values, the ranking clearly suggested superiority of the LR-PC2 classifier, with an average rank of 1.5, an estimated probability of 0.64 of being the best of all classifiers, and an estimated probability of 0.86 to be one of the best two. The two worst models were MLR and MKLR, which are both 'all-at-once' classifiers. When focusing on the weighted index M_s, the best classifiers were LR-PC2 and MLR.

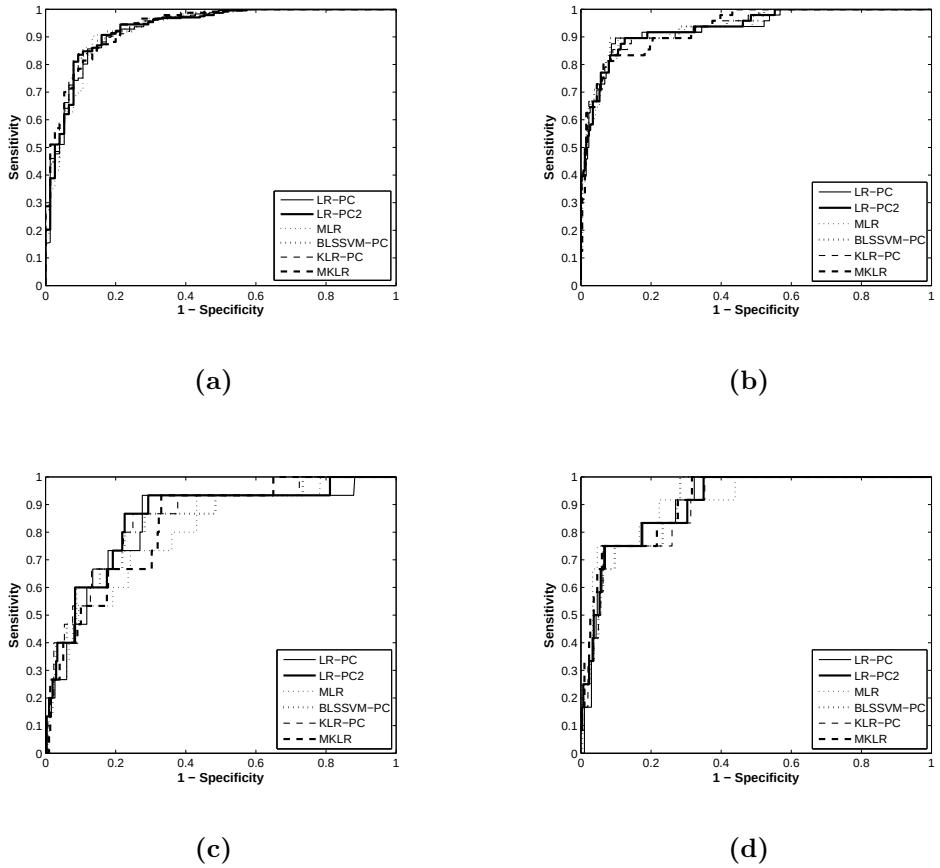


Figure 5.12: Test set ROC curves for (a) benign, (b) primary invasive, (c) borderline, and (d) metastatic tumors.

Table 5.13: Ranking of the mathematical multi-class models using bootstrap analysis.

	Ranking for M					Ranking for M_s			
Model	M	AR	Pm(1)	Pm(2)	Model	M_s	AR	Pm(1)	Pm(2)
LR-PC2	0.8501	1.5	.642	.863	MLR	0.300	1.9	.526	.730
LR-PC	0.8418	3.0	.113	.441	LR-PC2	0.298	1.8	.345	.873
KLR-PC	0.8415	2.9	.093	.325	BLSSVM-PC	0.291	2.8	.121	.325
BLSSVM-PC	0.8406	3.1	.143	.341	LR-PC	0.283	4.1	.003	.036
MLR	0.8265	5.1	.006	.017	MKLR	0.279	4.4	.005	.036
MKLR	0.8247	5.4	.003	.013	KLR-PC	0.267	5.9	.000	.000

KLR-PC performed poorest. Taken together, LR-PC2 had the best performance.

Since turning multi-class probabilistic classifiers into crisp predictions is not trivial, we decided to only develop a threshold system (on the training set) for LR-PC2. Several approaches can be adopted. For example, an automatic approach

Table 5.14: Cross-tabulation of true and predicted outcome using LR-PC2.

	Training set predictions (n, %)			Test set predictions (n, %)		
True class	Benign	Borderline	Invasive	Benign	Borderline	Invasive
Benign	439, 78.0%	47, 8.3%	77, 13.7%	179, 75.5%	26, 11.0%	32, 13.5%
Borderline	10, 25.0%	20, 50.0%	10, 25.0%	3, 20.0%	9, 60.0%	3, 20.0%
Invasive	7, 4.6%	27, 17.9%	117, 77.5%	3, 5.0%	10, 16.7%	47, 78.3%

could be to investigate for which class the probability has increased the most relative to the baseline probability which is taken to be the class proportions in the training set. Alternatively, one can order the classes and start with a threshold to distinguish between the first class and the rest. Next, a threshold is considered for the second class versus all other classes except the first, and so on. We decided to adopt the latter approach. Also, we combined the primary invasive and metastatic invasive tumors into one group of invasive malignant tumors. After all, the main goal of the multi-class analyses was to investigate how well mathematical models could identify borderline tumors. First, we chose to predict benignity when the probability of a benign tumor was at least 0.84. This threshold favored correct identification of malignant tumors over correct identification of benign tumors, as desired by the clinicians. Second, we chose to predict a borderline tumor when the probability of this tumor type was at least 0.12. This threshold favored correct identification of invasive malignant tumors over correct identification of borderline tumors. Taken together, these thresholds lead to the results in Table 5.14. In the test set, malignant tumors were correctly predicted in 92% of the cases (69/75). In the context of binary classification, this corresponds to the sensitivity. Benign, borderline, and invasive malignant tumors were correctly predicted in, respectively, 76, 60, and 78% of the cases. In the context of binary classification, the correction identification rate for benign tumors corresponds to the specificity.

Comparing algorithms

The algorithms rather than the classifiers are compared using 100 random splits of the data into a training and a test set. The splits were stratified for outcome. For each split, the six algorithms are trained on the training part and evaluated on the test part. The inputs used for each algorithm remained fixed. On the test part, the AUCs for each class were computed, as well as the multi-class AUC metrics VUS_2, M, and M_s. For each algorithm, the 100 test set results were summarized by taking their median. The results are presented in Table 5.15. A ranking analysis was performed for M and M_s using the 100 splits, the results of which are presented in Table 5.16. Using M as a multi-class measure of discriminatory ability, LR-PC and LR-PC2 were the best algorithms. When incorporating the actual probabilities suggested by the model, BLSSVM-PC was superior. Overall, however, LR-PC2 emerged as the best algorithm.

Table 5.15: Ranking of the six multi-class algorithms using 100 train-test splits.

	Median AUCs				Multi-class AUC metrics		
Algorithm	Benign	Pr. Inv.	Borderline	Metastatic	VUS_2	M	M_s
LR-PC	0.936	0.932	0.821	0.905	0.184	0.827	0.273
LR-PC2	0.936	0.932	0.821	0.900	0.181	0.831	0.278
MLR	0.933	0.922	0.800	0.899	0.149	0.813	0.281
BLSSVM-PC	0.926	0.931	0.793	0.899	0.152	0.812	0.290
KLR-PC	0.931	0.931	0.803	0.897	0.148	0.820	0.254
MKLR	0.930	0.926	0.801	0.895	0.135	0.805	0.261

Table 5.16: Ranking of the algorithms using 100 train-test splits.

	Ranking for M					Ranking for M_s			
Algorithm	M	AR	Pm(1)	Pm(2)	Model	M_s	AR	Pm(1)	Pm(2)
LR-PC2	0.831	1.7	.45	.88	BLSSVM-PC	0.290	1.6	.60	.87
LR-PC	0.827	2.1	.42	.73	MLR	0.281	2.4	.28	.63
KLR-PC	0.820	3.4	.06	.16	LR-PC2	0.278	2.6	.10	.42
MLR	0.813	4.0	.05	.14	LR-PC	0.273	3.7	.02	.07
BLSSVM-PC	0.812	4.5	.00	.03	MKLR	0.261	5.0	.00	.01
MKLR	0.805	5.2	.02	.06	KLR-PC	0.254	5.8	.00	.00

5.4 Discussion and conclusions

This chapter described binary and multi-class diagnostic models for adnexal masses. Such models use a set of measurements that are combined in some way to arrive at, in our case, probabilistic judgments about the nature of a condition. Interesting aspects of models are that they clarify relationships between a set of measurements and an outcome, that they make objective conclusions that may help human experts in reaching a diagnosis, and that sensitivity and specificity levels can be controlled to some extent by varying the threshold for translating the model's continuous output into a black-and-white, crisp prediction. Caveats are that good models need to be built in appropriate circumstances (appropriate in relation to the goals of the research), that models work in a rigid sense (which is the other side of the coin of a model's objectivity) as opposed to humans, and that it is not always clear how to arrive at a good crisp prediction, especially for multi-class models. These caveats show once more that models can assist yet not replace humans.

In our opinion, the first phase of the IOTA study provides good circumstances for model building. The data set is large and multi-centric, and data collection was standardized. This environment is likely to have influenced the development of robust and widely applicable models. Strictly spoken, in medical decision making, diagnostic models should be made separately for each clinical center and at regular points in time as characteristics of the patient population may differ and evolve over

time [180]. This is not practically feasible such that data sets such as that from IOTA phase I present good alternatives. It is necessary, however, to thoroughly evaluate the models using both prospective internal and prospective external validation. This is the subject of the next chapter. If possible, performance evaluation should be repeated regularly.

The linear and nonlinear classifiers that were developed on the IOTA phase I data set showed excellent performance on an independent test part of the data. Linear models are always good to start with, since simple methodologies tend to capture most of the diagnostic information available [180], and since they are preferred in case more flexible yet more complex methods such as support vector machines cannot improve performance. Note also that mathematically linear methods can be nonlinear with respect to the original measurements when these are transformed or are used in interaction or polynomial terms. Therefore, the general distinction between linear and nonlinear models is at times vague; we will not elaborate on this issue. The general observation is that the linear models, including logistic regression models but also perceptrons or support vector machines with linear kernel, perform similar to the nonlinear models. Of course, other complex methods could be developed on this data set, such as Bayesian networks [25], ensemble models based on bagging or boosting [312] or other fusion models as in [133], classification and regression trees [64], or the sequential non-uniform procedure presented in [363]. However, we do not think they will have much added value over the presented logistic regression models and LS-SVMs. For the multi-class models, we ended up without LS-SVM based models that used an RBF kernel. This was a result of the input selection analysis, which suggested that RBF kernels were not necessary. Also, no interaction terms or polynomial effects appeared useful in the multi-class logistic regression models. Thus, similar to binary classification, the multi-class classification of ovarian tumors does not entail a high level of nonlinearity.

For binary classification, there is no single classifier that outperformed the others when considering model complexity, AUC, and sAUC. The BLSSVMlin classifier had good performance, but a minor drawback is that the inputs are not optimal. Mainly the input representing whether or not the tumor is thought to be of ovarian origin (origin1) is undesirable as this is highly subjective. The color Doppler score (colscore) is also subjective, yet this variables appears to be more important. The use of the maximum diameter of the ovary (ovmax) is clinically not entirely logical, yet its inclusion is understandable because it is very highly related to the maximum diameter of the lesion. If another input set is desired, LR1 or BPER11 are good alternatives to BLSSVMlin. All classifiers outperformed the standard RMI model. For multi-class classification, the LR-PC2 classifier had the best performance. Also, its underlying algorithm had the best performance of all six algorithms.

Several interesting observations should be noted. First, the Bayesian LS-SVM models, both for binary and multi-class classification, produced high quality

probability estimates. This is clear from the results of the weighted AUC metrics and the calibration analysis. This is a promising result for LS-SVMs and their Bayesian approach. Second, the good discriminatory power of the multi-class classifiers for borderline tumors was unexpected. This is a promising result that needs prospective evaluation. Third, LR-PC2 performed better than LR-PC, although the differences were small when the underlying algorithms were compared. Nevertheless, this result lends support to the inputs selected using the R1U method. Fourth, the multi-class algorithms involving a combination of binary models had better discriminatory ability than their all-at-once counterparts (i.e. LR-PC versus MLR and KLR-PC versus MKLR). This supports the practice that uses inputs only where they are needed, even though a well regularized all-at-once algorithm should not suffer drastically from the use of a single input set. In medical applications, it can be an advantage to use a combination of binary models, as this may clarify the operation of a multi-class model for clinicians. We believe that clinicians' reluctancy to accept mathematical models whose operation is obscure is a healthy attitude. Note, however, that the weighted AUC metrics tended to give a small advantage to the all-at-once algorithms. In our opinion, this result is not convincing enough to overrule the result for the unweighted AUC metrics. A drawback in contrasting both types of algorithms is that the input selection analyses are subjective and may have an influence on the results. Fifth, the comparison of the neural network algorithms for binary classification suggested that tuning the number of hidden neurons for ARD input selection may positively influence the results.

It is clear that subjective assessment had excellent performance, and that the binary classifiers were not able to outperform the performance of the human experts. However, subjective assessment obtained higher specificity than sensitivity. This contrasts with the fact that clinicians prefer high sensitivity (i.e. high cancer identification rate). In a previous study, this is exactly what was obtained using subjective assessment [391]. In that sense, it is an asset of mathematical models that sensitivity and specificity levels are more easily controlled. Thresholds on the predicted probability of malignancy can be tuned so as to approximately obtain desired sensitivity and specificity levels. The results for the binary classifiers point at the problem of comparing classifiers based on sensitivity and specificity when the threshold level is determined by different researchers. This is for example the cause of the difference in specificity between LR1 and the Bayesian LS-SVM and RVM models. The statistically significant differences as obtained by McNemar's test should therefore not be considered too important. On top of this, optimal threshold values may well depend on the clinical center the model is used in, such that results based on thresholds should not be taken for granted. When applying the models to the IOTA phase Ib data set, we will investigate center differences in the optimal threshold.

Chapter 6

Prospective evaluation of diagnostic classifiers for ovarian tumors

Abstract – In this chapter, we will use the data from phases I, Ib and II of the IOTA study to prospectively test classifiers for ovarian tumors, both those developed on the IOTA phase I data set as those developed on other data and/or in other clinical centers. The data from IOTA phase I are used to evaluate existing models developed by other researchers. The data from IOTA phases Ib and II will mainly be used to evaluate classifiers developed on the phase I data.

6.1 IOTA phase I: prospective evaluation of old models

The first step of the work on prospective evaluation of classifiers for ovarian tumors dealt with the application of old and/or 'non-IOTA' models on the data from IOTA phase I [447]. The evaluated classifiers were based on scoring systems, logistic regression models, neural network models, least squares support vector machines, and relevance vector machines. Because this analysis focused on models that were not developed within the framework of this dissertation, we will discuss the results briefly.

6.1.1 The experimental setup

The classifiers selected for prospective evaluation were either classifiers that were developed at centers of members of the IOTA steering committee [212, 386, 396, 267, 268], or classifiers that were found using a literature search [203, 256, 399, 282, 329, 376]. Some models could not be used because the model formula was not found in the publication and the authors could not be contacted or could not find the necessary information, or because not all the inputs were available in the IOTA phase

I data set. This resulted in 17 models: three scoring systems [203, 399, 256] among which one was the widely used risk of malignancy index, seven logistic regression models [282, 329, 376, 386, 396, 268, 212], three multi-layer perceptrons [396, 268], three Bayesian LS-SVMs [267] and three RVMs [267]. The LS-SVMs and RVMs used the same inputs and were developed on the same data set. The inputs that are used by all of these models are listed in [447]. Models that could not be evaluated included four scoring systems [124, 153, 9, 348], one logistic regression model [8], and three neural networks [377, 49, 82]. Because some models use serum CA-125 levels, which are not available for all patients in the IOTA phase I data set, the classifiers were applied to the 809 patients for which this information was available.

Following the scope of this text, the AUC, pAUC, and sAUC were considered the most relevant performance measures. A ranking analysis similar to those presented in the previous chapter was performed for the AUC. No results involving thresholds are reported. These models were developed by different people which may have had different aims when seeking for a threshold value.

6.1.2 Results

The results are summarized in Table 6.1. AUCs ranged from 0.68 to 0.92. The RVM and LS-SVM models had best performance with AUCs between 0.91 and 0.92. The logistic regression model that was developed together with the LS-SVMs and RVMs, and that used the same inputs, also performed very well with an AUC of 0.91. The ranking analysis clearly shows that these five are the best of all 17 tested classifiers. Their average rank was markedly lower than that of any other model, as is the estimated probability of being among the best five classifiers. It should not come as a surprise that classifiers based on advanced methods such as LS-SVMs and RVMs perform well. Yet, since these models were developed using data from the University Hospitals Leuven, the center where the IOTA group was founded, it is possible that this may have had an additional positive influence on these models' performance. This could for example have happened due to similarities in data collection. Of course, the good performance may as well have been caused by the specific measurements that were included in the model building data set and by the input selection process performed on these measurements.

Finally, notice that RMI's performance was not impressive. The 17 classifiers from this section were not developed under optimal circumstances. For example, they were often developed on small to medium sized data sets (typically less than 200 and rarely higher than 400) that were collected at only one center. Often, only a limited number of measurements were taken that suffered from a lack of standardization. Therefore, the classifiers described in the previous chapter may still perform better.

Table 6.1: Prospective evaluation of old models on the IOTA phase I data set. The classifiers are ranked based on their AUC.

				Ranking results for AUC		
Model type	AUC	pAUC	sAUC	AR	Pm(3)	Pm(5)
RVM, RBF kernel [267]	0.923	0.193	0.623	1.0	1	1
RVM, lin. kernel [267]	0.921	0.191	0.616	2.0	1	1
Bay. LS-SVM, lin [267]	0.916	0.186	0.616	3.1	0.924	1
Bay. LS-SVM, RBF [267]	0.911	0.183	0.613	4.5	0.046	0.954
Log. Reg. [268]	0.910	0.181	0.632	4.8	0.000	0.912
Log. Reg. [212]	0.895	0.167	0.537	7.4	0.027	0.095
Log. Reg. [386]	0.894	0.171	0.611	7.5	0.000	0.021
Log. Reg. [396]	0.890	0.168	0.515	8.6	0.000	0.000
Bay. MLP [268]	0.890	0.172	0.541	8.9	0.001	0.006
MLP1 [396]	0.889	0.169	0.490	9.0	0.000	0.003
RMI scoring system [203]	0.883	0.174	–	10.0	0.002	0.009
MLP2 [396]	0.876	0.165	0.218	11.7	0.000	0.000
RMI2 scoring system [399]	0.860	0.172	–	13.2	0.000	0.000
Log. Reg. [376]	0.856	0.150	0.449	13.4	0.000	0.000
Log. Reg. [282]	0.813	0.105	0.315	15.1	0.000	0.000
Log. Reg. [329]	0.788	0.119	0.357	15.9	0.000	0.000
Scoring system [256]	0.676	0.063	–	17.0	0.000	0.000

Note: Since scoring systems do not output probabilities, sAUC could not be computed for these models.

6.2 IOTA phase Ib: prospective internal evaluation of IOTA models

6.2.1 Binary classifiers

Table 6.2 presents the performance of the binary models when applied to IOTA Ib. All models performed excellent, with LR1, LR2 and BLSSVMlin having the highest AUC. However, the difference between the highest and lowest AUC is only 0.006. These AUCs are higher than those obtained on the IOTA I test set. This may be caused by the fact that IOTAIb involves the three centers that contributed most to IOTA I. Yet, this does not explain why the RMI also performed markedly better. The ROC plots in Figure 6.1 show that, this time, the models lie closer to the performance of the expert operators' subjective assessment. Next to the AUC values, the pAUC and Sens75 values were also very good, with smaller differences compared to the IOTA I test set results. At a true negative rate of 75%, around 95% of the malignancies were identified by the models. The results for sAUC confirm that the Bayesian LS-SVMs yield good probabilities.

Three other observations are worth mentioning. First, the ROCCH classifier did not outperform the individual models, even though this classifier was developed on the IOTA I test set. Second, an interesting result is the similar performance of ObjLR and LR1, with the former using only objective variables. Third, large

Table 6.2: Performance of subjective assessment and mathematical models as assessed on the IOTA Ib data set.

Model	AUC (95% CI)	AUC_{CA}	pAUC	Sens75	sAUC	Sens Spec LR+ LR-
Subj. ass.						90.2 – 92.9 – 12.6 – 0.11
LR1	.950 (.929 – .967)	.940	.207	95.1	.631	95.1 – 74.2 – 3.68 – 0.07
LR2	.950 (.928 – .967)	.945	.208	95.1	.605	95.1 – 73.9 – 3.64 – 0.07
Reg. LR	.944 (.935 – .969)	.934	.203	95.8	.459	95.1 – 77.5 – 4.22 – 0.06
Obj. LR	.948 (.931 – .969)	.940	.205	96.5	.628	95.1 – 78.0 – 4.33 – 0.06
BMLP11-2a	.948 (.928 – .964)	.939	.204	94.4	.589	93.7 – 76.9 – 4.06 – 0.08
BMLP11-2b	.947 (.926 – .966)	.940	.205	95.8	.588	95.1 – 78.8 – 4.50 – 0.06
BPER11	.946 (.923 – .963)	.939	.206	95.8	.571	95.8 – 78.3 – 4.41 – 0.05
BLSSVMlin	.950 (.931 – .967)	.946	.207	95.1	.666	91.6 – 83.8 – 5.65 – 0.10
BLSSVMrbf	.949 (.930 – .966)	.946	.207	95.8	.679	91.6 – 83.2 – 5.47 – 0.10
BLSSVMaddrbf	.948 (.929 – .963)	.943	.202	95.1	.672	93.7 – 81.9 – 5.17 – 0.08
RVMlin	.948 (.926 – .965)	.943	.207	95.1	.628	91.6 – 82.1 – 5.13 – 0.10
RVMrbf	.945 (.923 – .964)	.941	.205	93.7	.635	92.3 – 80.2 – 4.67 – 0.10
RVMaddrbf	.947 (.927 – .964)	.943	.203	95.1	.634	93.7 – 81.3 – 5.02 – 0.08
RMI		.910				
ROCCH classifier	.947	.940	.203	97.9		

AUC_{CA}: AUC using only the patients for whom CA-125 levels were determined.

differences in performance between centers were obtained (results not shown). The classifiers performed near perfect in Malmö (AUCs between 0.98 and 0.99), excellent in Leuven (AUCs between 0.95 and 0.96), and good in Rome (AUCs between 0.88 and 0.91). AUC differences between Malmo and Rome were around 0.1.

The ranking results for the AUC, shown in Table 6.3, reveal that RVMlin was not the best classifier for the IOTA Ib data as opposed to the IOTA I test set data. In general, the ranking results reflect the minor differences between the classifiers. For example, the best classifier had an average rank of 4.2. The ranking for sAUC shows again that the Bayesian LS-SVM models provided very good probabilistic output.

The AUC results for the old models, as described in the previous section, were not as good as those for the IOTA models (results not shown). The old models had AUCs below 0.9, except the Bayesian LS-SVMs and RVMs that had AUCs between 0.93 and 0.94. Note that IOTA Ib involved three important centers from IOTA I, which may be a disadvantage for the old models. Interestingly, IOTA II involves totally new centers, such that the comparison of old models and IOTA models would be interesting for these centers.

An interesting question regarding thresholds to arrive at crisp predictions is whether they are highly center dependent. To tentatively answer this question, we plotted the weighted average of sensitivity and specificity versus the threshold value for the LR1 and BLSSVMlin models (Figure 6.2). Sensitivity and specificity were

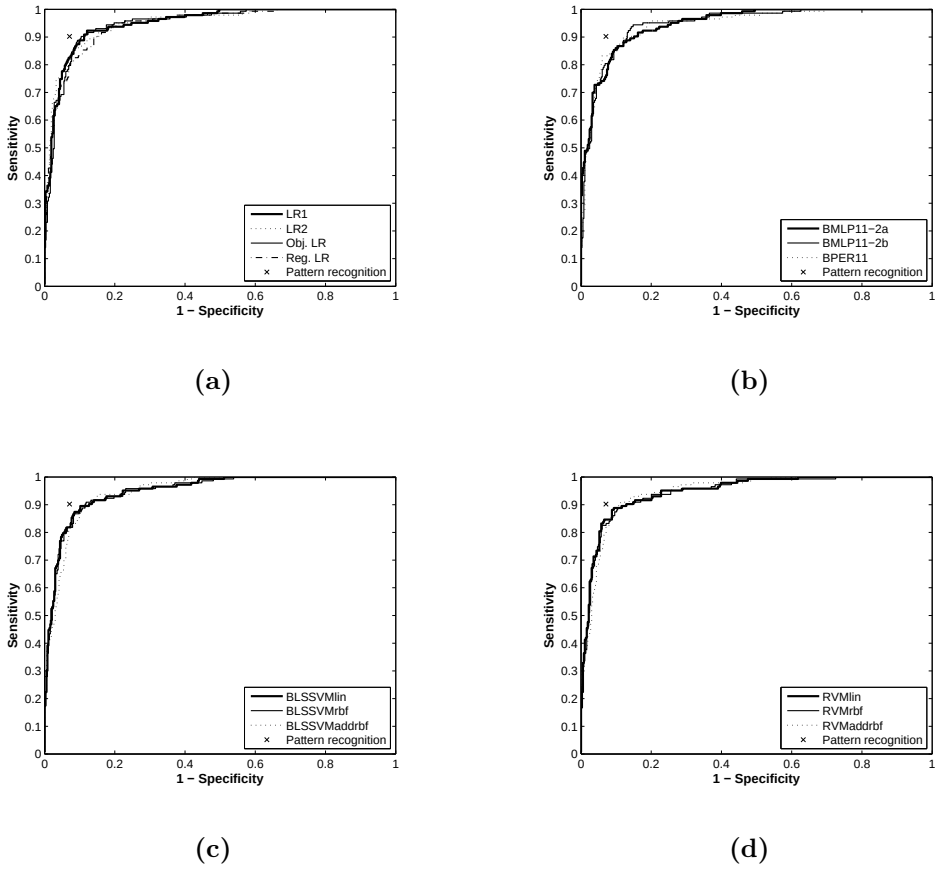


Figure 6.1: ROC curves of the mathematical models on the IOTA phase Ib database.

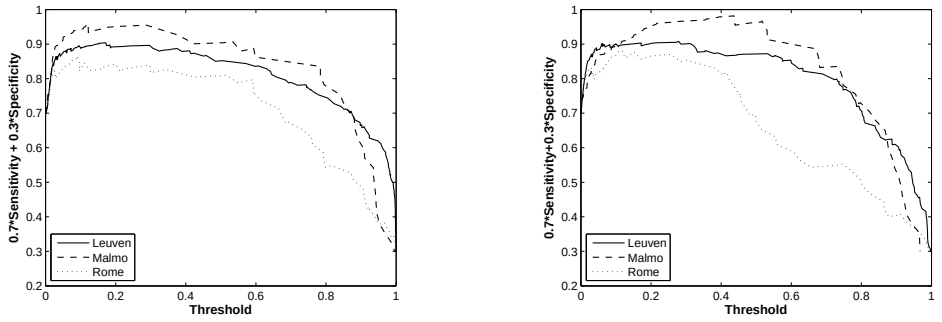
given the weights 0.7 and 0.3, respectively, to reflect the priority of sensitivity. Center differences regarding the optimal threshold appeared to depend on the classifier. LR1 had near perfect performance in Malmö when a threshold between 0.40 and 0.45 is used. In both other centers, the optimal threshold was much lower. For BLSSVMlin, the optimal threshold appeared similar for all centers.

It is necessary to investigate the quality of a classifier's probabilities. This is often done using calibration curves. As discussed in section 2.3.4, we used nonparametric regression to check the classifiers calibration. The IOTA phase I test set and the IOTA phase Ib data set were merged (n=819) for the construction of the calibration curves (see Figure 6.3). Combining the two data sets reduces the output, yet note that the IOTA I test set does not contain prospectively collected data. The curves show that most models have excellent calibration, with the Bayesian LS-SVM models having the nicest curves. The RegLR classifier performed worst. The shrinkage appeared too harsh, as was also suggested by the sAUC results.

The discriminatory potential of the models and of subjective assessment were

Table 6.3: Ranking of the mathematical models using bootstrap analysis.

	Ranking for AUC					Ranking for sAUC			
Model	AUC	AR	Pm(1)	Pm(3)	Model	sAUC	AR	Pm(1)	Pm(5)
BLSSVMlin	0.9501	4.2	.136	.484	BLSSVMrbf	0.6793	1.2	.818	1
LR1	0.9500	4.8	.138	.407	BLSSVMaddrbf	0.6720	2.1	.182	1
LR2	0.9497	5.0	.221	.426	BLSSVMlin	0.6664	2.7	.000	1
BLSSVMrbf	0.9495	5.1	.053	.322	RVMrbf	0.6352	5.1	.000	.713
BMLP11-2a	0.9479	7.0	.133	.270	RVMaddrbf	0.6338	5.6	.000	.535
BLSSVMaddrbf	0.9478	7.0	.074	.198	LR1	0.6313	5.9	.000	.416
RVMlin	0.9477	7.2	.028	.127	RVMlin	0.6280	6.9	.000	.098
Obj. LR	0.9476	7.0	.032	.186	Obj. LR	0.6275	6.6	.000	.237
BMLP11-2b	0.9474	7.3	.099	.240	LR2	0.6048	9.2	.000	.001
RVMaddrbf	0.9471	7.9	.023	.112	BMLP11-2a	0.5888	10.4	.000	.000
BPER11	0.9455	8.7	.047	.141	BMLP11-2b	0.5876	10.5	.000	.000
RVMrbf	0.9454	9.6	.001	.024	BPER11	0.5711	11.9	.000	.000
Reg. LR	0.9439	10.2	.013	.049	Reg. LR	0.4593	13.0	.000	.000



(a) **(b)**

Figure 6.2: Threshold curves showing weighted average of sensitivity and specificity by threshold for (a) BLSSVMlin and (b) LR1.

also investigated for different levels of diagnostic confidence of the expert operator. The operator had to indicate his confidence with respect to the outcome on a 5-point scale: certainly benign, probably benign, uncertain, probably malignant, and certainly malignant. Again, to reduce output, the IOTA I test data set and IOTA Ib data set were merged for this analysis. The results are presented in Table 6.4. The classifiers performed very good when the operator is certain, with AUCs around 0.98. Using the thresholds for the classifiers, similar sensitivity levels were obtained relative to subjective assessment, but subjective assessment was superior with respect to specificity. When the operator was quasi-certain, the models still performed very good, with AUCs between 0.90 and 0.92. Now, the classifiers had better sensitivity than subjective assessment, but specificity was clearly worse. This

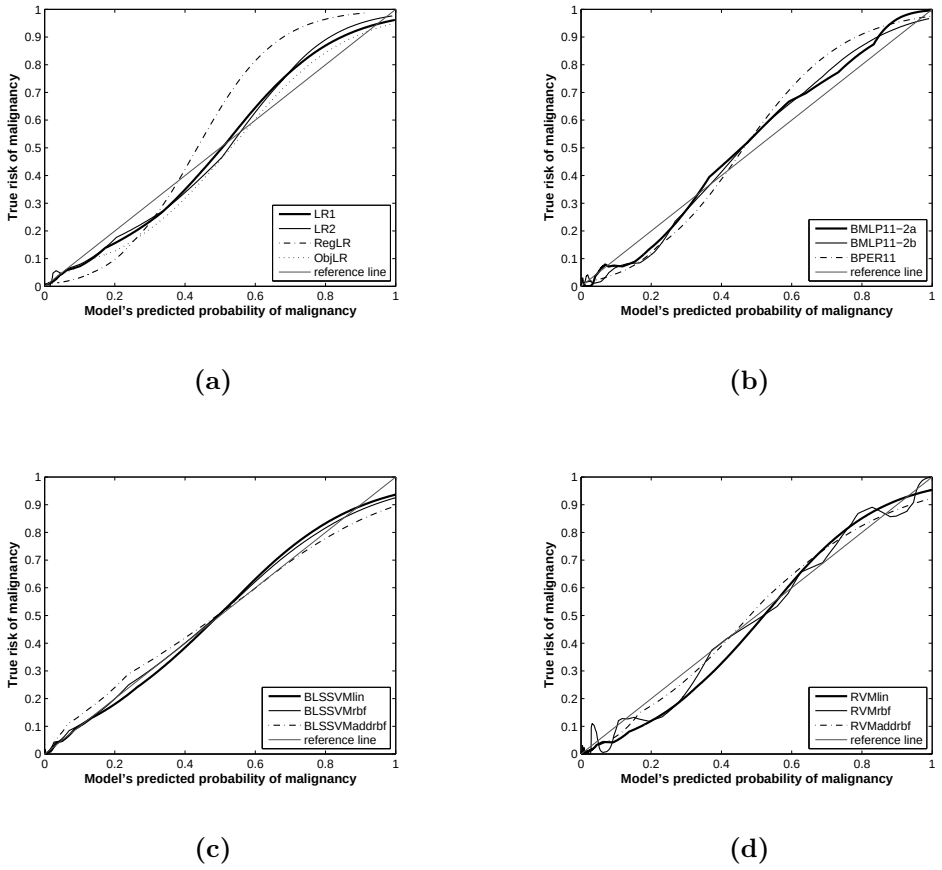


Figure 6.3: Calibration curves of the developed classifiers using the IOTA I test set and the IOTA phase Ib data set.

polarization was even stronger when the operator was uncertain. In this situation, the models still showed some discriminatory ability, with AUCs between 0.67 and 0.75. The results further demonstrate the high sensitivities obtained by applying the mathematical models and their thresholds, whereas the expert operator obtained higher specificity levels. Even though the AUCs are not directly comparable because not all patients had their CA-125 level measured, the results suggest that the new mathematical models performed better than the RMI irrespective of diagnostic confidence of the expert operator.

To end the IOTA Ib analysis for binary models, the performance of the simple rules are discussed (Table 6.5). There were 386 patients (76.1%) for which the simple rules resulted in a prediction. For these cases, sensitivity and specificity were 94.7% and 90.9%, respectively.

Table 6.4: Performance of subjective assessment and mathematical models by diagnostic confidence of the expert operator, using the IOTA I test data and the IOTA Ib data set.

	Certainly benign or malignant (n=428)		Probably benign or malignant (n=320)		Uncertain (n=71)	
Model	AUC	Sens Spec	AUC	Sens Spec	AUC	Sens Spec
Subj. ass.		98.1 – 98.8		85.4 – 91.3		60.9 – 64.6
LR1	0.984	97.2 – 87.3	0.920	94.4 – 67.1	0.718	82.6 – 27.1
LR2	0.980	96.2 – 86.0	0.905	92.1 – 64.9	0.739	82.6 – 33.3
Reg. LR	0.983	97.2 – 89.4	0.900	92.1 – 67.5	0.715	78.3 – 45.8
Obj. LR	0.981	96.2 – 89.8	0.916	93.3 – 68.8	0.721	82.6 – 33.3
BMLP11-2a	0.982	96.2 – 90.7	0.915	93.3 – 68.8	0.688	78.3 – 41.7
BMLP11-2b	0.980	97.2 – 88.8	0.925	95.5 – 70.6	0.671	73.9 – 31.3
BPER11	0.983	99.1 – 89.1	0.922	95.5 – 71.0	0.691	78.3 – 29.2
BLSSVMlin	0.982	95.3 – 95.0	0.920	89.9 – 79.2	0.743	78.3 – 47.9
BLSSVMrbf	0.981	95.3 – 93.2	0.920	92.1 – 75.8	0.740	78.3 – 41.7
BLSSVMaddrbf	0.982	97.2 – 93.2	0.918	91.0 – 74.5	0.730	78.3 – 43.8
RVMlin	0.983	96.2 – 93.8	0.916	89.9 – 74.9	0.746	78.3 – 39.6
RVMrbf	0.981	96.2 – 91.9	0.914	92.1 – 73.2	0.745	78.3 – 39.6
RVMaddrbf	0.981	97.2 – 92.9	0.922	92.5 – 77.1	0.732	82.6 – 43.8
RMI	0.947		0.864		0.616	

Note: for the RMI, only the cases with CA-125 levels were used. Therefore, the results for the RMI cannot be readily compared to those of other models.

Table 6.5: Applicability of M-rules and B-rules to the IOTA phase Ib data set.

		Outcome			
Situation	N	Benign	Primary invasive	Borderline	Metastatic
Prediction benign	255	249, 97.6%	5, 2.0%	1, 0.4%	0, 0.0%
0 M-rules, 1 B-rule	144	139, 96.5%	4, 2.8%	1, 0.7%	0, 0.0%
0 M-rules, 2 B-rules	98	97, 99.0%	1, 1.0%	0, 0.0%	0, 0.0%
0 M-rules, >2 B-rules	13	13, 100%	0, 0.0%	0, 0.0%	0, 0.0%
Prediction malignant	131	25, 19.1%	78, 59.5%	14, 10.7%	14, 10.7%
1 M-rule, 0 B-rules	62	19, 30.6%	30, 48.4%	6, 9.7%	7, 11.3%
2 M-rules, 0 B-rules	46	5, 10.9%	27, 58.7%	8, 17.4%	6, 13.0%
>2 M-rules, 0 B-rules	23	1, 4.3%	21, 91.3%	0, 0.0%	1, 4.3%
Undecided	116	85, 73.3%	20, 17.2%	5, 4.3%	6, 5.2%
Both types, undecided	18	10, 55.6%	4, 22.2%	2, 11.1%	2, 11.1%
No rules	98	75, 76.5%	16, 16.3%	3, 3.1%	4, 4.1%
Inconsistent	5	5, 100%	0, 0.0%	0, 0.0%	0, 0.0%
Both types, mostly B-rules	5	5, 100%	0, 0.0%	0, 0.0%	0, 0.0%
Both types, mostly M-rules	0				

6.2.2 Multi-class classifiers

We were happy to observe that the multi-class models performed very well when they were applied to IOTA Ib (Table 6.6 and Figure 6.4). For benign tumors, the

AUCs were around 0.94-0.95, which was similar to the AUCs for the binary models. For primary invasive tumors, the AUCs lied around 0.91-0.92, for metastatic tumors around 0.88-0.90. Borderline tumors, forming the most interesting class, obtained AUCs between 0.78-0.85. Thus, performance hardly decreased when compared to the IOTA I test set performance. Notice that the M index has mildly decreased, but binary and multi-class AUC metrics do not correspond in a linear sense [135]. For example, the AUC/VUS_2 values for perfect binary/multi-class classifiers are 1, but a random binary classifier has AUC 0.5, a random three-class classifier has VUS_2 1/6 and a random four-class classifier has VUS_2 1/24.

Table 6.6: Performance of the multi-class classifiers on the IOTA phase Ib database.

	Binary AUCs				Multi-class AUC metrics		
Model	Benign	Pr. Inv.	Borderline	Metastatic	VUS_2	M (95% CI)	M_s
LR-PC	0.941	0.919	0.782	0.884	0.154	0.808 (.765 – .848)	0.253
LR-PC2	0.946	0.918	0.827	0.881	0.187	0.823 (.779 – .860)	0.259
MLR	0.947	0.919	0.819	0.885	0.152	0.816 (.779 – .856)	0.266
BLSSVM-PC	0.938	0.916	0.834	0.886	0.147	0.819 (.777 – .856)	0.258
KLR-PC	0.938	0.914	0.829	0.891	0.161	0.823 (.783 – .860)	0.231
MKLR	0.948	0.918	0.846	0.904	0.157	0.824 (.789 – .858)	0.255

The ranking analysis for M placed three models on top with similar performance: MKLR, KLR-PC and LR-PC2 (Table 6.7). It is remarkable that MKLR now performed well relative to the other classifiers, while it performed relatively poor on the IOTA I test set. The ranking for M_s is very similar to that obtained on the IOTA I test set. Hence, the IOTA Ib analysis also suggested that LR-PC2 is the best multi-class classifier. Using the thresholds derived in the previous chapter, LR-PC2 identified 76% of the benign tumors, 55% of the borderline tumors, and 71% of the invasive tumors (primary invasive and metastatic) (Table 6.8). Misclassifications mainly dealt with borderline tumors that were predicted to be invasive and vice versa. Such misclassifications do not impair this model's sensitivity of malignancy, which was 95%.

Table 6.7: Ranking of the mathematical multi-class models using bootstrap analysis.

	Ranking for M					Ranking for M_s			
Model	M	AR	Pm(1)	Pm(2)	Model	M_s	AR	Pm(1)	Pm(2)
MKLR	0.8236	2.6	.335	.509	MLR	0.266	1.3	.793	.947
KLR-PC	0.8235	2.5	.211	.569	LR-PC2	0.259	2.6	.065	.514
LR-PC2	0.8226	2.8	.267	.441	BLSSVM-PC	0.258	3.1	.129	.366
BLSSVM-PC	0.8193	3.7	.079	.210	MKLR	0.255	3.7	.013	.149
MLR	0.8163	4.0	.098	.238	LR-PC	0.253	4.4	.000	.024
LR-PC	0.8076	5.5	.010	.033	KLR-PC	0.231	6.0	.000	.000

The calibration curves (Figure 6.5) were constructed after combining the IOTA I test data and the IOTA Ib data set (n=819). The curves were very good for benign

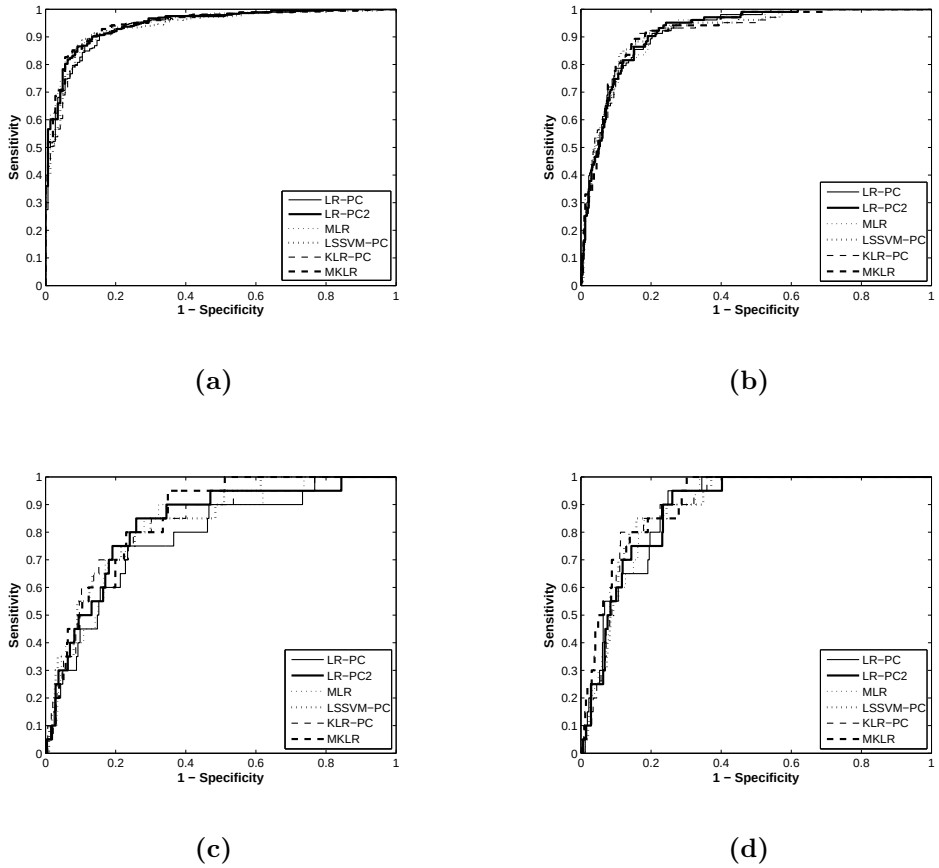


Figure 6.4: Test set ROC curves for (a) benign, (b) primary invasive, (c) borderline, and (d) metastatic tumors.

Table 6.8: Cross-tabulation of true and predicted outcome using LR-PC2.

	Predictions (n, %)		
True class	Benign	Borderline	Invasive
Benign	276, 75.8%	36, 9.9%	52, 14.3%
Borderline	2, 10.0%	11, 55.0%	7, 35.0%
Invasive	5, 4.1%	31, 25.2%	87, 70.7%

tumors and primary invasive tumors. Systematic deviations appeared to occur for borderline and metastatic tumors, suggesting that estimated probabilities of more than 0.20 for these classes were overestimated. Keep in mind, however, that not many instances of these classes were present, and that the low prevalence of these classes yield high predicted probabilities unlikely. For example, for LR-PC2 6.7% of the cases had an estimated probability for a borderline tumor above 0.20. The estimated probability for a metastatic tumor exceeded 0.20 in 6.6% of the cases.

Thus, the curves are not reliable for higher estimated probabilities.

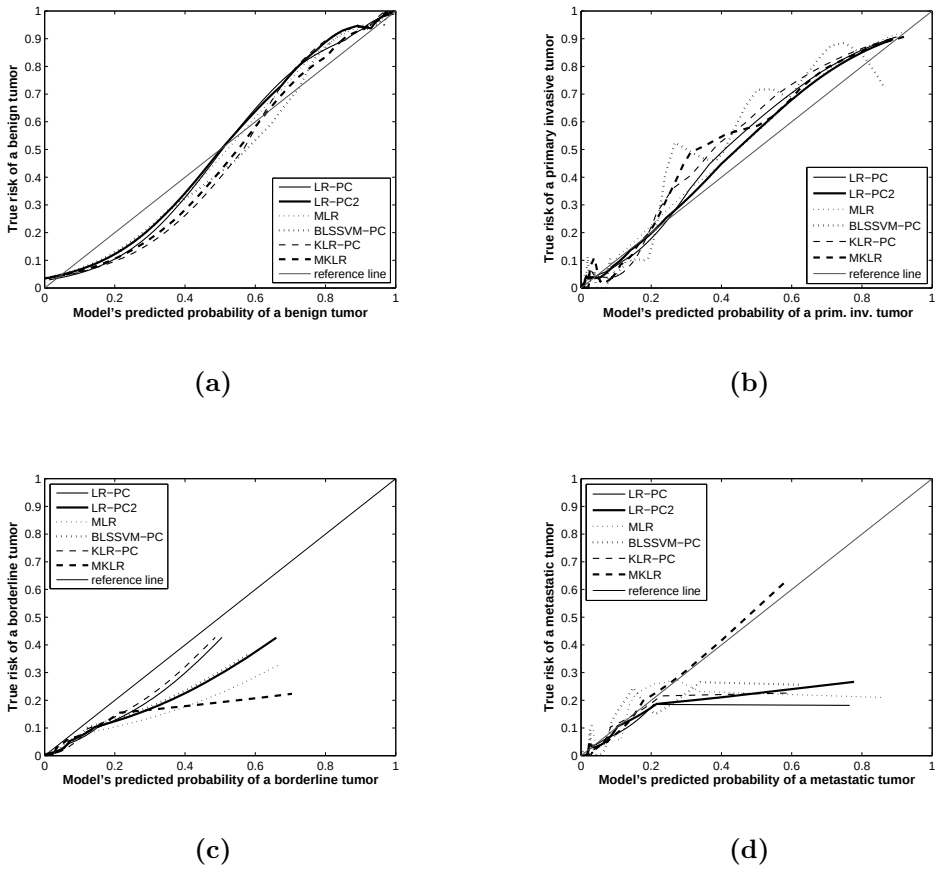


Figure 6.5: Calibration curves of the multi-class classifiers on the IOTA I test data and the IOTA Ib data set; curves for the probability of (a) a benign tumor, (b) a primary invasive tumor, (c) a borderline tumor, and (d) a metastatic tumor.

6.3 IOTA phase II: prospective internal and external validation

As mentioned in Chapter 3, there is at this moment only a preliminary IOTA phase II data set. Data collection ended in November 2007, and data cleaning is still in progress. Therefore, we will only perform a limited analysis using AUCs. The diagnostic models for ovarian tumors underwent prospective internal validation on the 993 tumors from seven centers that also contributed to IOTA phase I. A prospective external validation was performed using the 923 tumors from 11 newly recruited centers. The evaluation of the binary models is given in Table 6.9. For the Bayesian LS-SVMs and RVMs, only results for the models using linear kernel

are shown to reduce the output. Using all 1916 patients, the models have an AUC between 0.93 (LR2) and 0.95 (LR1, BPER11, BLSSVMrbf), which is very good. For the prospective internal validation, the AUCs were between 0.91 (LR2) and 0.94 (LR1, BMLP11-2a, BPER11). The AUCs for the prospective external validation were even higher, and ranged from 0.94 (BMLP11-2a) to 0.96 (LR1, BMLP11-2b, BLSSVMrbf, RVMrbf). All models performed well, yet LR1 and BPER11 had marginally better AUCs. Interesting to note is that all models performed good to excellent for all centers. Model performance was least good in Bologna (BIT) and King's College London (KUK).

Subjective assessment obtained 92.2% sensitivity and 93.0% specificity for prospective internal evaluation. Using thresholds, LR1 obtained 92.9% and 79.9%, respectively. BPER11 achieved 93.4% and 87.4%, respectively. For prospective external validation, subjective assessment had 89.4% sensitivity and 93.1% specificity. LR1 had 93.4% and 87.4%, respectively. BPER11 obtained 91.5% and 89.5%, respectively.

Table 6.9: Preliminary IOTA phase II results concerning a prospective internal and external validation of binary diagnostic models.

	AUC for							
Data	LR1	LR2	RegLR	ObjLR	BMLP11-2a	BPER11	BLSSVMlin	RVMlin
All data (n=1916)	.949	.932	.944	.943	.944	.948	.943	.943
Internal (n=993)	.941	.914	.933	.932	.943	.942	.931	.933
OIT (n=246)	.939	.907	.943	.920	.950	.936	.921	.933
LBE (n=228)	.953	.932	.944	.950	.952	.958	.941	.944
RIT (n=210)	.925	.891	.906	.905	.930	.932	.899	.897
MSW (n=136)	.977	.969	.981	.968	.973	.968	.974	.974
KUK (n=65)	.837	.817	.828	.836	.822	.841	.822	.824
NIT (n=63)	.954	.935	.943	.928	.946	.922	.952	.951
MIT (n=45)	.976	.982	.927	.963	.963	.970	.982	.982
External (n=923)	.957	.950	.954	.955	.944	.954	.955	.953
GBE (n=187)	.976	.974	.974	.968	.964	.982	.962	.960
SIT (n=154)	.972	.932	.959	.943	.953	.949	.969	.968
LPO (n=149)	.941	.932	.944	.940	.950	.941	.935	.939
BIT (n=132)	.902	.905	.905	.902	.840	.912	.911	.901
PCR (n=90)	.953	.958	.955	.964	.952	.953	.946	.938
CIT (n=71)	.944	.940	.928	.967	.919	.955	.943	.934
BCH (n=66)	.945	.950	.955	.954	.924	.945	.952	.950
LSW (n=38)	1	.948	.984	.995	.979	.995	1	1
UDI (n=17)	1	1	1	1	1	1	1	1
VIT (n=11)†	–	–	–	–	–	–	–	–
GIT (n=8)	.933	.933	.867	.933	.800	.933	.933	.933

† All patients from VIT had a benign tumor.

The evaluation of the multi-class models is presented in Table 6.10. Except for metastatic tumors, performance was again better for prospective external relative

to prospective internal validation. For benign tumors, AUCs were around 0.93 for prospective internal validation and varied between 0.94 and 0.96 (LR-PC2) for prospective external validation. These values are similar to those for the binary models. For primary invasive tumors, AUCs were between 0.90 and 0.92 (MLR) for prospective internal validation, and between 0.93 and 0.95 (LR-PC2) for prospective external validation. For borderline tumors, AUCs were between 0.79 and 0.84 (LR-PC, LR-PC2) for prospective internal validation, and between 0.87 and 0.92 (LR-PC2) for prospective external validation. For metastatic tumors, AUCs were between 0.84 and 0.86 (LR-PC, LR-PC2, BLSSVM-PC, MKLR) for prospective internal validation, and between 0.82 and 0.85 (LR-PC2) for prospective external validation. Overall, LR-PC2 again had the best performance as can also be concluded from the M values.

Table 6.10: Preliminary IOTA phase II results concerning a prospective internal and external validation of multi-class diagnostic models.

	AUCs				
	Benign	Prim. Inv.	Borderline	Metastatic	M
All data (n=1916)					
LR-PC	.937	.927	.856	.852	.816
LR-PC2	.945	.930	.872	.854	.825
MLR	.944	.930	.846	.833	.800
BLSSVM-PC	.936	.922	.856	.850	.807
KLR-PC	.932	.915	.859	.849	.810
MKLR	.936	.920	.830	.849	.792
Internal (n=993)					
LR-PC	.929	.911	.844	.861	.815
LR-PC2	.934	.913	.840	.857	.814
MLR	.934	.916	.812	.844	.799
BLSSVM-PC	.929	.911	.826	.855	.800
KLR-PC	.925	.902	.832	.854	.804
MKLR	.929	.910	.791	.861	.790
External (n=923)					
LR-PC	.945	.943	.868	.839	.813
LR-PC2	.956	.946	.918	.851	.839
MLR	.954	.944	.895	.821	.803
LSSVM-PC	.943	.933	.903	.844	.819
KLR-PC	.939	.929	.897	.842	.819
MKLR	.945	.930	.890	.836	.796

The threshold system for LR-PC2 resulted in a correct identification of 92.5% of malignant tumors as malignant, and correctly identified 83.5% of benign, 56.4 of borderline, and 75.3% of invasive malignant tumors. For prospective internal validation, 79.6% of benign, 54.6% of borderline, and 73.5% of invasive tumors were correctly identified. For prospective external validation, we obtained, 87.4%, 60.0%, and 77.5%, respectively.

Finally, we also used the IOTA II data set to investigate the performance of subjective assessment and mathematical models per tumor histology. To this end, the data from the IOTA I test set, IOTA Ib, and IOTA II were merged. This resulted in a data set with information on 2735 tumors in order to increase the number of cases with rare histologies. In Table 6.11, the average probability of malignancy given by the binary classifiers LR1 and BLSSVMlin is shown per tumor histology. The number of tumors that are correctly classified as benign or malignant is presented for both classifiers (using their threshold) and for subjective assessment. For the multi-class classifier LR-PC2, the average probability for each of the four outcomes is given.

Table 6.11: Results for subjective assessment and mathematical models by histology using the IOTA I test data, the IOTA Ib data, and the IOTA II data.

		Average prob.		N Correct			LR-PC2: average prob. for			
			BLSSVM	Subj.		BLSSVM		Prim.	Border	Meta
Histology	N	LR1	lin	ass.	LR1	lin	Benign	Inv.	line	static
Endometrioma	569	0.042	0.048	551	524	529	0.944	0.023	0.026	0.007
Dermoid/teratoma	325	0.033	0.049	319	301	305	0.927	0.029	0.028	0.016
Simple cyst	205	0.037	0.042	204	189	196	0.943	0.020	0.032	0.005
Functional cyst	96	0.051	0.049	91	84	89	0.934	0.025	0.033	0.009
Hydrosalpinx	66	0.077	0.044	65	54	62	0.918	0.035	0.039	0.009
Per. pseudocyst	19	0.045	0.020	19	18	19	0.930	0.028	0.037	0.005
Abscess	35	0.201	0.184	25	17	20	0.758	0.137	0.061	0.045
Fibroma	132	0.350	0.323	108	39	59	0.674	0.185	0.046	0.094
Cystadenoma	317	0.107	0.105	279	231	267	0.869	0.060	0.063	0.009
Muc. cystadenoma	198	0.101	0.105	174	142	162	0.873	0.046	0.073	0.009
Rare benign tumor	33	0.238	0.235	23	16	21	0.731	0.158	0.044	0.067
Primary invasive										
Stage I	106	0.666	0.714	96	103	97	0.384	0.413	0.124	0.079
Stage II	35	0.746	0.769	33	35	34	0.343	0.450	0.081	0.125
Stage III	274	0.751	0.767	258	264	258	0.295	0.510	0.069	0.126
Stage IV	48	0.791	0.798	48	48	46	0.253	0.541	0.053	0.153
Borderline										
Stage I-II	123	0.450	0.449	97	100	85	0.555	0.236	0.171	0.038
Stage III-IV	13	0.585	0.613	12	12	11	0.426	0.323	0.153	0.099
Metastatic	88	0.681	0.700	84	84	80	0.382	0.402	0.063	0.153
Rare primary inv.	53	0.621	0.616	40	46	45	0.474	0.336	0.063	0.127

A first observation is that BLSSVMlin yielded better probabilities (generally lower for benign cases and higher for malignant cases) relative to LR1, which was also suggested by the sAUC results. Apparently, fibromas, rare benign tumors, abscesses, and to a lesser extend cystadenomas and mucinous cystadenomas, are difficult benign histologies for the mathematical models. Considering the results for LR-PC2, fibromas and rare benign tumors were sometimes confused with metastatic or primary invasive tumors, abscesses with primary invasive tumors,

and (mucinous) cystadenomas were at times mildly confused with borderline tumors. Not surprisingly, borderline tumors got the lowest predicted probabilities of malignancy, even those of Stage III or IV. These tumors were sometimes confused with benign tumors. Primary invasive stage I tumors were given much higher probabilities. The probability of being benign or borderline decreased with increasing Stage of a primary invasive tumor, whereas the probability of being primary invasive or metastatic increased with Stage.

6.4 Conclusions

The main conclusion of this chapter is that both the binary and the multi-class diagnostic models performed very well on prospective evaluation. Very promising are the excellent results obtained on new centers (prospective external validation), and the consistent results for the small and difficult group of borderline tumors. Overall, the model LR1 consistently performed well for binary classification (even though other linear models such as BLSSVM-lin, ObjLR, and BPER11 had nearly similar performance), while LR-PC2 performed excellent for multi-class classification. Actually, the performance of LR-PC2 for binary classification was similar to that of LR1, while LR-PC2 was also able to disentangle the three types of malignant tumors. This is a reason to use LR-PC2 rather than LR1. The binary models performed good to excellent on all centers of IOTA phase II, except maybe for KUK.

Chapter 7

The role of CA-125 for ovarian tumor diagnosis

Abstract – In clinical practice, the CA-125 tumor marker is widely used for ovarian tumor diagnosis, even though a few skeptical voices emerged in the scientific literature. Using the IOTA phase I data, we investigated the value of CA-125 in various ways. First, we investigated whether CA-125 is indispensable as an input in multivariate mathematical models. For this analysis, models based on least squares support vector machines and logistic regression were implemented. We also applied imputation techniques to fill up missing CA-125 levels to avoid possibly biased conclusions. Second, we compared the performance of CA-125 to the subjective opinion of experts operators. Finally, the merits of a diagnostic strategy involving both subjective opinion and CA-125 levels are checked.

7.1 Introduction

The CA-125 antigen is a glycoprotein that is expressed by the monoclonal antibody OC 125. It was suggested as a tumor marker for ovarian cancer in 1983 by Bast and colleagues [42]. They reported that only 6% of patients with nonmalignant ovarian disease had raised CA-125 levels, i.e. levels above 35 U/ml. Patients with ovarian cancer, however, had raised levels in 82% of the cases. These findings resulted in the routine use of CA-125 in the preoperative diagnosis of ovarian masses. The importance of this tumor marker was stressed even more by its inclusion in the popular risk of malignancy index [203], a test that is used to select patients for referral to tertiary oncology centers. However, skepticism regarding the usefulness of CA-125 has emerged, based on findings that CA-125 can be misleading in women suffering from conditions that affect the peritoneal surface, such as endometriosis [404], and in women with early stage ovarian cancer or borderline malignant tumors [404, 144, 145]. Thus, CA-125 appears important yet not powerful enough on itself. In Table 3.2, the AUC of CA-125 on the IOTA I training set is shown to be 0.831. This is very good, yet not good enough to use CA-125 by itself. Using the IOTA phase I data set, we aimed to scrutinize the value of the serum CA-125 tumor

marker in several ways. Because this analysis of CA-125 is based on a large multicenter data set, the conclusions drawn from the results have more generality than those drawn from previous investigations. If we can conclude that CA-125 is by no means necessary to measure, this would be a clinically valuable finding given the controversy that surrounds this marker.

First, we wanted to check whether other information, based on gray-scale and Doppler ultrasound examination, could cover for the information that is available in CA-125 levels. Therefore, we developed and compared mathematical models with and without CA-125 as an input [387, 132, 131]. This was done for all patients of the IOTA phase I data set, and for pre- and postmenopausal patients separately. The models were developed on the training part, and tested on the test part. An initial analysis was performed using logistic regression analysis on the patients who had CA-125 levels measured. In a subsequent analysis, missing CA-125 levels were imputed to avoid the criticism that the complete cases with respect to CA-125 represent a biased subgroup of the population. In this analysis, Bayesian LS-SVMs and Bayesian perceptrons were the algorithms used to develop models. Second, we compared the performance of CA-125 and subjective assessment [424, 410]. This is particularly interesting since, hitherto, there were no reliable reports on CA-125 levels for different histologies. Third, in relation to the previous point, we wanted to know whether some combination of subjective opinion and CA-125 levels yielded better results than any of the strategies alone.

7.2 The importance of CA-125 in mathematical models

7.2.1 Round one: Logistic regression models based on complete cases

Experimental setup

The data were extensively inspected to detect multicollinearity and outliers, as described in section 3.2.1. The patients in the training set for which CA-125 levels were available were used to create new logistic regression models containing CA-125. Notice that the original logistic regression model LR1 did not retain CA-125 as an input. New models containing CA-125 were also developed for premenopausal and postmenopausal patients separately. The adopted procedure for input selection was identical to that used for developing multi-class classifiers in section 5.3.1. Automated and manual forward/backward selection methods were used to detect interesting sets of inputs. A final selection of input variables for a logistic regression model was made using the AIC and BIC model fitting indices, repeated 5CV results on the AUC, and the likelihood ratio p-value. Attention was given to the linearity in the logit assumption and to interactions (using a stepwise selection procedure). Interactions were only included if they clearly improved the cross-validation results.

The performance of the models developed in this section, as well as that of LR1, was then assessed on the respective test set patients with CA-125 information. The performance was also assessed on the prospective IOTA Ib data set. The AUC was used as the main measure of model performance. The models were also evaluated and compared using pAUC and Sens75. Ninety-five percent CIs for performance measures as well as 95% CIs for the difference between two models were calculated using the bias-corrected bootstrap method based on 500 bootstrap samples drawn from the test set.

Concerning the subgroup analyses on the pre- and postmenopausal patients, we decided a priori that, if the AUC of the model including CA-125 would not be larger than that of M1, we would conclude that CA-125 did not improve the diagnostic performance in that subgroup. If, however, the model including CA-125 would perform better than LR1, a new model without CA-125 would be developed (to replace LR1) and the test set AUC of the new models would be compared. The rationale for this strategy is that, if the best model including CA-125 does not outperform LR1, it is not necessary to develop a new model without CA-125 for each group studied. LR1 may not be an optimal model for both subgroups, because it is a model created for both pre- *and* postmenopausal patients. The new models containing CA-125 as an input, as well as LR1, were then prospectively tested on the IOTA Ib data. Model performance was evaluated in the same way as for the IOTA I test set. Finally, using the IOTA I data, the algorithms were compared using 100 random train-test splits of the data with stratification for outcome. Each time, the model without CA-125 (i.e. LR1 or if needed a newly developed model) was trained on all training cases, while the three CA-125 inclusive models were trained on the appropriate group of training cases. Then, the models were tested on the appropriate group of test cases. The 100 resulting test set AUCs for each model were summarized by taking the median.

Results

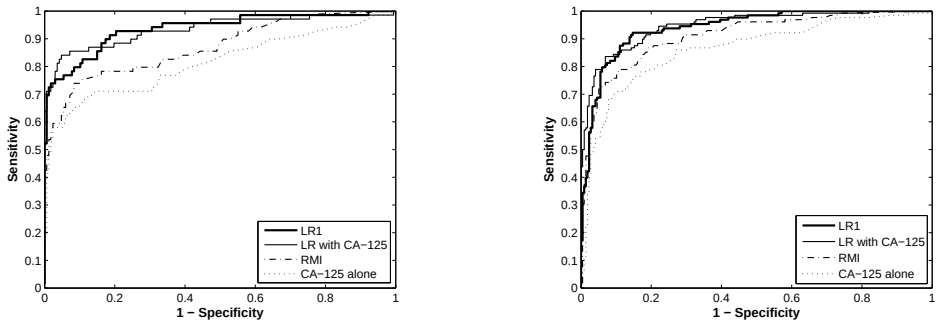
In the IOTA I data set, CA-125 levels were available for 809 patients (76%), 242 of which had a malignant tumor (30%). In the IOTA phase Ib data set, CA-125 results were available for 345 out of 507 patients (68%), of which 128 had a malignant tumor (37%).

All patients The model with CA-125 for all patients, LR_{all}, contained ten variables: log(ca125) (1), age (2), lesmax2 (3), solidmax2 (4), colscore (5), papflow (6), wallreg (7), pershistovca (8), ascites (9), hormtherapy (10). The probability of malignancy is computed as $\frac{1}{1+e^{-z}}$, where

$$\begin{aligned} z &= -9.1291 + 0.5471(1) + 0.0287(2) + 0.0109(3) + 0.0496(4) \\ &\quad +0.6490(5) + 1.323(6) + 0.9599(7) + 2.5447(8) \\ &\quad +0.9589(9) - 0.8239(10). \end{aligned} \tag{7.1}$$

The AUC performance of LR_{all}, LR1, RMI, and CA-125, and the comparisons between them, are presented in Table 7.1, both for the IOTA I test set (n=236) and the IOTA Ib data (n=345). Results for pAUC and Sens75 can be found in [387]. The ROC curves are shown in Figure 7.1. To summarize the results, LR_{all} was not superior to LR1, LR_{all} and LR1 were superior to RMI and CA-125, and RMI was superior to CA-125 alone. The analysis to compare the algorithms underlying classifiers LR1 and LR_{all} resulted in median AUCs of 0.946 and 0.949, respectively.

Thus, while CA-125 proved a helpful predictor in many previously developed multi-variable models, the IOTA analyses show that CA-125 is not necessary when (a combination of) other inputs is present. By looking at changes in the likelihood ratio χ^2 values of LR1's inputs when adding CA-125 to the model, variables that could cause CA-125 to become superfluous were identified. Further investigations showed that a model containing only ascites, lesmax2 and solidmax2 as inputs has a test set AUC of 0.887 while adding CA-125 to this model results in a test set AUC of 0.890. (For those interested, this difference in AUC was not significant using [120]'s method, p=0.7830.) When these three variables were omitted from LR1 the test set AUC became 0.893. Adding CA-125 to the nine remaining variables raised the test set AUC to 0.922 (for those interested, p=0.0313). Using canonical correlation analysis, the multiple correlation of log(CA-125) with ascites, solidmax2, and lesmax2 was 0.62. Associations between CA-125 and both ascites and tumor size have been reported on in the literature [404, 111]. Another input that may be partially responsible for making CA-125 superfluous is locul5.



(a) **(b)**

Figure 7.1: ROC curves of LR1, LR_{all}, RMI, and CA-125 when applied to patients from (a) the IOTA I test set and (b) the IOTA Ib data for which CA-125 levels were available.

Premenopausal patients The training set contained 315 premenopausal patients with CA-125 information; the test set contained 130 such patients. The model with CA-125 information developed for the group of premenopausal patients (LR_{pre})

Table 7.1: Performance of logistic regression models with and without CA-125, RMI, and CA-125 alone on the IOTA I test set and the IOTA Ib data.

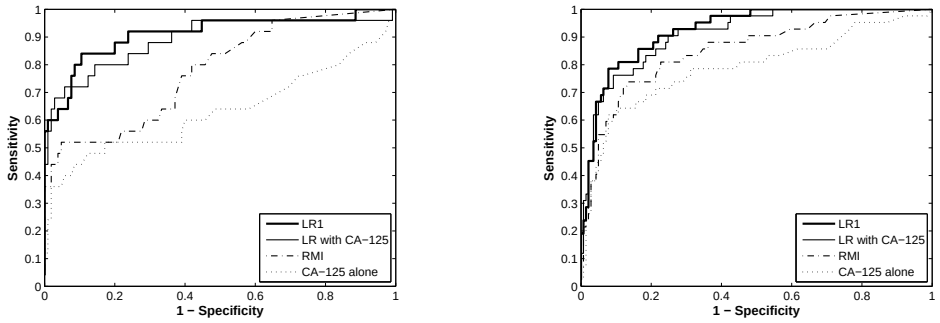
	IOTA I test set		IOTA Ib	
Comparison	AUC	95% CI for difference in AUC	AUC	95% CI for difference in AUC
All patients				
LR1 vs LR_{all}	.936 vs .934	−.013 to .018	.940 vs .947	−.021 to .008
LR1 vs RMI	.936 vs .870	.025 to .125	.940 vs .907	.000 to .069
LR1 vs CA-125	.936 vs .814	.059 to .200	.940 vs .862	.032 to .129
LR_{all} vs RMI	.934 vs .870	.026 to .117	.947 vs .907	.012 to .065
LR_{all} vs CA-125	.934 vs .814	.057 to .185	.947 vs .862	.049 to .120
RMI vs CA-125	.870 vs .814	.023 to .095	.907 vs .862	.022 to .071
Premenopausal				
LR1 vs LR_{pre}	.911 vs .891	−.001 to .051	.924 vs .912	−.017 to .046
LR1 vs RMI	.911 vs .778	.061 to .239	.924 vs .850	.014 to .138
LR1 vs CA-125	.911 vs .632	.155 to .444	.924 vs .790	.056 to .234
LR_{pre} vs RMI	.891 vs .778	.040 to .202	.912 vs .850	.005 to .111
LR_{pre} vs CA-125	.891 vs .632	.135 to .412	.912 vs .790	.054 to .198
RMI vs CA-125	.778 vs .632	.075 to .238	.850 vs .790	.023 to .111
Postmenopausal				
LR1 vs LR_{post}	.949 vs .975	−.057 to .003	.945 vs .959	−.046 to .018
LR1 vs RMI	.949 vs .924	−.032 to .098	.945 vs .928	−.028 to .074
LR1 vs CA-125	.949 vs .920	−.030 to .094	.945 vs .912	−.015 to .098
LR_{post} vs RMI	.975 vs .924	.012 to .115	.959 vs .928	.005 to .066
LR_{post} vs CA-125	.975 vs .920	.013 to .111	.959 vs .912	.018 to .089
RMI vs CA-125	.924 vs .920	−.028 to .033	.928 vs .912	−.010 to .044

contained nine variables: log(ca125) (1), lesmax2 (2), solidmax2 (3), papnr (4), colscore (5), pershistovca (6), ascites (7), hormtherapy (8), and pain (9). The probability of malignancy is computed as $\frac{1}{1+e^{-z}}$, where

$$\begin{aligned} z &= -10.7906 + 0.6284(1) + 0.0254(2) + 0.0767(3) + 0.6979(4) \\ &\quad +1.0104(5) + 2.5655(6) + 2.3792(7) - 1.4966(8) - 1.8999(9). \end{aligned} \tag{7.2}$$

The AUC performance of LR_{pre}, LR1, RMI, and CA-125, and the comparisons between them, are presented in Table 7.1, both for the IOTA I test set (n=130) and the IOTA Ib data (n=183). The ROC curves are shown in Figure 7.2. To summarize the results, LR_{pre} was not superior to LR1, both LR_{pre} and LR1 were superior to RMI and CA-125, and RMI was superior to CA-125 alone. The analysis to compare the algorithms underlying classifiers LR1 and LR_{pre} resulted in median AUCs of 0.944 and 0.949, respectively.

Postmenopausal patients The training set contained 258 postmenopausal patients with CA-125 information; the test set contained 106 such patients. The model with CA-125 information (LR_{post}) developed for the postmenopausal patients



(a) **(b)**

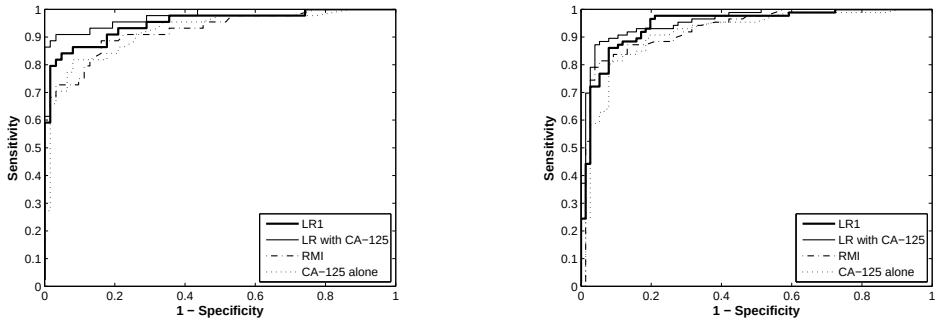
Figure 7.2: ROC curves of LR1, LR_{pre}, RMI, and CA-125 when applied to premenopausal patients from (a) the IOTA I test set and (b) the IOTA Ib data for which CA-125 levels were available.

used seven parameters: log(ca125) (1), solidmax2 (2), colscore (3), papflow (4), wallreg (5), hormtherapy (6), and presence of a multilocular tumor (locul3) (7). The probability of malignancy is computed as $\frac{1}{1+e^{-z}}$, where

$$\begin{aligned} z &= -7.2217 + 0.6887(1) + 0.0595(2) + 0.5588(3) + 1.2271(4) \\ &\quad +1.4862(5) - 0.9252(6) + 2.2282(7). \end{aligned} \tag{7.3}$$

The AUC performance of LR_{post}, LR1, RMI, and CA-125, and the comparisons between them, are presented in Table 7.1, both for the IOTA I test set (n=106) and the IOTA Ib data (n=162). The ROC curves are shown in Figure 7.3. To summarize the results, the performance of LR_{post} was similar to that of LR1. Even though all models performed very well, LR_{post} was superior to RMI and CA-125 in both data sets. The same was true for LR1 but differences in performance were smaller, especially the difference between LR1 and RMI. RMI and CA-125 had very similar performance. The analysis to compare the algorithms underlying classifiers LR1 and LR_{post} resulted in median AUCs of 0.928 and 0.945, respectively.

Because the AUC for LR_{post} was 0.026 higher than the AUC for LR1 (0.975 versus 0.949), we did develop a new model without CA-125 and compared its performance to that of LR_{post}. The difference in AUC on the IOTA I test set was 0.024 (0.951 v 0.975; 95% CI for the difference in AUC: -0.054 to 0.009; p=0.13) whereas the difference in AUC on the IOTA Ib data was 0.023 (0.936 v 0.959; 95% CI for the difference: -0.050 to 0.007; p=0.12). Evaluating the underlying algorithm resulted in a median AUC of 0.931.



(a) **(b)**

Figure 7.3: ROC curves of LR1, LR_{post}, RMI, and CA-125 when applied to postmenopausal patients from (a) the IOTA I test set and (b) the IOTA Ib data for which CA-125 levels were available.

Discussion and conclusions

A wealth of publications have concluded that CA-125 is a necessary tool for assessing ovarian tumors, see [42, 203] and the references in [387]. Therefore, our findings are remarkable. According to our data, CA-125 does not enhance our ability to characterize an ovarian mass using logistic regression models either in pre- or postmenopausal patients. In premenopausal patients, a logistic regression model that does not use CA-125 levels (LR1) performed best, yet differences with a logistic regression model with CA-125 were not large. The RMI and CA-125 alone clearly performed poorer. LR1 was developed on all training set patients, such that it may not even be optimal for premenopausal patients only. The situation is different in postmenopausal patients, where the logistic regression model including CA-125 performed best but differences with LR1 were again small. This is demonstrated by the higher Sens75 for LR1 in the IOTA Ib data set (97.7% for LR1 versus 93.0% for LR_{post}). The RMI and CA-125 alone performed very well too. When comparing the logistic regression algorithms in each of the three situations (i.e. all patients, premenopausal patients, and postmenopausal patients), the median AUCs that were obtained for LR1 and the model with CA-125 were very similar. The differences were not clinically relevant. Hence, it was found that CA-125 can be replaced by other measurements when diagnosing ovarian tumors.

There are several reasons that may explain why we have found CA-125 to be of little value, while other researchers did not come to this conclusion. Previous models were always based on data involving a relatively small number of measurements and patients. In the IOTA study, over 50 measurements were recorded for more than 1000 patients. It seems that, in LR1, the information on ascites, lesmax2, and solidmax2 made CA-125 levels redundant. If these three variables were removed from the

model, CA-125 did have added value. Hence, ultrasound information appears more informative than CA-125 when evaluating the risk of an adnexal mass being benign or malignant.

It should be acknowledged that CA-125 was not measured in all 1066 patients of the IOTA I data set, and that therefore the study population of 809 patients may be a nonrandom subsample from the data. However, an analysis of the full IOTA I data set showed that CA-125 was often not measured when the ultrasound examiner was certain about whether the mass was benign or malignant. When CA-125 was not measured, subjective assessment with regard to benignity versus malignancy performed very well. It can be argued that the exclusion of cases in whom the diagnosis was very clear does not invalidate the conclusion of this study. The study population is likely to be representative of patients in whom clinicians consider CA-125 to be helpful. It is good practice, however, to consider imputation techniques to fill in empty CA-125 levels. This allows us to use all patients for developing diagnostic models. This is the core rationale for the analysis described in the next section.

7.2.2 Round two: LS-SVMs and perceptron models based on imputed data sets

Introduction

This analysis differs from the previous one for two reasons. First, to avoid possibly biased results related to the missing values for CA-125, it was investigated whether CA-125 is redundant in mathematical models by imputing the missing values. It is known that caution is necessary when confronted with missing values, because the complete cases may not represent a random subsample of the whole data set. This may result in biased conclusions. However, in medical research missing values are often not dealt with appropriately [68]. For example, measurements and or cases with missing values are discarded, or missing values are simply imputed with the mean of the available values. Many imputation algorithms require the missing values to be missing at random (MAR). This term may wrongly suggest that it deals with missingness that has nothing to do with the other data or the study in general. In such a situation, missing values are said to be missing completely at random (MCAR). MAR means that the missing values depend on already observed data values, but not on yet to perform measurements. In case of the IOTA data, it means that missing CA-125 levels are not caused by the levels that would have been observed, but rather on other measurements that were already performed. Of course, MAR does not imply that CA-125 missingness is unrelated to the CA-125 levels itself, yet such a relationship should only be indirect through associations with other measurements [352]. The missingness patterns for CA-125 are likely to follow the MAR assumption. CA-125 is more often missing when the tumor looks clearly benign to the expert operator. Hence, whether CA-125 was measured or not depended on other observed data values. Apart from that, missingness patterns

also follow center-specific procedures. Some centers routinely measure CA-125 levels, while others do not.

Second, LS-SVMs with RBF kernel were used as the core algorithm in this analysis to investigate whether the use of a flexible algorithm would bring about different conclusions. Next to the LS-SVMs, Bayesian perceptrons were used as a linear alternative.

Imputation methods

Four imputation methods were applied: regression imputation [352], expectation-maximization (EM) imputation [121, 354], data augmentation (DA) [351], and hot-deck imputation [156].

Regression imputation Regression imputation aims at imputing the missing values using a regression model based on the cases with CA-125 information. CA-125 is predicted using the other variables (except the tumor outcome since that is the outcome of the diagnostic models developed later). A regression model was built using stepwise variable selection. This procedure tends to select many inputs, yet a rich imputation model is desired to cover up for the post-imputation analyses [352]: The regression imputation of CA-125 ignores the possible relationship between CA-125 and variables that are not included in the imputation model. CA-125 was bilog-transformed to approach a normal distribution for the regression analysis. This approach depends both on the choice of good regressors, and on the linearity of the dependence of the outcome (i.e. CA-125) on the regressors. Intuitively, one would exclude a significant missingness predictor from a regression imputation model, because the regression parameter for the available cases could be different from the one that could be expected for the imputed cases. However, it has been argued that the use of missingness predictors renders the MAR assumption more plausible such that this practice should be encouraged [352].

EM imputation The input vector of patient n, $\mathbf{x}_n$, can be split up in the observed values $\mathbf{a}_n$ and the missing values $\mathbf{m}_n$. Then, each $\mathbf{x}_n$ is generated by the underlying distribution $p(\mathbf{x}|\theta) = p(\mathbf{a}, \mathbf{m}|\theta)$ where θ are the distribution's parameters. If there are no missing values, θ can be easily estimated. When missing values are observed, the EM algorithm is a deterministic method to circumvent the problem of estimating θ without the missing values $\mathbf{m}$ by integrating $\log[p(\mathbf{a}, \mathbf{m}|\theta)]$ over $\mathbf{m}$ yielding the expected log-likelihood $E(\ell)$ [121, 354]. In practice, one iteratively alternates between an E-step in which the missing values $\mathbf{m}$ are estimated using initial parameter values θ_0 (first iteration) or estimated parameter values θ_{t-1} (iteration t), and an M-step in which $E(\ell)$ is maximized with respect to θ. The M-step yields θ_t, the updated estimate of the model parameters to be used in the next E-step. A variation of the EM algorithm, called regularized EM, was used in this study [354]. For EM imputation, the multivariate normality assumption

is required. A bilog-transformation was introduced for the highly skewed CA-125 levels in order to approximate a normal distribution.

Data augmentation Data augmentation (DA) [351] is a stochastic rather than deterministic method to circumvent the problem of estimating θ without the missing values $\mathbf{m}$. In DA, missing values are randomly imputed given a missing data distribution based on assumed values for θ. Given $\mathbf{a}$ and the imputed values for $\mathbf{m}$, a posterior distribution for θ is constructed from which updated estimates for θ are drawn. This procedure was implemented using the Gibbs sampling framework. In this two-step iterative process, the distributions for the missing values and the model parameters stabilize, leading to convergence in distribution whereas EM converges in its estimates. The difference with EM imputation is that, in DA, the values still change after convergence but their distribution remains fixed.

Both the EM and DA algorithms are based on a Bayesian estimation framework. As a consequence, they rely on the choice of a prior distribution for the model parameters that reflects our prior knowledge. In the existing literature, often an uninformative prior is selected that reflects our lack of a priori knowledge. Little variation was found in our study by varying the prior.

Hot-deck imputation Using a similarity metric, hot-deck imputation consists of selecting the closest complete sample to a sample with missing CA-125 and using it to impute CA-125. The most common similarity metric used is the sum of absolute differences for all variables (i.e. excluding CA-125 and the outcome). This method depends on the existence of similar cases in the data set. In a large data set it is more likely to obtain sufficient similarity to arrive at good imputations. In order to obtain a balanced similarity metric over all types of variables, the numerical variables were re-scaled to the $[0, 1]$ interval. Define C as the number of samples with CA-125 information in the IOTA I data set. Such samples are denoted as $(\mathbf{x}_c, y_c)$, $c = 1, ..., C$. The missing CA-125 level of patient m ($m = 1, ..., N - C$) is imputed using a sample $\mathbf{x}_c$, which is chosen as

$$\mathbf{x}_c = \arg\min_C \left[d(\mathbf{a}_c, \mathbf{a}_m) \right], \tag{7.4}$$

where the distance, as explained above, is

$$d(\mathbf{a}_c, \mathbf{a}_m) = |\mathbf{a}_c - \mathbf{a}_m|. \tag{7.5}$$

Of course, other distance metrics can be used for hot-deck imputation.

Experimental setup

Five types of diagnostic models were developed: four including one of the four imputed versions of the CA-125 variable, and one excluding CA-125 altogether.

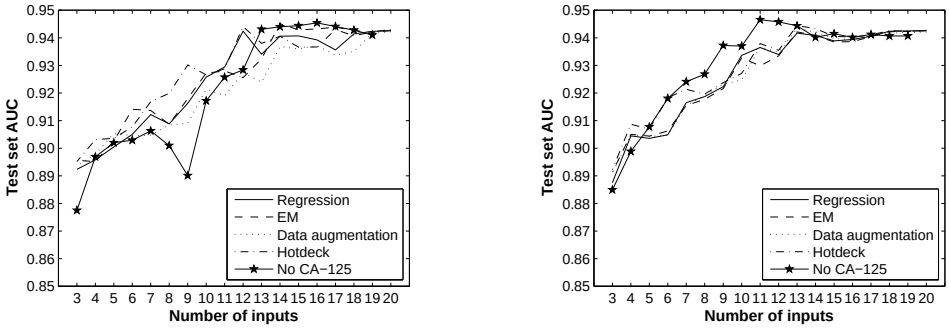
Input selection proceeded in two steps. First, 20 variables were selected based on previous input selection processes done on the training set using logistic regression and LS-SVMs, cf section 5.2. These 20 variables (including CA-125) can be seen as the most important ones for predicting tumor malignancy. The second step of the input selection process was done separately for each of the five model types: four times using an imputed version of CA-125, and once without CA-125. The 20 variables (19 when CA-125 is omitted, cf. the fifth type) were ranked according to their importance for predicting malignancy. This was done on the training set using ARD [270], an input selection method that is a natural part of the Bayesian framework. For the first four model types, CA-125 was by definition ranked on top. After the input selection process, and again separately for all five model types, we developed Bayesian LS-SVM and Bayesian perceptron models using the 3, 4, ..., or 20 (19 for the fifth model type) most important inputs, resulting in 18 (or 17) models. These models were then applied to the test set. Test set performance was evaluated using the AUC. In a second step, we constructed 100 new random train-test set splits of the IOTA I data set (cf. several previous chapters). Using the input ranking derived from ARD, the 18 (or 17) models were trained on the training set and tested on the test set. The 100 test set AUCs were summarized by their median.

Results and discussion

CA-125 was unavailable in 233 patients with a benign tumor (29%) and in 24 of the patients with malignant tumor (9%). This reflects the fact that some centers did not measure CA-125 when the tumor looked clearly benign. Patients with a benign tumor had a median CA-125 of 17 U/ml (IQR = 20.5) while the median CA-125 level was 167 U/ml (IQR = 570) when the tumor was malignant. For patients with a benign tumor, the imputed CA-125 levels had a median of 27 U/ml (IQR = 19.3) for regression imputation, 20 U/ml (IQR = 11.0) for EM imputation, 21 U/ml (IQR = 43.5) for DA imputation and 19 (IQR = 34.0) for hot-deck imputation. Median CA-125 levels for patients with a malignant tumour were, respectively, 42 U/ml (IQR = 67.6), 35 U/ml (IQR = 49.7), 39 U/ml (IQR = 40.6) and 59 U/ml (IQR = 315.8). The test set results are given in Figure 7.4. The curves are rather noisy, but the main observation was that ignoring CA-125 did not result in inferior performing classifiers compared to the use of an imputed version of CA-125.

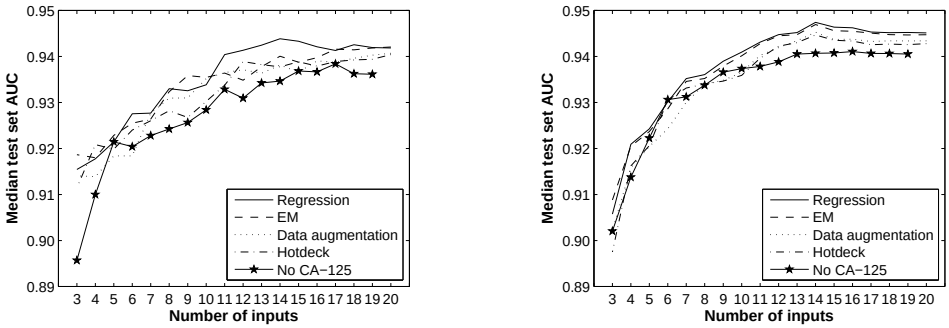
The results based on the 100 train-test splits yielded less noisy curves that could be interpreted more easily (Figure 7.5). There were only minor differences between the various imputation techniques, yet the regression imputation very often had the best median test set AUC. Both plots in Figure 7.5 show that the use of imputed versions of CA-125 resulted in slightly higher median AUCs than when CA-125 was omitted, yet differences were never larger than 0.01. An interesting observation in the AUC plot based on LS-SVMs, is that CA-125 mattered when only a few, albeit important, variables were included in the model. From a clinical point of view, this suggests that, by measuring CA-125, gynaecologists can filter patients with

increased accuracy only when very few measurements are available. Otherwise, CA-125 information becomes largely unnecessary.



(a) **(b)**

Figure 7.4: Test set AUC results based on (a) Bayesian LS-SVMs and (b) Bayesian perceptrons.



(a) **(b)**

Figure 7.5: Median test set AUCs based on 100 train-test splits for (a) Bayesian LS-SVMs and (b) Bayesian perceptrons.

This analysis confirmed and strengthened the findings presented in the previous section, based on logistic regression models for complete cases. CA-125 is not necessary to obtain an accurate prediction. One remark is that, for the 100 train-test splits, we always used the same input rankings, i.e. those obtained by applying ARD to the original IOTA I training set. This may limit the conclusions. Therefore, in an additional analysis ARD was redone for 20 of the 100 train-test splits within the Bayesian perceptron framework. The median test set AUCs resulted in the

same conclusion: CA-125 is not necessary, apart from a minor advantage when only a few other inputs were in the model. Finally, we did not apply a multiple imputation procedure, an approach often used imputation analyses, because it was not considered necessary for our primary focus of predicting malignancy.

7.3 CA-125 versus subjective ultrasound evaluation

7.3.1 Introduction and experimental setup

Using the complete IOTA I data set, we compared the performance of subjective assessment with that of CA-125 for predicting whether a tumor is benign or malignant. CA-125 is used to classify tumors by using a threshold. Often used thresholds are 30 U/ml or 35 U/ml, or 35 U/ml for postmenopausal patients and 65 for premenopausal petients [42, 204, 53, 404]. The comparison of subjective assessment and CA-125 entails the computation and comparison of proportions including 95% confidence intervals (CIs). CIs for single proportions were computed using Wilson's score interval method with continuity correction [302]. CIs for the difference between paired proportions were computed using a method that is based on Wilson's score interval (method 10 from [301]).

7.3.2 Results and discussion

Figure 7.6 shows box plots of CA-125 levels for different histologic diagnoses. For this figure, the IOTA I and IOTA Ib data sets were combined to get more accurate results. The figure based on IOTA I can be found in [424]. Overall, the results are similar. The CA-125 serum levels in women with endometriomas or abscesses were higher than those in women with other benign tumors or with borderline tumors of stage I, and were even not much lower than in women with rare primary invasive tumors. CA-125 levels were highest in women with stage IIIV primary invasive or borderline malignancies. CA-125 serum levels were 30 U/mL or higher in 664 of the 838 women with malignancies (79%, 95% CI = 76.4% to 81.9%), 307 of the 456 with endometriomas (67%, 95% CI = 62.9% to 71.5%), and in 25 of the 44 with abscesses (57%, 95% CI = 42.2% to 70.3%). In contrast, CA-125 serum levels were 30 U/mL or higher in only 40 of the 291 women with dermoid cysts (14%, 95% CI = 10.3% to 18.2%), 27 of the 200 with simple cysts (13%, 95% CI = 9.5% to 18.9%), and 84 of the 513 with (mucinous) cystadenomas (16%, 95% CI = 13.4% to 19.8%). See Table 7.2 for these results.

Subjective assessment correctly classified 92% (95% CI = 91.0% to 93.1%) of the tumors as benign or malignant, whereas a CA-125 threshold of 30 U/mL classified 72% (95% CI = 69.9% to 73.4%) of the tumors correctly (Table 7.2). By the use of CA-125 serum levels, at best 82% (95% CI = 80.2% to 83.1%) of the tumors were correctly classified with a threshold value of 100 U/mL. However, when a threshold

of 100 U/mL was used instead of 30 U/mL, CA-125 correctly classified 91% (95% CI = 89.9% to 92.5%) of the benign tumors as benign but only 61% (95% CI = 58.1% to 64.7%) of the malignant tumors as malignant.

Table 7.2: Median CA-125 levels and rate of correct classifications for subjective assessment and CA-125 using the data from IOTA I, IOTA Ib, and IOTA II.

		Median	Correctly classified as ben. or mal., N (%)			
		CA-125	Subjective	using CA-125 threshold (U/ml)		
Histology	*N*	(U/ml)	assessment	30	65	100
Endometrioma	456	44.5	440 (96)	149 (33)	307 (67)	366 (80)
Dermoid/teratoma	291	13	287 (99)	251 (86)	286 (98)	287 (99)
Simple cyst	200	14	198 (99)	173 (87)	192 (96)	197 (99)
Functional cyst	67	17	65 (97)	51 (76)	60 (90)	61 (91)
Hydrosalpinx	50	15.5	48 (96)	36 (72)	40 (80)	42 (84)
Peritoneal pseudocyst	12	17.5	12 (100)	8 (67)	10 (83)	11 (92)
Abscess	44	40	32 (73)	19 (43)	29 (66)	32 (73)
Fibroma	110	18	85 (77)	73 (66)	91 (83)	98 (89)
Cystadenoma	307	15	274 (89)	257 (84)	291 (95)	298 (97)
Mucinous cystadenoma	206	15	187 (91)	172 (84)	196 (95)	199 (97)
Rare benign tumor	30	18	20 (67)	18 (60)	26 (87)	27 (90)
All benign tumors	1773	18	1648 (93)	1207 (68)	1528 (86)	1618 (91)
Primary invasive	512	347	488 (95)	457 (89)	416 (81)	385 (75)
Stage I	118	86.5	110 (93)	82 (69)	65 (55)	55 (47)
Stage II	37	204	35 (95)	32 (86)	26 (70)	25 (68)
Stage III	306	402	292 (95)	297 (97)	279 (91)	259 (85)
Stage IV	51	730	51 (100)	46 (90)	46 (90)	46 (90)
Borderline	160	35	120 (75)	87 (54)	61 (38)	46 (29)
Stage I	142	31.5	103 (73)	73 (51)	48 (34)	34 (24)
Stage II	6	189.5	5 (83)	5 (83)	5 (83)	5 (83)
Stage III-IV	12	496	12 (100)	9 (75)	8 (67)	7 (58)
Metastatic	103	120	97 (94)	78 (76)	65 (63)	56 (54)
Rare primary invasive	63	63	52 (83)	42 (67)	31 (49)	28 (44)
All malignant tumors	838	180	757 (90)	664 (79)	573 (68)	515 (61)
All tumors	2611	26	2405 (92)	1871 (72)	2101 (80)	2133 (82)

Subjective assessment was superior to CA-125 (with a threshold of 30 U/mL) for correctly classifying benign (93% of benign tumors correctly classified versus 68%: difference = 25%, 95% CI = 22.4% to 27.3%) and malignant tumors (90% of malignant tumors correctly classified versus 79%: difference = 11%, 95% CI = 8.2% to 14.1%). Subjective assessment remained superior after stratifying for menopausal status, with 96% of benign tumors in premenopausal women being correctly classified by subjective assessment versus 62% by CA-125 (difference = 33%, 95% CI = 30.2% to 36.1%), with 87% of malignant tumors in premenopausal women being correctly classified by subjective assessment versus 73% by CA-125 (difference = 14%, 95%

CI = 8.5% to 20.0%), with 88% of benign tumors in postmenopausal women being correctly classified by subjective assessment versus 80% by CA-125 (difference = 8%, 95% CI = 3.4% to 11.6%), and 92% of malignant tumors in postmenopausal women being correctly classified by subjective assessment versus 83% by CA-125 (difference = 9%, 95% CI = 6.1% to 12.8%).

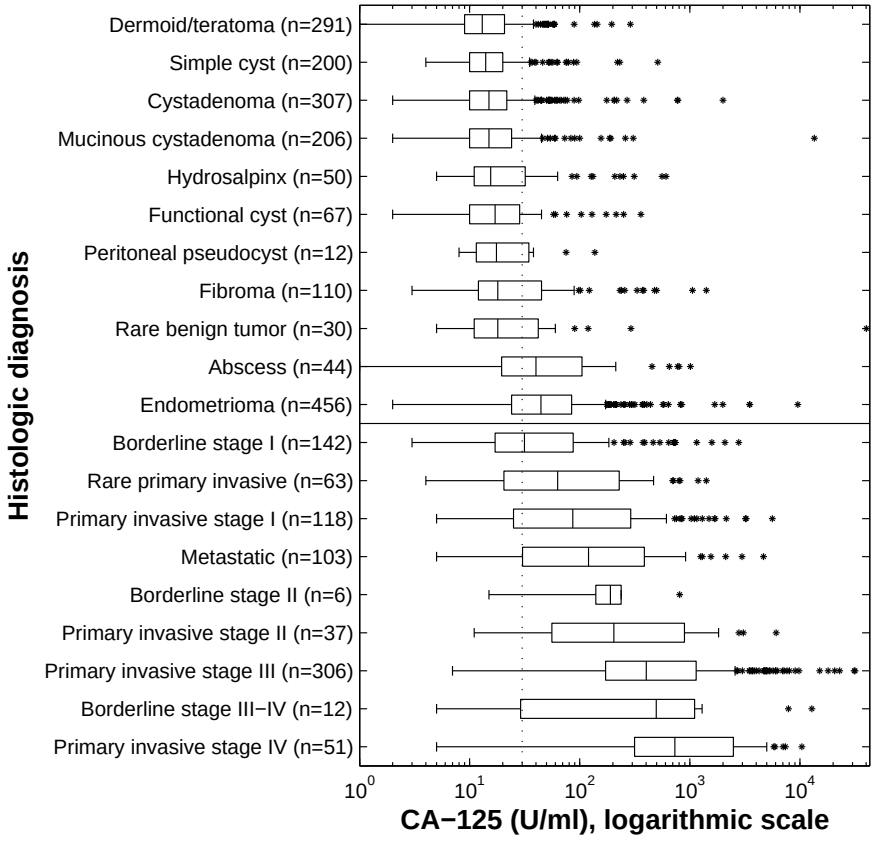


Figure 7.6: Boxplots showing CA-125 levels by histologic diagnosis using the data from IOTA I, IOTA Ib, and IOTA II (n=2611).

For nearly all specific histologic diagnoses, subjective assessment was superior to CA-125. Subjective assessment correctly classified most benign tumors associated with high levels of CA-125 as benign (i.e. endometriomas and abscesses), and it correctly classified more borderline stage I tumors as malignant than did CA-125 at any cutoff value. For example, subjective assessment correctly classified 90% (95% CI = 87.0% to 91.7%) of these three histologies, whereas a CA-125 threshold of 30 U/mL classified only 38% correctly (95% CI = 33.9% to 41.4%). Moreover, subjective assessment was superior to CA-125 for discriminating between benign tumors and borderline tumors and between benign tumors and stage I primary

invasive malignancies in both pre- and postmenopausal women, irrespective of the CA-125 threshold chosen to indicate malignancy (results can be found in [424]).

Subjective assessment was superior to CA-125 irrespective of the certainty of the expert operator regarding the nature of the mass. This was true both for pre- and postmenopausal patients. On top of that, we investigate the use of CA-125 as a second-stage test after subjective assessment. This was done in two scenarios: in the first, CA-125 could overrule a benign diagnosis made by the expert operator if the CA-125 level exceeded a threshold; in the second, CA-125 could overrule an operator's diagnosis of malignancy if CA-125 levels were below the threshold. In both scenarios, the use of CA-125 as a second stage test decreased diagnostic performance, and this was true both for pre- and postmenopausal patients and for different levels of certainty on behalf of the expert regarding the nature of the mass.

7.4 Conclusions

The results in this chapter clearly demonstrate that ovarian masses are more accurately diagnosed by using ultrasound examination than by evaluating CA-125 levels. First, mathematical models using ultrasound information and patient characteristics was not inferior to models that also used CA-125. This was investigated using logistic regression, Bayesian LS-SVMs, and Bayesian perceptrons, and using imputation techniques to fill in the missing CA-125 levels. Second, the subjective assessment of an expert operator was much better than CA-125 in diagnosing ovarian masses. This entails an important message for clinical practice. It is not needed to routinely measure CA-125, whereas it is very interesting to offer good ultrasound education programs for gynecologists.

Chapter 8

Diagnostic models for pregnancies of unknown location

Abstract – In this chapter, diagnostic models for pregnancies of unknown location (PUL) are developed. The aim is to accurately predict whether the PUL is a failing pregnancy, a normal intra-uterine pregnancy, or an ectopic pregnancy. The models are based on logistic regression, multi-layer perceptrons, least squares support vector machines, and kernel logistic regression. The models are developed on the training part of a PUL data set containing 1003 such pregnancies. Model performance and comparison was done on the test part of this data. At this very moment, the newly established International Pregnancy of Unknown Location Analysis (IPULA) group is collecting data from a large number of PULs at multiple centers around the world. The resulting data set will be used to prospectively test the models developed in this chapter, and to develop new models.

8.1 Introduction

As explained in Chapter 3, it is important to predict the nature of PULs as soon as possible. Some pregnancies need more attention than others. More specifically, ectopic pregnancies need careful monitoring as they can rupture and possibly jeopardize the life of the mother. Women with an intra-uterine pregnancy need standard pregnancy follow-up procedures, whereas women with failing pregnancies need no further follow-up and can be discharged. The early diagnosis of PULs using mathematical models is the topic of this chapter. Using rather small data sets, initial logistic regression models were built [104, 91, 90]. A first analysis was performed by Lu [104] using 388 women with a PUL recruited at the early pregnancy unit (EPU) of St Georges Hospital in London (UK) between June 2001 and December 2002. The resulting model, M1, contained only the hCG ratio. The obtained parameter estimates can be found in [104]. Using a set of 376 PULs collected at the same center between March 2002 and July 2003, we built two new logistic regression models [91, 90]. Finally, using a larger set of PULs, an extensive analysis was performed in which logistic regression, Bayesian MLPs, Bayesian LS-SVMs, and multi-class KLR

models were developed and compared [415]. The training set of the larger set of PULs comprised the smaller data sets used for the initial logistic regression models. The descriptive statistics for the measurements recorded for all patients in the larger data set are presented in Table 3.9.

8.2 Initial logistic regression models

The training set consisted of 201 PULs (53.5%) examined between March 2002 and November 2002, of which 102 failing PULs (54.2%), 76 IUPs (37.8%), 12 EPs (6.0%), and 4 persisting PULs (2.0%). The test set had 175 PULs examined between November 2002 and July 2003: 94 failing PULs (53.7%), 64 IUPs (36.6%), 15 EPs (8.6%), and 2 persisting PULs (1.1%). Persisting PULs were discarded from all model building analyses.

8.2.1 M4: the model using hCG information

The following variables were considered in the input selection analyses: hCG ratio, logarithm of the average hCG, progesterone ratio, endometrial thickness, and maternal age. The logarithm of the average progesterone was not considered because it correlated very strongly with the hCG ratio (Pearson $r = 0.83$), such that multicollinearity was present. Also, gestational age was not considered because it had some missing values, for example for 3 out of 12 ectopic pregnancies in the training set. Promising input sets were selected using stepwise selection and manual selection procedures. A final input set was chosen based on AIC and BIC, and to a lesser extent also based on the likelihood ratio p-value for the inputs. This was followed by checks for linearity in the logit and checks for useful interaction terms. This procedure is fairly similar to that which was followed for the logistic regression models to predict the nature of ovarian masses, cf. sections 5.3.1 and 7.2.1. The biggest difference is that input selection for the ovarian tumor models was partly based on cross-validation performance, whereas the input selection analysis described in this paragraph was not (and was performed some time before the input selection analyses for ovarian tumor models).

The final model, M4, contained the hCG ratio (a), the square of the hCG ratio (included as a result of the checks for linearity in the logit), and the log of the hCG average (b). The hCG ratio was centered to artificially suppress the correlation between hCG ratio and its square in order to avoid instability of the parameter estimation. Centering was done by subtracting the training set average of hCG ratio (1.17). The model is as follows:

$$\begin{cases} \log\left(\frac{\hat{\pi}_{failing}}{\hat{\pi}_{IUP}}\right) &= 5.88 - 5.56(a-1.17) + 2.05(a-1.17)(a-1.17) \\ & \quad -1.18(b) = z_1 \\ \log\left(\frac{\hat{\pi}_{ectopic}}{\hat{\pi}_{IUP}}\right) &= 0.39 - 0.26(a-1.17) - 3.93(a-1.17)(a-1.17) \\ & \quad -0.06(b) = z_2 \\ \log\left(\frac{\hat{\pi}_{ectopic}}{\hat{\pi}_{failing}}\right) &= z_2 - z_1 = -5.49 + 5.30(a-1.17) - 5.98(a-1.17)(a-1.17) \\ & \quad +1.12(b). \end{cases} \tag{8.1}$$

The average hCG is clearly lower for failing PULs than for both other classes. The effect of hCG ratio is rather complicated due to the quadratic effect. Low ratios yield high probabilities for failing PUL, high ratios for IUP, and average ratios (around 1-1.5) for EP. However, very high ratios (3.5 and above) again increase the likelihood of a failing PUL. The reason is that hCG ratio is highly skewed for failing PULs.

8.2.2 M5: does clinical information matter?

In another analysis using the same PULs, we investigated whether clinical information could improve the prediction. Clinical data considered were the level of vaginal bleeding (ordinary variable with three levels: none, without clots, with clots), presence and location of abdominal pain (categorical variable with four categories: no pain, lower abdomen, left iliac fossa, right iliac fossa), character of the midline echo (disrupted or intact), presence of free fluid in the Pouch of Douglas (yes or no), number of risk factors for ectopic pregnancy based on gynecologic history, presence of abdominal tenderness during examination (yes or no), four variables indicating site-specific tenderness on TVS over the uterus, right adnexa, left adnexa, and cul-de-sac (yes or no), and an overall pain score using a visual analogue scale from 0 (no pain during scan) to 100 (worst pain ever). Risk factors for ectopic pregnancy were pelvic inflammatory disease, chlamydia and/or gonorrhea, previous ectopic pregnancy, endometriosis, infertility, fertility treatment, tubal surgery, operative laparoscopy, evacuation of retained products of conception, suction termination of pregnancy, sterilization, and use of an intra-uterine conceptive device.

Again, gestational age was not considered in the input selection analysis. A multicollinearity analysis resulted in the following variables that were used in the input selection analysis: hCG ratio, logarithm of the average hCG, age, endometrial thickness, midline echo, free fluid, level of vaginal bleeding, number of risk factors for EP, tenderness over the uterus, and abdominal pain. Univariate analysis on the training set suggested that the level of vaginal bleeding may be helpful to predict ectopic pregnancy. The other clinical factors appeared to be of little, if any, importance. The final model included these inputs: hCG ratio (a), square of the hCG ratio, logarithm of the average hCG (b), and level of vaginal bleedings

(c). Again, hCG ratio was centered to suppress the correlation with its square. The model was as follows:

$$\begin{cases} \log\left(\frac{\hat{\pi}_{failing}}{\hat{\pi}_{IUP}}\right) &= 5.97 - 4.86(a-1.17) + 1.85(a-1.17)(a-1.17) \\ & \quad -1.47(b) + 1.75(c) = z_1 \\ \log\left(\frac{\hat{\pi}_{ectopic}}{\hat{\pi}_{IUP}}\right) &= -0.79 + 0.25(a-1.17) - 3.77(a-1.17)(a-1.17) \\ & \quad -0.05(b) + 1.33(c) = z_2 \\ \log\left(\frac{\hat{\pi}_{ectopic}}{\hat{\pi}_{failing}}\right) &= z_2 - z_1 = -6.76 + 5.11(a-1.17) - 5.63(a-1.17)(a-1.17) \\ & \quad +1.42(b) - 0.42(c). \end{cases} \tag{8.2}$$

8.2.3 Test set performance and conclusions

M4 had test set AUCs of 0.978, 0.974, and 0.900 for failing PULs, IUPs, and EPs, respectively. For M5, the AUCs were 0.979, 0.979, and 0.912. The corresponding results for M1 were 0.965, 0.969, and 0.842. Using the method described in [120], the difference in AUC for EP between M1 and M4 was statistically significant ($p = 0.03$), but the difference between M4 and M5 was not ($p = 0.18$). Using the training set results for M4, it was decided to make crisp predictions using weights 1, 1, and 4, for the predicted probabilities of failing PUL, IUP, and EP, respectively. The crisp prediction is the class with the highest weighted probability. On the test set, sensitivities for failing PULs, IUPs, and EPs were 87.2, 85.9, and 80.0%, respectively. For M5, the weights were set at 2, 1, and 7, achieving test set sensitivities of 90.4, 85.9, and 78.6%, respectively. For M1, weights 1, 1, and 4 were derived from its training set results. Resulting sensitivities on the present test set were 86.2, 85.9, and 73.3, respectively. The conclusions are that hCG information achieves good diagnostic results, and that clinical factors are most likely not necessary to predict the outcome of PULs. Moreover, the only clinical factor that could be useful, is the level of vaginal bleeding.

8.3 New diagnostic models using a large set of PULs

The aim of this study was to use various methods with probabilistic output based on logistic regression (LR), neural networks, and LS-SVMs to predict the outcome of PULs. Similar to the models developed for ovarian tumors, we were interested in the comparison of several algorithms for probabilistic multi-class classification when applied to a real life medical decision making problem.

The initial logistic regression models showed that the hCG ratio and the average hCG level may be crucial measurements to predict the PUL outcome. The amount

of vaginal bleeding also appeared interesting even though its inclusion did not seem to improve the discriminative performance of the model. In sum, these models suggested that biochemical markers related to hCG levels are the key indices to predict PUL outcome. Progesterone levels were also interesting biochemical markers, but they were not used in the logistic regression models since they are strongly correlated with hCG ratio. Another analysis suggested that creatine kinase and CA-125 were not at all useful for diagnosing PULs [100]. Undoubtedly, the single most important parameter is the hCG ratio. These models, however, were developed on small data sets using a basic statistical method. Even though simple methods already perform very well in many cases, it was necessary to apply more advanced techniques.

The data set is described in section 3.2.2. Twelve variables were considered: hCG ratio, logarithm of the average hCG, logarithm of the average progesterone, endometrial thickness, maternal age, gestational age, level of vaginal bleeding, pain at examination, previous EP, previous miscarriage, character of the midline echo, and presence of free fluid in the Pouch of Douglas.

8.3.1 Algorithms and input selection

We applied standard statistical multi-class logistic regression (MLR) with manual checks for transformations, interactions, and the linearity in the logit assumption, a combination of standard binary logistic regression models with similar checks (LR-PC), a Bayesian MLP (BMLP), a Bayesian perceptron (BPER), combinations of binary Bayesian LS-SVMs using either the linear or RBF kernel (BLSSVMlin, BLSSVMrbf), and multi-class kernel logistic regression using RBF kernel (MKLR). The binary logistic regression models for LR-PC were combined using pairwise coupling, a method that we already explained in previous chapters. Binary LS-SVM models were combined using either pairwise coupling or ensembles of nested dichotomies (END). END was applied as described in section 4.1.2 (method `endall`. In this method, the multi-class problem is presented as a tree of nested binary classifiers. First, all classes are divided into two sets for which a binary classifier is constructed. If a set consists of more than one class it is again split up into two sets. This is repeated until all classes form a set of their own. The multi-class probability for a class is obtained by multiplying all binary probabilities for sets involving that class. Several different trees can be constructed from multi-class problems, each of which is usually not more important than any other tree a priori. We averaged the multi-class probabilities of all possible trees to obtain the final estimated probabilities for the three outcomes. Both pairwise coupling and END performed well in the comparison of methods for generating multi-class probabilities that was presented in section 4.1. This comparison study involved the training set of the data used in this section. Thus, we implemented nine methods: MLR, LR-PC, BMLP, BPER, BLSSVMlin-PC, BLSSVMlin-END, BLSSVMrbf-PC, BLSSVMrbf-END, and MKLR.

Input selection was performed separately for the different models. Remember that MLR and LR-PC involved manual checks, possibly resulting in adaptations of the input set (e.g. transformations, interaction effects, or polynomial effects). As a consequence, input selection was by definition different for different models. For the MLR model and the three binary LR models for the LR-PC approach, several interesting input sets were selected from all available measurements based on three criteria: AIC, BIC, and the p-value for each input based on the likelihood ratio test given the other inputs in the model. Then, for each promising input set, the linearity in the logit assumption was checked for all continuous/ordinal inputs. Next, we checked for interesting interaction effects to add to the input set. Finally, 20 5CV runs (R5CV) were performed on the training set to select the best input set. Each 5CV run resulted in 5 validation set areas under the receiver operating characteristic curve (AUC) per class. The use of R5CV resulted in 100 validation set AUCs per class. These 100 AUCs per class were summarized by their mean. The selection of the final input set was based on the mean AUCs with extra importance for the mean AUC for the ectopic pregnancy class because that is the most difficult, yet most important, class.

For the BMLP, BPER, and the binary BLSSVMs, the ARD method was used within the context of each model (as explained in Chapter 2). For BMLP, the ARD analysis was repeated ten times with different initializations of the model parameters, and the median rank of each input over the ten ARD runs was computed. The input with worst median rank was dropped and ten new ARD runs were done. This procedure allowed us to make a final ranking of input importance by recording the order in which inputs were dropped. The ARD analysis was done using a hidden layer with 10 hidden neurons. Then, using R5CV, we simultaneously determined how many of the most important inputs should be used and what number of hidden neurons would be optimal for the final BMLP model (i.e. the architecture of the network). For the BPER model, a similar strategy was adopted (except that it was not necessary to use different initializations of the model parameters).

For binary BLSSVMs, 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. Again, R5CV was used to determine how many of the most important inputs were to be chosen. This was done for all three 1-versus-1 BLSSVM classifiers. The END approach for combining binary probabilities into multi-class probabilities also needs to train 1-versus-2 classifiers (i.e. 1-vs-All classifiers in case of a three-class problem). The inputs for these classifiers were those selected for both 1-versus-1 classifiers involving the class that is compared to both other classes. For the MKLR model, all inputs selected for any of the binary BLSSVM models were used.

8.3.2 Model evaluation

Input selection and model training was done using the training set. The resulting classifiers were then applied to the test set, which was used solely for model testing. The main performance evaluation measure was the AUC. This is not straightforward in multi-class problems. We chose to look first at each class' separate AUC, as we did for the multi-class models for ovarian tumors. In addition, we applied multi-class AUC metrics, as suggested in Chapter 4. pAUC and Sens75 were considered as well. As is the case for AUC, both pAUC and Sens75 can be computed for each class. Applying many 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. Confidence intervals for performance measures on the test set data were obtained using the bias-corrected bootstrap method based on 1000 bootstrap samples. Using calibration curves, the classifiers were compared with respect to the calibration of the estimated probabilities.

Next to classifier evaluation, we also evaluated and compared the algorithms. One hundred random splits of the data into training and test set were made with stratification for outcome, as was done many times before in this thesis. Each time, the model was trained on the training set and tested on the test set. The performance measures for the 100 test sets were summarized by their median. Note, however, that we did not perform separate input selection for all models on each of the 100 training sets but simply used the inputs as selected on the original training set to reduce computational load.

Comparing the classifiers/algorithms for a particular performance measure was done using the ranking technique used previously.

8.3.3 Results

Multicollinearity and input selection

First, the level of multicollinearity among the variables was assessed using the training data. This was not straightforward, because we do not know of a method to directly expand the method of weighted ordinary least squares regression to multi-class problems. Therefore, the multicollinearity analysis was performed for six binary problems: failing PUL vs rest, IUP vs rest, EP vs rest, failing PUL vs IUP, failing PUL vs EP, and IUP vs EP. For each variable, the maximum VIF over the six analyses was recorded (Table 8.1). The VIFs for hCGratio and log(progmean) exceed 2.5, all other are low. hCGratio and log(progmean) were strongly correlated (Spearman rho = 0.78). Given this correlation, the VIFs for these variables were expected to be higher. As was done for the initial logistic regression models, it was defendable to exclude log(progmean) from further analyses. Based on the very high AUCs for this variable in the training set (Table 3.9), the variable was kept to check

if it has additional predictive value over hCGratio.

Table 8.1: Maximum VIF for the PUL variables.

Variable	Max VIF	Variable	Max VIF
log(progmean)	3.58	freefluid	1.37
hCGratio	2.61	endthick	1.36
prevect	1.75	pain	1.25
bleeding	1.61	age	1.19
gestation	1.52	midline	1.17
log(hCGmean)	1.48	prevmisc	1.11

These are the selected inputs per model:

- MLR: log(hCGratio) (1), log(hCGmean) (2), log(progmean) (3), bleeding (4), age (5), age*age (6), log(hCGratio)*log(progmean) (8);
- LR-PC: log(hCGratio)^{1,2,3} (1), log(hCGmean)^{1,2} (2), log(progmean)^{1,2,3} (3), bleeding^{3} (4), age^{2} (5), age*age^{2} (6), prevect^{3} (7), log(hCGratio)*log(progmean)^{2} (8);
- BMLP: hCGratio, log(hCGmean), log(progmean), bleeding, prevect;
- BPER: hCGratio, log(hCGmean), log(progmean), bleeding;
- BLSSVM models: hCGratio^{1,2,3}, log(hCGmean)^{1,2}, log(progmean)^{1,2,3}, bleeding^{2,3}, age^{1,3}, prevect^{3};
- MKLR: hCGratio, log(hCGmean), log(progmean), bleeding, age, prevect.

Models consisting of a combination of binary models (i.e., LR-PC and all BLSSVM models) can use different input sets for the different binary models. For these models, it was indicated in the above list in which binary model each input was used: ^{1} means the input is used in the model contrasting failing PUL and IUP, ^{2} means the input is used in the model contrasting failing PUL and EP, and ^{3} means the input is used in the model contrasting IUP and EP. The input selection analyses tended to yield similar results: hCG ratio, hCG average, progesterone average, vaginal bleedings, age, and the presence of previous ectopic pregnancies were the important variables. The linearity in the logit analyses for MLR and LR-PC suggested the use of log(hCGratio) or of the square of hCGratio, and the use of age and its square. R5CV suggested the use of log(hCGratio) rather than the square of hCGratio. The interaction between log(hCGratio) and log(progmean) also appeared useful. These results for the logistic regression models suggested that the decision boundaries between the three classes are nonlinear. This is supported by the cross-validation results for the BMLP method, which suggested the use of three hidden neurons.

This is the obtained MLR classifier:

$$\begin{cases} \log\left(\frac{\hat{\pi}_{failing}}{\hat{\pi}_{ectopic}}\right) &= 8.18 - 9.44(1) - 0.86(2) - 1.05(3) - 0.86(4) \\ & \quad +0.086(5) + 0.024(6) + 2.34(8) = z_1 \\ \log\left(\frac{\hat{\pi}_{IUP}}{\hat{\pi}_{ectopic}}\right) &= -2.34 - 5.00(1) + 0.17(2) + 0.30(3) - 1.41(4) \\ & \quad -0.018(5) + 0.018(6) + 2.17(8) = z_2 \\ \log\left(\frac{\hat{\pi}_{IUP}}{\hat{\pi}_{failing}}\right) &= z_2 - z_1 = -10.51 + 4.44(1) + 1.03(2) + 1.35(3) - 0.55(4) \\ & \quad -0.10(5) - 0.0062(6) - 0.17(8). \end{cases} \tag{8.3}$$

and these are the binary models used for the LR-PC classifier:

$$\begin{aligned} \log\left(\frac{\hat{\pi}_{fp}}{\hat{\pi}_{ep}}\right) &= 7.67 - 9.78(1) - 0.97(2) - 0.92(3) + 0.11(5) + 0.031(6) + 2.36(8), \\ \log\left(\frac{\hat{\pi}_{iup}}{\hat{\pi}_{ep}}\right) &= -2.52 + 3.95(1) + 0.80(3) - 1.65(4) - 2.73(7), \\ \log\left(\frac{\hat{\pi}_{uip}}{\hat{\pi}_{fp}}\right) &= -9.49 + 3.40(1) + 0.79(2) + 1.35(3). \end{aligned}$$

For each case, results for all three models are transformed into multi-class probabilities using pairwise coupling.

Classifier performance on the test set

The test set results are shown in Table 8.2. The AUC, pAUC, and Sens75 values for failing PUL and IUP are not shown, but their results were extremely good with minor variations between classifiers. For failing PUL, these indices were between 0.985 and 0.990, 0.238 and 0.241, and 98.9 and 100%, respectively. For IUP, they were between 0.984 and 0.990, 0.234 and 0.240, and 99.3 and 100%, respectively. The ROC curves for these two classes are nearly identical for all models (Figure 8.1). Since ectopic pregnancy can be dangerous but is difficult to predict, we focused on this class. The AUC for ectopic pregnancy were good, ranging from 0.87 to 0.92. Sensitivity levels at 75% specificity were around 92% for several classifiers. The ROC curves for ectopic pregnancy are shown in Figure 8.1. The results confirmed that linear classifiers are suboptimal for the diagnosis of PULs. The three linear classifiers performed worst with respect to discriminatory ability. Note that we do not regard the logistic regression models as linear models based on their use of interaction and polynomial effects. The Bayesian LS-SVM classifiers showed once more to yield good probabilistic output, as they obtained high values for M_s and AOT. The triangle whose weighted area is the AOT (cf. Chapter 4) is shown in Figure 8.2 for a few classifiers.

Table 8.2: Test set results for the PUL classifiers.

	For EP only			Multi-class metrics			
Classifier	AUC (95% CI)	pAUC	Sens75	VUS_2	M (95% CI)	AOT	M_s
MLR	.920 (.874-.954)	.172	91.7	.835	.954 (.930-.972)	.315	.626
LR-PC	.924 (.872-.965)	.182	87.5	.838	.955 (.925-.978)	.347	.640
BMLP	.910 (.853-.954)	.174	91.7	.801	.949 (.920-.972)	.121	.401
BPER	.866 (.799-.923)	.141	75.0	.737	.930 (.895-.957)	.128	.489
BLSSVMlin-PC	.896 (.821-.954)	.168	83.3	.794	.940 (.902-.969)	.256	.599
BLSSVMlin-END	.875 (.795-.945)	.161	79.2	.760	.931 (.889-.964)	.190	.558
BLSSVMrbf-PC	.907 (.835-.958)	.177	91.7	.807	.946 (.906-.972)	.317	.630
BLSSVMrbf-END	.914 (.852-.959)	.178	91.7	.821	.951 (.918-.974)	.345	.648
MKLR	.917 (.864-.962)	.176	91.7	.816	.952 (.922-.973)	.273	.593

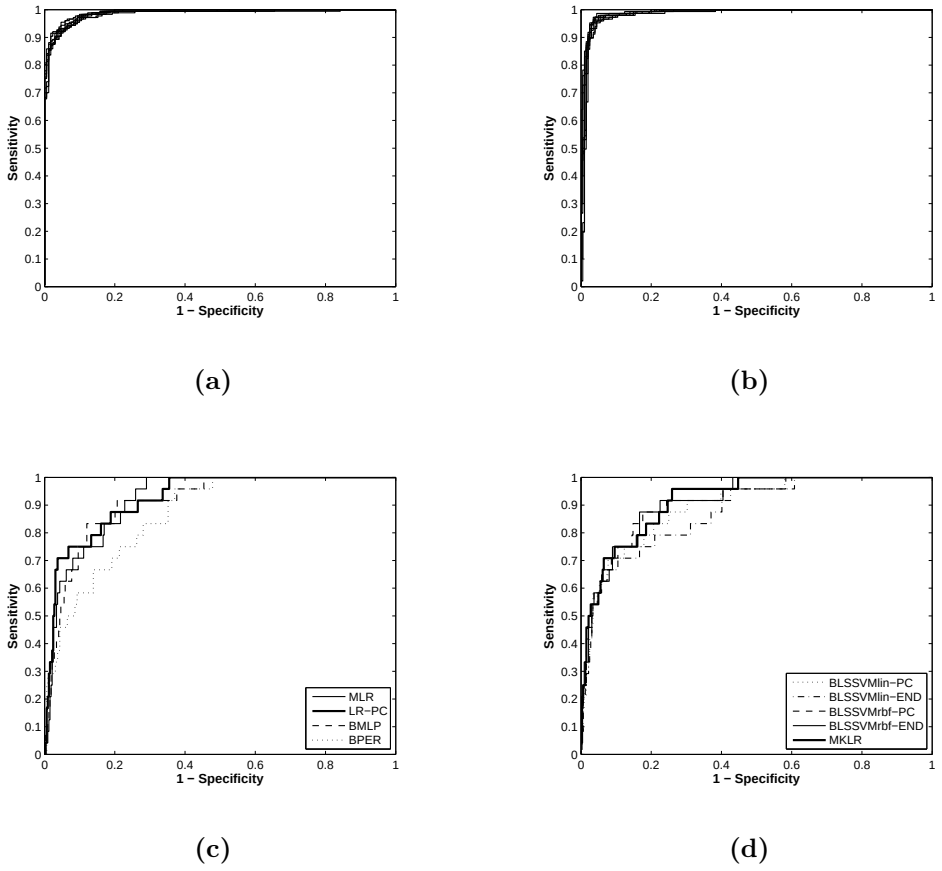


Figure 8.1: Test set ROC curves for (a) failing PULs, (b) IUPs, and (c-d) EPs .

The classifiers are ranked for the following performance measures: AUC for ectopic pregnancy (AUC_{EP}), M, M_s, and AOT (Table 8.3). Considering AUC_{EP}, LR-PC and MLR are ranked at the top, with good values for AR, Pm(1) and

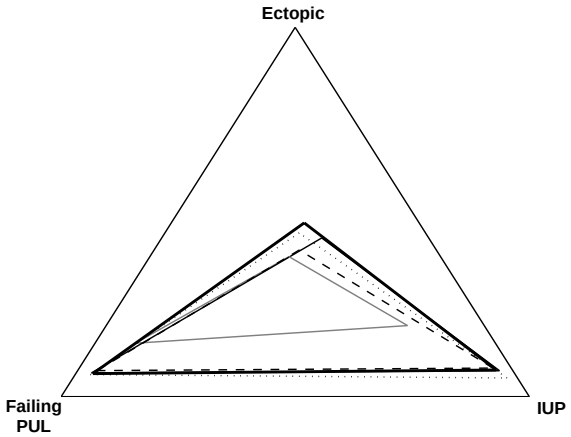


Figure 8.2: Triangle plots for MLR (normal line), LR-PC (bold line, largest triangle), BLSSVMrbf-END (dotted line), MKLR (dashed), and BMLP (gray line; smallest triangle).

Table 8.3: Ranking results for the PUL classifiers.

Classifier	AUC_{EP}	AR-Pm(1)-Pm(3)	Classifier	M	AR-Pm(1)-Pm(3)
LR-PC	.924	2.1-.38-.89	LR-PC	.955	2.2-.33-.86
MLR	.920	2.6-.42-.74	MLR	.954	2.3-.50-.79
MKLR	.917	3.5-.08-.56	MKLR	.952	3.8-.07-.48
BLSSVMrbf-END	.914	3.7-.04-.36	BLSSVMrbf-END	.951	3.6-.04-.44
BMLP	.910	4.5-.04-.23	BMLP	.949	4.4-.03-.27
BLSSVMrbf-PC	.907	4.9-.05-.20	BLSSVMrbf-PC	.946	5.3-.02-.15
BLSSVMlin-PC	.896	6.8-.00-.01	BLSSVMlin-PC	.940	6.8-.00-.02
BLSSVMlin-END	.875	8.3-.00-.00	BLSSVMlin-END	.931	8.5-.00-.00
BPER	.866	8.6-.00-.00	BPER	.930	8.1-.00-.00
Classifier	M_s	AR-Pm(1)-Pm(3)	Classifier	AOT	AR-Pm(1)-Pm(3)
BLSSVMrbf-END	.648	1.5-.61-.99	LR-PC	.347	1.8-.46-.98
LR-PC	.640	2.1-.28-.93	BLSSVMrbf-END	.345	1.8-.43-.98
BLSSVMrbf-PC	.630	3.1-.01-.61	BLSSVMrbf-PC	.317	3.3-.01-.56
MLR	.626	3.3-.09-.47	MLR	.315	3.3-.10-.49
BLSSVMlin-PC	.599	5.2-.00-.00	MKLR	.273	5.1-.00-.00
MKLR	.593	5.7-.00-.00	BLSSVMlin-PC	.256	5.8-.00-.00
BLSSVMlin-END	.558	7.0-.00-.00	BLSSVMlin-END	.190	7.0-.00-.00
BPER	.489	8.0-.00-.00	BPER	.128	8.3-.00-.00
BMLP	.401	9.0-.00-.00	BMLP	.121	8.7-.00-.00

Pm(3). For example, there is an estimated chance of 80% that one of these two is the best of all classifiers. MKLR comes in third, closely followed by BLSSVMrbf-END. The ranking results for M are highly similar, which is not surprising since the AUCs for failing PUL and IUP are nearly identical for all classifiers. The ranking for both weighted AUC metrics placed BLSSVMrbf-END and LR-PC clearly at

the top, followed by BLSSVMrbf-PC and MLR. Overall, LR-PC has produced a classifier that performed very well on all performance measures. MKLR, MLR, and BLSSVMrbf-END also performed well, but BLSSVMrbf-END is clearly better with respect to weighted AUC metrics.

To turn the probabilities into crisp predictions, we could weight the multi-class probabilities and choose the outcome with the largest weighted probability. This only requires the choice of the weights for each outcome. However, a more easily applied system was used, in which first a threshold probability for predicting ectopic pregnancy was used, followed by a threshold probability for predicting IUP. For the LR-PC classifier, we chose to predict ectopic pregnancy if the probability for this outcome was at least 0.1. Else, IUP was predicted if the probability for this outcome was at least 0.5. Else, failing PUL was predicted. Table 8.4 presents the results. In the test set, this system correctly identified 82% of the failing PULs, 80% of the IUPs, and 83% of the ectopic pregnancies.

Table 8.4: Cross-tabulation of true and predicted outcome using LR-PC.

	Training set predictions (n, %)			Test set predictions (n, %)		
True class	Failing	IUP	Ectopic	Failing	IUP	Ectopic
Failing	249, 88.0%	5, 1.8%	29, 10.2%	146, 82.5%	1, 0.6%	30, 16.9%
IUP	2, 1.1%	144, 78.7%	37, 20.2%	0, 0.0%	117, 79.6%	30, 20.4%
Ectopic	1, 2.4%	3, 7.1%	38, 90.5%	2, 8.3%	2, 8.3%	20, 83.3%

Calibration plots for MLR, LR-PC, MKLR, and BLSSVMrbf-END are given in Figure 8.3. The probabilities given by these classifiers were acceptable. No substantial differences between classifiers were observed for failing PUL, even though MKLR seems to have the best correspondence between probability and true outcome. In general, the probabilities tend to be slightly underestimated for failing PULs. For IUP, BLSSVMrbf-END's probabilities seem to correspond best to the outcome. One can observe here that low probabilities are overestimated while high probabilities are slightly underestimated. With respect to ectopic pregnancies, the probabilities tend to be slightly overestimated, with MKLR having the best result.

Since this work is an extension of three previous studies in which multi-class logistic regression models for PULs were developed [104, 91, 90], we applied these classifiers to the present test set (Table 8.5). As mentioned before, the three old models (M1, M4, M5) were developed on a part of the training set used in this section. When comparing the performances of these models with that of the present model (MLR), it is clear that the old models have less good performance. For AUC_{EP}, the differences between MLR and the old models were 0.070 (95% CI: -0.003 to 0.188) for M1, 0.028 (95% CI: -0.036 to 0.127) for M4, and 0.034 (95% CI: -0.029 to 0.125) for M5. Differences for M yielded similar results since differences in this index were almost exclusively accounted for by differences in AUC_{EP}. For,

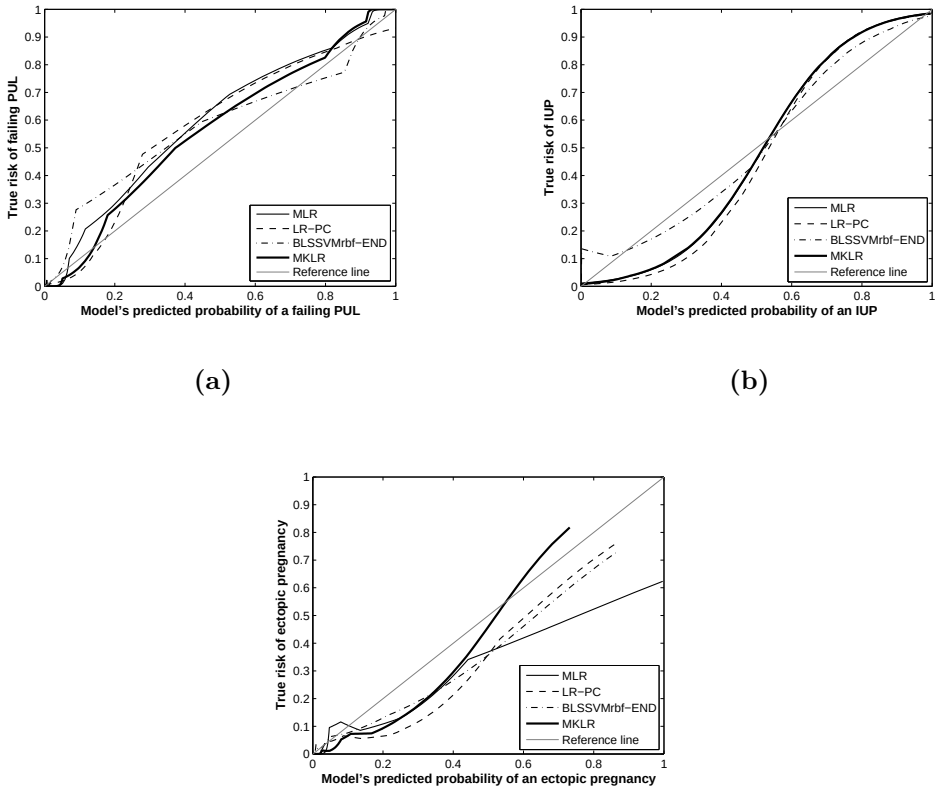


Figure 8.3: Calibration curves of the multi-class classifiers for PULs; curves for the probability of (a) a failing PUL, (b) an IUP, and (c) an EP.

AOT, differences were 0.148 (95% CI: 0.092 to 0.256) for M1, 0.085 (95% CI: 0.024 to 0.197) for M4, and 0.128 (95% CI: 0.074 to 0.233) for M5.

Table 8.5: Test set results for the previously developed logistic regression models.

	AUC	pAUC	Sens75	Multi-class metrics			
Classifier	Failing-IUP-EP	Failing-IUP-EP	Failing-IUP-EP	VUS_2	M	AOT	M_s
M1	.979-.978-.850	.236-.228-.161	99.4-100-87.5	.756	.918	.167	.536
M4	.985-.977-.892	.235-.233-.184	100-98.6-91.7	.800	.938	.230	.566
M5	.984-.975-.886	.234-.231-.174	98.3-98.0-87.5	.774	.934	.188	.538
MLR	.989-.989-.920	.241-.240-.172	99.4-99.3-91.7	.835	.954	.315	.626

Comparing algorithms using repeated data splitting

The median values of the performance measures are presented in Table 8.6. Again, the AUC, pAUC, and Sens75 values for failing PUL and IUP are not shown. Very good performance was noted for the prediction of both classes and no big differences between algorithms were noted. For failing PUL, these indices were between 0.982 and 0.988, 0.237 and 0.240, and 98.9 and 99.5%, respectively. For IUP, they were between 0.981 and 0.985, 0.231 and 0.235, and 99.2 and 100%, respectively. The linear models (and BMLP) performed poorly on the algorithm level as well. Concerning discriminatory ability (AUC_{EP} and M), LR-PC, MLR, MKLR, and BLSSVMrbf-END performed best. When the probabilities are taken into account as well, MKLR performed much less but BLSSVMrbf-PC performed very well. Overall, LR-PC and MLR were the best algorithms as they were always ranked in the top three.

Table 8.6: Performance results for the PUL algorithms using 100 train-test splits.

	For EP only			Multi-class metrics			
Classifier	AUC	pAUC	Sens75	VUS_2	M	AOT	M_s
MLR	.931	.185	92.3	.834	.955	.290	.613
LR-PC	.929	.187	92.3	.836	.956	.277	.604
BMLP	.910	.170	88.5	.791	.946	.122	.420
BPER	.903	.168	90.4	.782	.944	.138	.496
BLSSVMlin-PC	.912	.174	92.3	.807	.946	.236	.582
BLSSVMlin-END	.901	.170	88.5	.791	.942	.167	.542
BLSSVMrbf-PC	.916	.180	92.3	.813	.949	.280	.614
BLSSVMrbf-END	.924	.183	92.3	.823	.953	.236	.590
MKLR	.927	.181	92.3	.823	.953	.230	.571

For PULs belonging to each outcome class, the average test set probability the classifier predicted for each outcome was computed for each train-test split. The 100 resulting cross-tabulations were averaged into one general confusion table per algorithm (not shown). Except for BMLP and BPER, failing PULs were given on average a probability of being failing PUL between 0.91 and 0.94, IUPs an average probability between 0.89 and 0.92 of really being IUP, and ectopic pregnancies an average probability between 0.24 and 0.38 of really being ectopic. Failing PULs were considered more likely to be ectopic than to be an IUP. IUPs were considered more likely to be ectopic than to be failing. Ectopic pregnancies were given a relatively high probability for both failing PUL and IUP.

8.3.4 Does hCG information suffice or do we need more?

The models developed in this chapter used hCGratio, hCGmean, progmean, age, bleeding, and prevect. These inputs were selected by the model selection algorithms, but this does not mean that they are all truly indispensable for discriminating

Table 8.7: Ranking results for the PUL algorithms based on 100 train-test splits.

Classifier	AUC_{EP}	AR-Pm(1)-Pm(3)	Classifier	M	AR-Pm(1)-Pm(3)
MLR	.931	2.7-.39-.72	LR-PC	.9561	2.3-.35-.80
LR-PC	.929	2.4-.29-.80	MLR	.9547	2.9-.37-.68
MKLR	.927	3.7-.17-.54	MKLR	.9534	3.6-.15-.56
BLSSVMrbf-END	.924	3.6-.08-.50	BLSSVMrbf-END	.9526	3.5-.04-.55
BLSSVMrbf-PC	.916	5.3-.03-.16	BLSSVMrbf-PC	.9490	5.4-.04-.16
BLSSVMlin-PC	.912	5.7-.02-.11	BMLP	.9463	6.3-.03-.13
BMLP	.910	6.7-.00-.10	BLSSVMlin-PC	.9462	6.0-.02-.05
BPER	.903	7.4-.00-.01	BPER	.9439	7.4-.00-.03
BLSSVMlin-END	.901	7.5-.02-.06	BLSSVMlin-END	.9416	7.7-.00-.04
Classifier	M_s	AR-Pm(1)-Pm(3)	Classifier	AOT	AR-Pm(1)-Pm(3)
BLSSVMrbf-PC	.614	2.1-.42-.90	MLR	.2897	2.0-.39-.93
MLR	.613	2.1-.36-.93	BLSSVMrbf-PC	.2805	2.3-.39-.87
LR-PC	.604	2.9-.10-.68	LR-PC	.2775	2.6-.14-.74
BLSSVMrbf-END	.590	3.8-.12-.39	BLSSVMrbf-END	.2360	4.3-.08-.34
BLSSVMlin-PC	.582	4.7-.00-.06	BLSSVMlin-PC	.2356	4.9-.00-.05
MKLR	.571	5.5-.00-.04	MKLR	.2302	5.1-.00-.07
BLSSVMlin-END	.542	6.9-.00-.00	BLSSVMlin-END	.1666	7.0-.00-.00
BPER	.496	8.0-.00-.00	BPER	.1382	8.3-.00-.00
BMLP	.420	9.0-.00-.00	BMLP	.1216	8.6-.00-.00

between all classes. In this section, we try to investigate more thoroughly which information is likely to be sufficient. In [90], we already noted that bleeding was selected as an input, but that this did not resulted in a clear increase in AUC results. Therefore, MKLR models (using the RBF kernel) with three different input sets were fed to the 100 training sets based on repeated data splitting. Each time, the resulting models were applied to the appropriate test set, and various performance measures were computed. The 100 test set measures were summarized by their median. In this way, different input sets can be compared while at the same time superseding the influence of one specific train-test split. The three input sets were: hCGratio and log(hCGmean) (hCG model); hCGratio, log(hCGmean), and log(progmean) (hormonal model), and the full model for MKLR (see inputs in Table 8.3.3). The results are shown in Table 8.8. As already described in Chapter 4, progesterone, despite its correlation with hCGratio, still improved the model. Clinical information mildly improved performance on top of progesterone, even though weighted AUC metrics suggested little improvement.

We further investigated which inputs are sufficient by constructing resampled ROC curves, which consist of a summary of each test set's ROC curve. For various levels of specificity, the average sensitivity over the 100 test set ROC curves was computed. This results in a smoothed curve that estimates the true ROC curve for a given algorithm, independent of a given training and test set. The resampled ROC curves for each outcome and each algorithm are shown in Figure 8.4. These

Table 8.8: Performance results for MKLR algorithms with different input sets using 100 train-test splits.

	median AUC			Multi-class metrics (median)				
Classifier	Failing	IUP	EP	VUS_2	M	AOT	M_s	$wVUS_2$
hCG	.984	.981	.886	.759	.932	.178	.541	.187
Hormonal	.988	.984	.914	.806	.947	.219	.570	.228
Full	.986	.984	.927	.823	.953	.230	.571	.238

show that progesterone was useful to use together with hCG information. Clinical information, on the other hand, appeared superfluous. The resampled ROC curves for the hormonal model and the full model are identical in the crucial top left part of the curve.

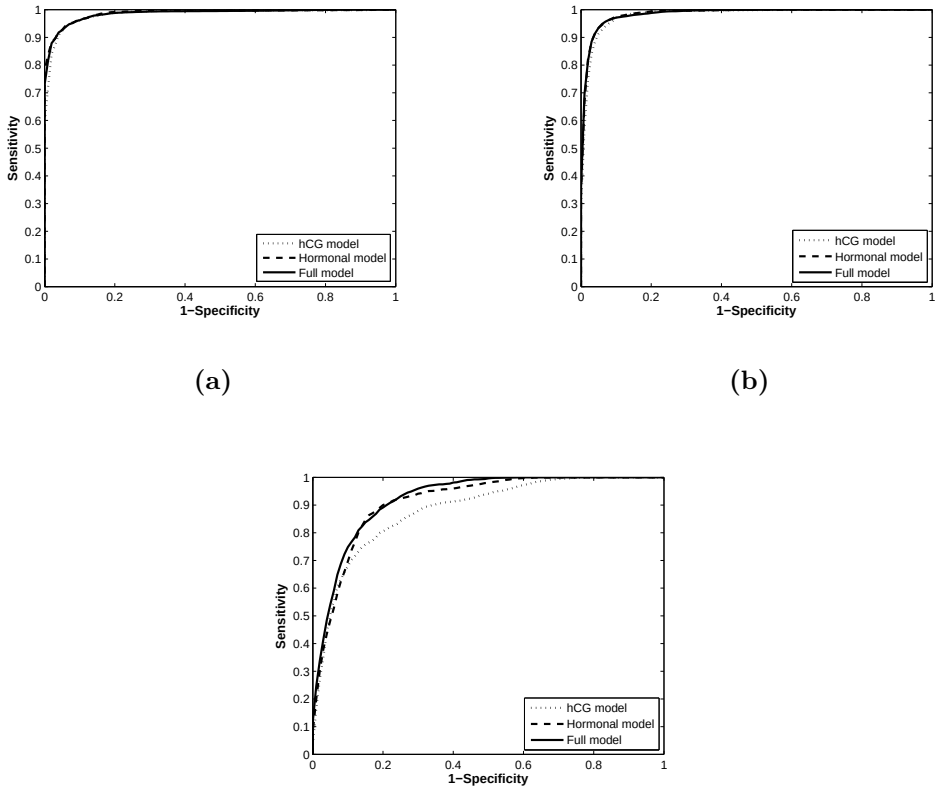


Figure 8.4: Resampled ROC curves for MKLR models with different input sets based on 100 train-test splits: (a) failing PULs, (b) IUPs, and (c) ectopic pregnancies.

8.4 Discussion and conclusions

In this chapter, several multi-class models were used for the early characterization of PULs. The results can be summarized into five general findings. First, on both the classifier and algorithm level, the logistic regression approaches (LR-PC and MLR) appeared to be the best methods. MKLR and BLSSVMrbf-END also performed well, yet MKLR was suboptimal when probabilities were incorporated in the AUC metrics. Second, MKLR performed very well, with nonlinearity being automatically detected and multi-class probabilistic outputs being directly given without Bayesian analysis or combination of binary results. With respect to discriminatory ability, it performed as good as, or better than, the algorithms involving the combination of binary Bayesian least squares support vector machines (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 problems is fruitful. Third, even though differences are small, the performance of LR-PC suggests that combining binary probabilistic models with separate input selection may be useful as compared to one single multi-class model (MLR). This conclusion was also drawn from the multi-class models for ovarian tumors. Fourth, linear approaches (BPER, BLSSVMlin-PC, BLSSVMlin-END) performed less than nonlinear ones. It is useful to detect whether a classification task can be successfully tackled using linear methods because it would result in models that are more interpretable. Fifth, among PULs, failing PULs and IUPs can be predicted with very high level of reliability. Even though the prediction of the potentially dangerous condition [224, 22] of ectopic pregnancy is good, it remains the class that is most difficult to predict. For ectopic pregnancies, the predicted probability of being ectopic is on average around 0.3 to 0.4 such that mildly elevated probabilities for ectopic pregnancy should raise concern. The drawback is that it will cause several failing PULs and IUPs to be wrongly predicted as ectopic. Since ectopic pregnancy can be dangerous for the mother, clinicians may be willing to accept this drawback in the hope of achieving a high sensitivity for ectopic pregnancy.

When evaluating classifiers, often threshold levels 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, thresholds on the probabilities cannot reliably be determined on the training set because they may suffer from overfitting. In practice, however, this is often done. Moreover, thresholds may depend on the distribution of outcomes in a population such that they may differ between centers. It is known that ultrasound scanning abilities can influence the class distribution in PULs (see e.g. [466]). Good ultrasound operators may detect more ectopic pregnancies on ultrasound such that they will have less such pregnancies in their PUL population. Therefore, our threshold system to obtain crisp predictions using the LR-PC classifier may perform poor in other centers. On our test data, this system correctly identifies between 80 and 83% of the cases from each class.

We believe that the distinction between the performance of a classifier and the performance of an algorithm is crucial [129]. A classifier is dependent on the particular set of training cases on which it is developed. When clinicians want to use the predictions of a model in clinical practice, they need a classifier to obtain fixed probability estimates for new cases. Therefore, evaluating classifier performance is important. When one is interested in algorithm comparison, one can resort to repeated data splitting or a related technique.

The test set performance of the classifiers developed in this paper is better than the performance of the old multi-class logistic regression classifiers when applied to the present test set. The use of more data and more advanced algorithms thus resulted in better classifiers that can hopefully result in better decisions. A Bayesian network including expert prior information was previously derived for the prediction of ectopic pregnancy [162]. The network used gestational age, hCG ratio and the progesterone level at 48 hours for its prediction. The test set AUC was 0.88, which is lower than the AUCs obtained for the present classifiers. We have to be careful to compare the AUCs directly, however. Even though the same data set was used for both analyses, the split of the data into a training and test set was different.

An essential issue that is raised by a first prospective multi-center evaluation of a mathematical model is that of the robustness of the models. In this first prospective evaluation [240], M4 was applied to PULs from three new centers in the UK (London and Bedford) and Norway (Oslo). The AUC for failing PUL varied between 0.82 and 0.97, the AUC for IUP between 0.93 and 0.97, and the AUC for EP between 0.74 and 0.79. These are clear decreases in performance as compared to the data from St. George's. However, M4 proved useful for PULs encountered at St. George's Hospital, as observed in a prospective internal interventional study [236]. The study demonstrated that the use of M4 in clinical practice reduced the number of patient visits, scans, blood tests, and interventions, relative to standard clinical practice. As described in Chapter 1, this is a very important result considering the possible use of such a model as a CDS system.

Finally, additional analyses showed that hCG information and progesterone information is likely to be sufficient to diagnose PULs using mathematical models. Other clinical information such as age, history of ectopic pregnancy, and level of vaginal bleeding can be added, as the input analyses showed. Yet, their effect on the model's performance seemed very limited when examining resampled ROC curves. The clinical information that is used in our models is very cheap to collect, however.

Chapter 9

Conclusions and further research

9.1 Overview of results

In medicine, CDS systems can help medical personnel in taking good decisions concerning diagnosis and patient treatment. The development of well-performing CDS systems is a complex task. These systems need to be developed carefully, their operation needs to be clear, and evidence of their satisfactory performance documented. These aspects are relevant in order to convince people of their validity and persuade them to actually put the system into use. In the development process of CDS systems in medical applications, computational intelligence can play an important role. Through the use of mathematical models, robust predictive models for diagnosis and prognosis can be derived. Next to the modeling part of a knowledge discovery process, crucial parts are data collection and performance evaluation. In this thesis, we have developed mathematical models for ovarian tumors and pregnancies of unknown location (PULs). The main focus was on multi-class classification of ovarian tumors as benign, primary invasive, borderline, or metastatic, and of PULs as failing, intra-uterine, or ectopic.

In the first part of Chapter 4, a several methods to obtain multi-class probabilities were compared. All methods were based on LS-SVMs. The data used for this comparison were the training sets of the IOTA phase I data set on ovarian tumors and the data set on PULs used in Chapter 8. We concluded that the Bayesian formulation of the LS-SVM with sparseness induction was a good approach to obtain binary probabilities (as opposed to a similar model without sparseness induction, Platt's method to transform the latent output into probabilities, and isotonic regression). Good methods to combine binary probabilities into multi-class probabilities were the pairwise coupling methods described in [472] and the method using ensembles of nested dichotomies (END) [157]. LS-SVM-based multi-class kernel logistic regression [220] was an all-at-once method with excellent performance, and as such constitutes an interesting option as it does not involve some combination of binary models. The second part of Chapter 4 discussed weighted AUC-based metrics for multi-class classification. Extending suggestions for multi-class AUC

metrics [290, 181] and weighted AUC metrics for binary classification [154, 470], suggestions for weighted multi-class metrics were given. The correlations between these methods was assessed using the PUL data set. Most weighted metrics correlated extremely well with each other (Spearman rho > 0.90) and mildly with unweighted metrics (Spearman rho between 0.40 and 0.60). The most easy measures are the combination of binary weighted AUC metrics. For three-class classification, an interesting visual measure is the area of the triangle spanned by the estimated average probability vectors of each class. These weighted multi-class metrics can be used for model selection and model comparison. For example, when confronted with two models that have similar AUC results but differing weighted AUC results, the model with the best weighted AUC results can be thought of as the most robust model where small perturbations in the predicted probabilities will have least influence on the AUC results [471].

Chapter 5 described binary and multi-class models for ovarian tumors using the training part of the IOTA phase I data set. The binary models aim to distinguish between benign and malignant tumors. All classifiers performed very well on the test set, with AUCs between 0.92 and 0.95. When weighted AUC metrics were considered, the Bayesian LS-SVM classifiers clearly performed best, followed by the relevance vector machine classifiers and a logistic regression model. Overall, the models did not outperform the subjective impression of the expert operators. However, the experts taking part in IOTA phase I were very experienced. It is possible that the models can be of more use for less experienced operators. As an alternative to mathematical models, a limited set of simple rules was developed that could be immediately used by the operator when investigating an ovarian tumor. The rules that were selected could be applied to about three fourth of the tumors, with excellent sensitivity and specificity levels.

The multi-class models aim to distinguish between benign, primary invasive, borderline, and metastatic tumors. The last two classes of ovarian tumors are rather rare, and may not always be easy to distinguish from benign and primary invasive tumors, respectively. Models were built using logistic regression, LS-SVMs and kernel logistic regression. An input selection method for LS-SVMs with linear kernel based on rank-one updates of the kernel matrix (R1U) compared favorably to ARD input selection using either a linear or an RBF kernel. Important variables were age, ascites, acoustic shadows, maximum diameter of the largest solid component, irregular internal cyst walls, personal history of ovarian cancer, maximum diameter of the lesion, presence of blood flow in the papillary structures, number of papillary structures, unilocular tumor, entirely solid tumor, and the presence of bilateral tumors. The classifiers were well able to identify all four types of tumors, considering their AUC results on the test set. The AUCs for both benign and primary invasive tumors varied between 0.93 and 0.94. The AUC for borderline tumors varied between 0.80 and 0.85. The AUC for metastatic tumors varied between 0.90 and 0.92. Classifiers based on a combination (using pairwise coupling) of binary classifiers

had the best discriminatory potential on the test set. Overall, the best classifier was a combination of binary (1-vs-1) logistic regression models using inputs based on the R1U method for LS-SVMs (LR-PC2). When comparing the algorithms rather than the classifiers, similar results were obtained. Combining binary models resulted in higher discriminatory ability. However, when weighted AUC measures were considered, the multi-class logistic regression algorithm performed well too.

In Chapter 6, the models from Chapter 5 were prospectively validated on new data. The IOTA Ib data set contains 507 patients from the three most contributing centers of IOTA phase I. This data set allowed for a first prospective internal validation of the models. The IOTA phase II study contained both centers that participated in IOTA phase I and new centers. Therefore, this data set allowed for both a prospective internal and prospective external validation of the models. Because data cleaning for the IOTA phase II data set is still in progress, a preliminary data set with 1916 patients with complete and correct data was used. The results on these data sets showed that the binary and multi-class models performed very well on both the prospective internal and prospective external validations. Actually, the prospective external validation yielded higher AUCs than the prospective internal validation. Overall, for binary classification, one could opt for a simple linear model such as LR1, ObjLR (which does not use the subjective variables pain and colscore), BLSSVM-lin, or BPER11. In my opinion, BLSSVM-lin has the disadvantage of using a less optimal input set. For multi-class classification, the combination of 1-vs-1 logistic regressions using pairwise coupling (LR-PC2) stood out. LR-PC2 differs from LR-PC with respect to the inputs used. LR-PC2's inputs were based on the R1U method for LS-SVMs with linear kernel. Apparently, this yielded good results.

Chapter 7 analyzed the need to use the CA-125 tumor marker for ovarian tumor diagnosis. The results showed that, notwithstanding the fact that CA-125 on itself conveys a lot of information about tumor malignancy, CA-125 is not strictly necessary.In mathematical models, CA-125 can be replaced by cheaper information based on an ultrasound examination of the tumor. On top of that, the subjective impression of an expert operator clearly outperformed CA-125 in the diagnosis of malignancy. It is confirmed that relying solely on CA-125 for gauging whether a tumor is malignant or not can be dangerous. For example, it was observed that CA-125 levels of benign tumors such as endometriomas or abscesses were slightly higher than the levels of borderline stage I malignancies. Stage I primary invasive ovarian cancers, rare malignancies, metastases, or borderline malignancies of stage II and III typically resulted in higher CA-125 levels than in women with benign or borderline stage I tumors. CA-125 levels were highest for stage II−IV primary invasive ovarian cancers and for stage III borderline tumors.

The models for diagnosing PULs are presented in Chapter 8. Models were built using logistic regression, Bayesian multi-layer perceptrons, Bayesian LS-SVMs, and

kernel logistic regression. Important variables were the hCG ratio, the average hCG level, the average progesterone level, the level of vaginal bleeding, age, and history of ectopic pregnancy. Whereas linear models appeared sufficient for ovarian cancer diagnosis, nonlinear models performed better for diagnosing PULs. The test set AUCs for failing PULs and IUPs were around 0.98-0.99 for all models. These two classes can thus be predicted with very high accuracy. For ectopic pregnancy, the test set AUC varied between 0.87 and 0.92. Concerning discriminatory ability, the logistic regression classifiers and algorithms (with interaction effects and polynomial effects to introduce nonlinearity) were the best, together with the multi-class kernel logistic regression classifier and a combination of binary Bayesian LS-SVM classifiers using RBF kernel. Concerning the weighted AUC metrics, the Bayesian LS-SVM based classifiers and algorithms showed good performance, as they did in all analyses throughout this thesis. Additionally, using resampled ROC curves, we observed that the clinical variables (level of vaginal bleeding, age, history of ectopic pregnancy) do not have much added value over the hCG and progesterone information. It is likely that, out of all measurements considered in the analyses, the hormonal variables suffice for accurate PUL prediction.

In all, it is observed that ovarian tumors and PULs could be accurately diagnosed using the data and variables at hand. For ovarian tumors, linear approaches were sufficient, whereas PULs required more complex models. Logistic regression models, if used well, can result in diagnostic models with a discriminatory ability that is equal to more advanced models. It is necessary, however, to pay ample attention to input selection when using logistic regression. When considering AUC measures that also incorporate the actual predicted probabilities, the Bayesian LS-SVM models showed superior performance. In my opinion, this is a nontrivial argument in favor of their use. Finally, we have often applied all-at-once methods (e.g. multi-class logistic regression) as well as similar methods based on a combination of binary models (e.g. a combination of binary logistic regression models). In the latter approach, inputs were selected for each binary model separately such that inputs are used only where they are needed. Results showed that the combination of binary models performed very well, such that this is an interesting approach to consider when building diagnostic models for multi-class problems. They come, however, with additional computational effort. The good performance of all diagnostic models for ovarian tumors after prospective internal and prospective external validation is an important finding towards the development of a CDS system for ovarian tumor diagnosis.

9.2 Future research

For both medical applications, there are plans to investigate the importance of new markers for diagnosis. These markers will be of biochemical nature, and will stem from areas such as proteomics and genomics. This is an important area of research that may lead to even more discriminating models [58, 119, 379, 469]. At

the same time, it will be interesting to investigate to what extent ensemble and fusion approaches can improve performance. More specifically, it is interesting to apply boosting [353, 89], regularized boosting [335], multi-class boosting extensions [158, 370], or the fusion approaches as applied in [133].

When estimated probabilities are used to generate a crisp, black-and-white prediction, a threshold system is typically used. However, thresholds were developed for the whole set of training data, even if these data stem from multiple centers. Therefore, it is necessary to evaluate to what extent optimal thresholds depend on center. This will be done for ovarian tumors (using the IOTA phase II data set) and PULs (using the IPULA consortium). Also, for multi-class problems, more complex systems are needed to arrive at crisp predictions. More attention can be devoted to this, for example based on explicit cost matrices supplied by clinicians. Such matrices provide information about the seriousness of different types of misclassifications.

For ovarian tumors, it is necessary to perform an extensive performance evaluation on the finalized IOTA phase II database. This also encompasses the evaluation of a strategy in which simple rules are combined with mathematical models. As plans for a third phase are in an advanced stage at the moment, it is clinically very useful to merge the data sets from phases I, Ib, and II in order to obtain data on approximately 3500 ovarian tumors. This data set can be used to derive more reliable models. IOTA phase III aims to collect data on 2000 tumors. These data can be used in the first place to test the existing models (developed on IOTA I or on the joint data from IOTA I, Ib, and II). Additionally, it is the aim to improve performance by applying second stage tests to those cases for which the mathematical models are uncertain. Second stage tests include the use of intravenous contrast agents, proteomics-based indices, new tumor markers (see e.g. [397]), 3D power Doppler indices, and models that are specifically designed for 'difficult' tumors. Of course, the overall usefulness of these second stage tests to distinguish between benign and malignant tumors will be examined as well.

Other issues for future research concerning the development of a CDS system for ovarian tumors deal with the comparison of the models with expert operators with varying levels of experience, and with the examination of the extent to which the use of mathematical models can have real benefits in practice. With respect to the first topic, one could investigate the effect of experience by documenting the approximate number of scans performed by several operators (cf. [391]), or one could examine the implementation of mathematical models in regional medical centers. A CDS system for ovarian tumors will probably needed most by regional centers. Such centers, as opposed to academic medical centers, have less experienced operators that also have less scientific interest in the topic of ovarian tumors. For such a study, one needs to be aware of the effect of substandard data collection on model performance. Therefore, it would be interesting to additionally investigate the effect

of an educational program that is targeted at the scoring of an ultrasound scan. With respect to the second topic, real benefits could be the reduction of morbidity, costs, and hospitalization length for women with a benign ovarian tumor, and improved survival for women with ovarian cancer.

For pregnancies of unknown location, the next step is to apply the models described in Chapter 8 to the prospective multi-center data set collected by the IPULA study group in 7 countries (Belgium, Italy, UK, Australia, Singapore, Canada, USA). Main issues are the robustness of the models and of the threshold system developed for the LR-PC2 classifier. It is possible that performance degradation (as observed in [240]) originates from country- and center-related differences in clinical practice. For example, if pregnant people attend the hospital on average at a higher gestational age in some centers as compared to others, this may have its influence on hCG and progesterone levels. This could initiate a study on the importance of gestational age for building models that successfully apply to different centers worldwide. The IPULA database thus allows the development of new, hopefully more robust classifiers.

Finally, it would be interesting to compare the weighted multi-class AUC metrics from Chapter 4 for model selection. So far, they were applied only for the sake of model evaluation. Also, the resampled ROC curves presented at the end of Chapter 8 are interesting tools to determine the necessity of specific measurements when building classifiers. In that sense, they could be applied to the CA-125 models presented in section 7.2. Other resampling techniques than repeated data-splitting can be investigated, and efforts for creating confidence intervals on the AUC or confidence bands on the whole ROC curve could be undertaken. Recall that the results obtained using repeated data splitting are not independent, such that one cannot directly use them for the construction of confidence intervals.

Appendix: Description of IOTA variables

This appendix presents an overview of the IOTA variables that are used throughout this thesis. The measurement scale is always indicated.

Clinical information

Family history

pershistovca Personal history of ovarian cancer (Binary: 0 - no, 1 - yes)

pershistbrca Personal history of breast cancer (Binary: 0 - no, 1 - yes)

famhistovca Number of family members (first-degree relatives) with history of ovarian cancer (Ordinal)

famhistbrca Number of family members (first-degree relatives) with history of breast cancer (Ordinal)

famovca Family history of ovarian cancer (Binary: 0 - no, 1 - yes)

fambrca Family history of breast cancer (Binary: 0 - no, 1 - yes)

Medical history

age Age of patient at examination (Continuous: in years).

parity Number of children (Ordinal).

nullipara Nulliparous (Binary: 0 - no, 1 - yes).

menoyn Menopausal status (Binary: 0 - pre, 1 - post).

postmenoy Years postmenopausal (Continuous: in years).

pmb Presence of postmenopausal bleeding within one month prior to the ultrasonography (Binary: 0 - no, 1 - yes).

hysterectomy Previous hysterectomy (Binary: 0 - no, 1 - yes).

hormtherapy Current use of hormonal therapy (Binary: 0 - no, 1 - yes).

pain Pelvic pain during examination (Binary: 0 - no, 1 - yes).

Ultrasound characteristics of the tumor

bilateral Presence of masses on both ovaries (Binary: 0 - no, 1 - yes).

ascites Presence of ascites (Binary: 0 - no, 1 - yes).

fluid Amount of fluid in the Pouch of Douglas, measured as the largest anteroposterior diameter in a sagittal plane (Continuous: in mm).

ovmax Maximum diameter of the ovary (Continuous: in mm). This is the largest of the diameters measured in three perpendicular planes.

lesmax Maximum diameter of the lesion (Continuous: in mm). This is the largest of the diameters measured in three perpendicular planes. In the mathematical models, the maximum diameter is often bounded at 60 and 170 mm. Values below 60 are replaced by 60, values above 170 by 170. This bounded version is labeled lesmax2.

lesvol Volume of the lesion (Continuous: in ml). The volume is approximated by multiplying the product of the diameters in three perpendicular planes with $\pi/6000$.

septum Thickness of the thickest septum where it appears to be at its widest (Continuous: in mm). A septum is defined as a thin strand of tissue running across the cyst cavity from one internal surface to the contralateral side.

incomplseptum Presence of an incomplete septum, i.e. a septum that is not complete in some scanning planes (Binary: 0 - no, 1 - yes).

papillation Presence of solid papillary projections (Binary: 0 - no, 1 - yes). Solid papillary projections or structures are defined as solid projections from the cyst wall into the cyst cavity that are at least 3 mm in height.

papnr Number of papillary projections (ordinal: 0 - none, 1 - one, 2 - two, 3 - three, 4 - more than three).

papheight The height of the largest papillary projection (Continuous: in mm).

papflow The presence of blood flow within at least one papillary projection (Binary: 0 - no, 1 - yes).

papsmooth Indicates whether the largest papillary projection is irregular or not (Binary: 0 - smooth, 1 - irregular).

papmax Maximum diameter of the largest papillary structure (Continuous: in mm). This is the largest of the diameters measured in three perpendicular planes.

papvol Volume of the largest papillary structure (Continuous: in ml). The volume is approximated by multiplying the product of the diameters in three perpendicular planes with $frac\pi 6000$.

solidmax Maximum diameter of the largest solid component (Continuous: in mm). This is the largest of the diameters measured in three perpendicular planes. In the mathematical models, the maximum diameter is often bounded at 50 mm. Values above 50 are replaced by 50. This bounded version is labeled solidmax2.

solidvol Volume of the largest solid component (Continuous: in ml). The volume is approximated by multiplying the product of the diameters in three perpendicular planes with $\pi/6000$.

ratiopaples papvol divided by lesvol (Continuous: number between 0 and 1).

ratiosolles solidvol divided by lesvol (Continuous: number between 0 and 1).

nrlocules The number of locules in the whole volume of the tumor (Ordinal: 0 - none, 1 - one, 2 - two, 3 - three, 4 - four, 5 - five to ten, 6 - more than 10).

origin1 Indicates whether the tumor is thought to be of ovarian origin (Binary: 0 - no, 1 - yes).

wallreg Indicates whether the internal cyst wall is irregular or not (Binary: 0 - smooth, 1 - irregular).

shadows Presence of acoustic shadows (Binary: 0 - no, 1 - yes). Acoustic shadows are defined as a loss of acoustic echo behind a sound-absorbing structure.

Locularity

Locularity describes the morphology of the lesion. It can be any of five types, or otherwise it can be non-classifiable. This variable is split up into five dummy variables.

locul1 Unilocular tumor (Binary: 0 - no, 1 - yes). This is defined as a unilocular tumor without septa and without solid parts or papillary structures.

locul2 Unilocular-solid tumor (Binary: 0 - no, 1 - yes). This is defined as a unilocular tumor with a measurable solid component or at least one papillary structure.

locul3 Multilocular tumor (Binary: 0 - no, 1 - yes). This is defined as a tumor with at least one septum but no measurable solid components or papillary projections.

locul4 Multilocular-solid tumor (Binary: 0 - no, 1 - yes). This is defined as a multilocular tumor with a measurable solid component or at least one papillary projection.

locul5 Entirely solid tumor (Binary: 0 - no, 1 - yes). This is defined as a tumor where the solid components comprise at least 80% of the tumor when assessed in a two-dimensional section.

Echogenicity

Echogenicity describes the dominant feature of the cystic contents. It can be any of five types, or it is unclassifiable when there is no cyst fluid. This variable is split up into five dummy variables.

echo1 Anechoic (black) cystic contents.

echo2 Homogeneous low-level echogenic cystic contents.

echo3 Ground glass appearance of cystic contents. In this case, the cystic contents is homogeneously dispersed echogenic.

echo4 Hemorrhagic cystic contents. Cystic contents with internal thread-like structures, representing fibrin strands.

echo5 Mixed cystic contents. In this case, it is possible to describe the echogenicity as star-shaped, cobweb-like (or jelly-like), or mixed echogenic.

Color Doppler imaging and blood flow indices

With color Doppler imaging it is possible to identify the presence of blood flow and vascular changes in new tissue growth. Such changes are imaged as fluctuating areas of color. Calculating indices derived from the Doppler flow velocity waveform gives the possibility to estimate blood flow impedance quantitatively.

colscore A subjective semi-quantitative measurement of the amount of intratumoral blood flow (Ordinal: 1 - no blood flow, 2 - minimal flow, 3 - moderate flow, 4 - very strong flow).

col4 Indicates whether there is very strong intratumoral blood flow (Binary: 0 - no, 1 - yes).

venous Presence of venous blood flow only (Binary: 0 - no, 1 - yes).

psv Peak systolic velocity (Continuous: in cm/s). This is the peak Doppler shifted frequency.

tamx Time-averaged maximum velocity (Continuous: in cm/s). This is the mean Doppler shifted frequency.

pi Pulsatility index (Continuous). This is (psv - minimum Doppler shifted frequency)/tamx.

ri Resistance index (Continuous: between 0 and 1). This is (psv - minimum Doppler shifted frequency)/psv.

Serum tumor marker

ca125 Level of the CA-125 tumor marker (Continuous: in U/ml).

Bibliography

[1] S. Abe. *Support vector machines for pattern classification*. Springer-Verlag, London, 2005.

[2] N.M. Adams and D.J. Hand. Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32:1139–1147, 1999.

[3] N.M. Adams and D.J. Hand. Improving the practice of classifier performance assessment. *Neural Computation*, 12:305–311, 2000.

[4] S. Aerts, P. Antal, D. Timmerman, B. De Moor, and Y. Moreau. Web-based data collection for ovarian cancer: a case study. In P. Kokol, B. Stiglic, M. Zorman, and D. Zazula, editors, *Proceedings of the fifteenth IEEE symposium on Computer Based Medical Systems*, pages 282–287, Los Alamitos, 2002. IEEE Computer Society Press.

[5] A. Agresti. *An introduction to categorical data analysis*. Wiley, New York, 1996.

[6] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[7] A. Albert and J.A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10, 1984.

[8] J.L. Alcázar and M. Jurado. Using a logistic model to predict malignancy of adnexal masses based on menopausal status, ultrsound morphology, and color Doppler findings. *Gynecologic Oncology*, 69:146–150, 1998.

[9] J.L. Alcázar, L.T. Mercé, C. Laparte, M. Jurado, and G. López-Garcia. A new scoring system to differentiate benign from malignant adnexal masses. *American Journal of Obstetrics and Gynecology*, 188:685–692, 2003.

[10] P.D. Allison. *Logistic regression using the SAS system: theory and applications*. SAS Institute, Cary, 1999.

[11] E.L. Allwein, R.E. Schapire, and Y. Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[12] E. Alpaydin. Combined 5×2 CV F test for comparing supervised classification learning algorithms. *Neural Computation*, 11:1885–1892, 1999.

[13] D.G. Altman and P.K. Andersen. Bootstrap investigation of the stability of the Cox regression model. *Statistics in Medicine*, 8:771–783, 1989.

[14] D.G. Altman and J.M. Bland. Absence of evidence is not evidence of absence. *British Medical Journal*, 311:485, 1995.

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

[16] F. Amant, A. De Knijf, B. Van Calster, K. Leunen, P. Neven, P. Berteloot, I. Vergote, S. Van Huffel, and P. Moerman. In contrast to surgical castration and lymphadenectomy, adjuvant hormonal treatment should be considered in the treatment for endometrial stromal sarcoma. Poster presented at the 15th International Meeting of the European Society of Gynaecological Oncology, Berlin (Germany), Oct. 2007.

[17] F. Amant, A. De Knijf, B. Van Calster, K. Leunen, P. Neven, P. Berteloot, I. Vergote, S. Van Huffel, and P. Moerman. Clinical study investigating the role of lymphadenectomy, surgical castration and adjuvant hormonal treatment in endometrial stromal sarcoma. *British Journal of Cancer*, 97:1194–1199, 2007.

[18] American Cancer Society. Learn about breast cancer. Webpage, American Cancer Society Inc., USA, 2007. Available at http://www.cancer.org/ (accessed on Sep. 21, 2007).

[19] American Cancer Society. Learn about ovarian cancer. Webpage, American Cancer Society Inc., USA, 2007. Available at http://www.cancer.org/ (accessed on Sep. 21, 2007).

[20] L. Ameye. *Predictive models for classification in gynecology.* PhD thesis, Department of Electrical Engineering (ESAT-SISTA), Katholieke Universiteit Leuven, Belgium, July 2005.

[21] L. Ameye, C. Lu, L. Lukas, J. De Brabanter, J.A.K. Suykens, S. Van Huffel, H. Daniels, G. Naulaers, and H. Devlieger. Prediction of mental development of preterm newborns at birth time using LS-SVM. In *Proceedings of the tenth Eureopean Symposium on Artificial Neural Networks (ESANN2002)*, pages 167–172, Bruges (Belgium), 2002.

[22] F.W.J. Anderson, J.G. Hogan, and R. Ansbacher. Sudden death: ectopic pregnancy mortality. *Obstetrics & Gynecology*, 103:1218–1223, 2004.

[23] W.M. Ankum, P.J. Hajenius, L.S. Schrevel, and F. Van der Veen. Management of suspected ectopic pregnancy. Impact of new diagnostic tools in 686 consecutive cases. *Journal of Reproductive Medicine*, 41:724–728, 1996.

[24] R. Ansbacher, E.M. Mills, J.C. Thrush, and L.D. Stevenson. Ectopic pregnancy and maternal mortality in Michigan. *American Journal of Gynecological Health*, 3:118–123, 1989.

[25] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, 30:257–281, 2004.

[26] N. Aslam, S. Banerjee, J.V. Carr, M. Savvas, R. Hooper, and D. Jurkovic. Prospective evaluation of logistic regression models for the diagnosis of ovarian cancer. *Obstetrics & Gynecology*, 96:75–80, 2000.

[27] A. Asuncion and D.J. Newman. UCI machine learning repository. Department of Information Science, University of California at Irvine, USA, 2007. Accessible at http://www.ics.uci.edu/∼mlearn/MLRepository.html (accessed on Sep. 7, 2007).

[28] M.S.H. Aung, P.J.G. Lisboa, T.A.E. Etchells, A.C. Testa, B. Van Calster, S. Van Huffel, L. Valentin, and D. Timmerman. Comparing analytical decision support models through Boolean rule extraction: a case study of ovarian tumor malignancy. In D. Liu, S. Fei, Z.G. Hou, H. Zhang, and C. Sun, editors, *Advances in Neural Networks: ISNN2007*, volume 4492 of *Lecture Notes in Computer Science*, pages 1177–1186. Berlin, Springer, 2007.

[29] M. Ayer, H. Brunk, G. Ewing, W. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26:641–647, 1955.

[30] J.A. Baker, K. Boakye, S.E. McCann, G.P. Beehler, K.J. Rodabaugh, J.A. Villella, and K.B. Moysich. Consumption of black tea or coffee and risk of ovarian cancer. *International Journal of Gynecological Cancer*, 17:50–54, 2007.

[31] I.J. Bakken and F.E. Skjeldestad. Incidence and treatment of extrauterine pregnancies in Norway 1990-2001. *Tidsskrift for den Norske Lægeforening*, 123:3016–3020, 2003.

[32] I.J. Bakken and F.E. Skjeldestad. Reduced number of extrauterine pregnancies - increased fertility of women during the 1990s? *Tidsskrift for den Norske Lægeforening*, 123:3011–3014, 2003.

[33] I.J. Bakken and F.E. Skjeldestad. Time trends in ectopic pregnancies in a Norwegian county 1970-2004 - a population-based study. *Human Reproduction*, 21:3132–3136, 2006.

[34] S. Banerjee, N. Aslam, B. Woelfer, A. Lawrence, J. Elson, and D. Jurkovic. Expectant management of early pregnancies of unknown location: a prospective evaluation of methods to predict spontaneous resolution of pregnancy. *British Journal of Obstetrics and Gynaecology*, 108:158–163, 2001.

[35] S. Banerjee, N. Aslam, N. Zosmer, B. Woelfer, and D. Jurkovic. The expectant management of women with early pregnancy of unknown location. *Ultrasound in Obstetrics and Gynecology*, 14:231–236, 1999.

[36] D.J. Barker, J.G Eriksson, T. Forsén, and C. Osmond. Fetal origins of adult disease: strength of effects and biological basis. *International Journal of Epidemiology*, 31:1235–1239, 2002.

[37] K. Barnhart, M.T. Mennuti, I. Benjamin, S. Jacobson, D. Goodman, and C. Coutifaris. Prompt diagnosis of ectopic pregnancy in an emergency department setting. *Obstetrics & Gynecology*, 84:1010–1015, 1994.

[38] K. Barnhart, M.D. Sammel, K. Chung, L. Zhou, A.C. Hummel, and W. Guo. Decline of serum human chorionic gonadotropin and spontaneous complete abortion: defining the normal curve. *Obstetrics & Gynecology*, 104:975–981, 1994.

[39] K.T. Barnhart, M.D. Sammel, P.F. Rinaudo, L. Zhou, A.C. Hummel, and W. Guo. Symptomatic patients with an early viable intrauterine pregnancy: hCG curves redefined. *Obstetrics & Gynecology*, 104:50–55, 2004.

[40] K.T. Barnhart, H. Simhan, and S.A. Kamelle. Diagnostic accuracy of ultrasound above and below the beta-hCG discriminatory zone. *Obstetrics & Gynecology*, 94:583–587, 1999.

[41] R.M. Baron and D.A. Kenny. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182, 1986.

[42] R.C. Bast, T.L. Klug, E. St John, E. Jenison, J.M. Niloff, H. Lazarus, R.S. Berkowitz, T. Leavitt, C.T. Griffiths, L. Parker, V.R. Zurawski Jr, and R.C. Knapp. A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. *The New England Journal of Medicine*, 309:883–887, 1983.

[43] D.W. Bates, G.J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, C. Spurr, R. Khorasani, M. Tanasijevic, and B. Middleton. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association*, 10:523–530, 2003.

[44] Belgian Cancer Registry. Belgium: females, age-specific and age-standardised incidence rates per localisation and age group in 2003. Webpage, Stichting Kankerregister, Belgium, 2007. Available at http://www.kankerregister.org/ (accessed on Sep. 21, 2007).

[45] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

[46] J.O. Berger. *Statistical design theory and Bayesian analysis*. Springer-Verlag, Berlin, 1985.

[47] J. Berkson. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33:526–542, 1938.

[48] D.A. Berry. *Statistics: a Bayesian perspective*. Duxbury Press, Belmont, 1996.

[49] R. Biagiotti, C. Desii, E. Vanzi, and G. Gacci. Predicting ovarian malignancy: application of artificial neural networks to transvaginal and color Doppler flow US. *Journal of the American Statistical Association*, 33:526–542, 1938.

[50] C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, Oxford, 1995.

[51] C.M. Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006.

[52] C.M. Bishop and M.E. Tipping. Bayesian regression and classification. In J. Suykens, G. Horváth, S. Basu, C. Micchelli, and J. Vandewalle, editors, *Advances in learning theory: methods, models and applications*, pages 267–288. IOS Press, Amsterdam, 2003.

[53] G.G. Bon, P. Kenemans, R. Verstraeten, G.J. van Kamp, and J. Hilgers. Serum tumor marker immunoassays in gynecologic oncology: establishment of reference values. *American Journal of Obstetrics and Gynecology*, 174:107–114, 1996.

[54] R.R. Bouckaert. Choosing between two learning algorithms based on calibrated tests. In *Proceedings of the 20th International Conference on Machine Learning (ICML2003)*, pages 51–58. Menlo Park, AAAI Press, 2003.

[55] S. Boufous, M. Quartararo, M. Mohsin, and J. Parker. Trends in the incidence of ectopic pregnancy in New South Wales between 1990-1998. *The Australian and New Zealand Journal of Obstetrics and Gynaecology*, 41:436–438, 2001.

[56] J. Bouyer, J. Coste, H. Fernandez, J.-L. Pouly, and N. Job-Spira. Sites of ectopic pregnancy: a 10 year population-based study of 1800 cases. *Human Reproduction*, 17:3224–3230, 2002.

[57] G.E.P. Box and P.W. Tidwell. Transformations of the independent variables. *Technometrics*, 4:531–550, 1962.

[58] E.A. Boyce and E.C. Kohn. Ovarian cancer in the proteomics era: diagnosis, prognosis, and therapeutics targets. *International Journal of Gynecological Cancer*, 15:266–273, 2005.

[59] R. Brachman and T. Anand. The process of knowledge discovery in databases: a human centered approach. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in knowledge discovery and data mining*, pages 37–58. MIT Press, Cambridge, 1996.

[60] A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[61] R.A. Bradley and M. Terry. The rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345, 1952.

[62] P.B. Brazdil and C. Soares. A comparison of ranking methods for classification algorithm selection. In R. López de Mántaras and E. Plaza, editors, *Machine learning: ECML2000*, volume 1810 of *Lecture Notes in Computer Science*, pages 63–75, Berlin, 2000. Springer.

[63] Breastcancer.org. Symptoms and diagnosis. Webpage, breastcancer.org, USA, 2007. Available at http://www.breastcancer.org/ (accessed on Sep. 21, 2007).

[64] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Chapman & Hall, New York, 1984.

[65] G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

[66] C.E. Brodley and P. Smyth. Applying classification algorithms in practice. *Statistics and Computing*, 7:45–56, 1997.

[67] K.P. Burnham and D.R. Anderson. *Model selection and inference: a practical information-theoretic approach*. Springer, New York, 1998.

[68] A. Burton and D.G. Altman. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer*, 91:4–8, 2004.

[69] B. Cacciatore, U.H. Stenman, and P. Ylöstalo. Diagnosis of ectopic pregnancy by vaginal ultrasonography in combination with a discriminatory serum hCG level of 1000 IU/L (IRP). *British Journal of Obstetrics and Gynaecology*, 95:497–502, 1990.

[70] B. Cacciatore, U.H. Stenman, and P. Ylöstalo. Early screening for ectopic pregnancy in high-risk symptom-free women. *Lancet*, 343:517–518, 1994.

[71] B. Cacciatore, P. Ylöstalo, U.H. Stenman, and O. Widholm. Suspected ectopic pregnancy: Ultrasound findings and hCG levels assessed by an immunofluorometric assay. *British Journal of Obstetrics and Gynaecology*, 95:497–502, 1988.

[72] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[73] M.E. Carley, C.J. Klingele, J.B. Gebhart, M.J. Webb, and T.O. Wilson. Laparoscopy versus laparotomy in the management of benign unilateral adnexal masses. *Journal of the American Association of Gynecologic Laparoscopists*, 9:321–326, 2002.

[74] R. Carver. The case against statistical significance testing. *Harvard Educational Review*, 48:378–399, 1978.

[75] G.C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proceedings of the 19th International Joint Conference on Neural Networks (IJCNN2006)*, pages 2970–2977, Vancouver (Canada), 2006.

[76] G.C. Cawley and N.L. Talbot. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17:1467–1475, 2004.

[77] Centers for Disease Control and Prevention. Ectopic pregnancy - United States, 1990-1992. *MMWR*, 44:46–48, 1995.

[78] C. Chatfield. Model uncertainty, data mining, and statistical inference. *Journal of the Royal Statistical Society A*, 158:419–466, 1995.

[79] W.H. Chow, J.R. Daling, W. Cates Jr, and R.S. Greenberg. Epidemiology of ectopic pregnancy. *Epidemiologic Reviews*, 9:70–94, 1987.

[80] A. Christmann and P.J. Rousseeuw. Measuring overlap in logistic regression. *Computational Statistics & Data Analysis*, 37:65–75, 2001.

[81] G.P. Chrousos and P.W. Gold. Editorial: a healthy body in a healthy mind - and vice versa - the damaging power of "uncontrollable" stress. *Journal of Clinical Endocrinology and Metabolism*, 83:1842–1845, 1998.

[82] R.D. Clayton, S. Snowden, M.J. Weston, O. Mogensen, J. Eastaugh, and G. Lane. Neural networks in the diagnosis of malignant ovarian tumours. *British Journal of Obstetrics and Gynaecology*, 106:1078–1082, 1999.

[83] W.S. Cleveland, S.J. Devlin, and E. Grosse. Regression by local fitting. *Journal of Econometrics*, 37:87–114, 1988.

[84] J. Cohen. The cost of dichotomization. *Applied Psychological Measurement*, 7:249–253, 1983.

[85] J. Cohen. Things I have learned (so far). *The American Psychologist*, 45:1304–1312, 1990.

[86] J. Cohen. The earth is round ($p < .05$). *The American Psychologist*, 49:997–1003, 1994.

[87] E. Coiera. Question the assumptions. In P. Barahona and J.P. Christensen, editors, *Knowledge and decisions in health telematics - the next decade*, pages 61–66. IOS Press, Amsterdam, 1994.

[88] E. Coiera. *Guide to health informatics*. Hodder Arnold, London, 2nd edition, 2003.

[89] M. Collins, R.E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48:253–285, 2002.

[90] G. Condous, B. Van Calster, E. Kirk, Z. Haider, D. Timmerman, S. Van Huffel, and T. Bourne. Clinical information does not improve the performance of mathematical models in predicting the outcome of pregnancies of unknown location. *Fertility and Sterility*, 88:572–580, 2007.

[91] G. Condous, B. Van Calster, E. Kirk, Z. Haider, D. Timmerman, S. Van Huffel, and T. Bourne. Prediction of ectopic pregnancy in women with a pregnancy of unknown location. *Ultrasound in Obstetrics and Gynecology*, 29:680–687, 2007.

[92] G. Condous, E. Kirk, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Failing pregnancies of unknown location: a prospective evaluation of the human chorionic gonadotrophin ratio. *British Journal of Obstetrics and Gynaecology*, 113:521–527, 2006.

[93] G. Condous, E. Kirk, Z. Haider, B. Van Calster, and T. Bourne. Can we use the ultrasonographic appearance of an ectopic pregnancy to predict the likelihood of success for expectant and medical management? Talk given at the 35th Annual Scientific Meeting of the Australasian Society for Ultrasound in Medicine, Adelaide (Australia), Oct. 2005.

[94] G. Condous, E. Kirk, Z. Haider, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Can we improve the performance of diagnostic tests to predict the outcome of pregnancies of unknown location? Talk given at the 35th Annual Scientific Meeting of the Australasian Society for Ultrasound in Medicine, Adelaide (Australia), Oct. 2005.

[95] G. Condous, E. Kirk, Z. Haider, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Can we predict the outcome of medical management of ectopic pregnancies earlier than one week? Talk given at the 35th Annual Scientific Meeting of the Australasian Society for Ultrasound in Medicine, Adelaide (Australia), Oct. 2005.

[96] G. Condous, E. Kirk, Z. Haider, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Can we improve the performance of logistic regression analysis for predicting the outcome of pregnancies of unknown location (PULs)? *Ultrasound in Obstetrics and Gynecology*, 26:455, 2005. Abstract presented at ISUOG2005, Vancouver (Canada).

[97] G. Condous, E. Kirk, Z. Haider, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. What is the optimal approach to accurately classifying failing pregnancies of unknown location (PULs)? *Ultrasound in Obstetrics and Gynecology*, 26:362, 2005. Abstract presented at ISUOG2005, Vancouver (Canada).

[98] G. Condous, E. Kirk, C. Lu, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. There is no role for uterine curettage in the contemporary diagnostic work-up of women with a pregnancy of unknown location. *Human Reproduction*, 21:2706–2710, 2006.

[99] G. Condous, E. Kirk, C. Lu, S. Van Huffel, O. Gevaert, B. De Moor, F. De Smet, D. Timmerman, and T. Bourne. Diagnostic accuracy of varying discriminatory zones for the prediction of ectopic pregnancy in women with a pregnancy of unknown location. *Ultrasound in Obstetrics and Gynecology*, 26:770–775, 2005.

[100] G. Condous, E. Kirk, A. Syed, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Do levels of serum cancer antigen 125 and creatine kinase predict the outcome in pregnancies of unknown location? *Human Reproduction*, 20:3348–3354, 2005.

[101] G. Condous, C. Lu, S. Van Huffel, D. Timmerman, and T. Bourne. Human chorionic gonadotrophin and progesterone levels in pregnancies of unknown location. *International Journal of Gynaecology and Obstetrics*, 86:351–357, 2004.

[102] G. Condous, E. Okaro, and T. Bourne. Ultrasound diagnosis and management of pregnancy of unknown location and ectopic pregnancy. In D. Timmerman, J. Deprest, and T.H. Bourne, editors, *Ultrasound and endoscopic surgery in obstetrics and gynecology*. Springer, Berlin, 2002.

[103] G. Condous, E. Okaro, A. Khalid, C. Lu, S. Van Huffel, D. Timmerman, and T. Bourne. The accuracy of transvaginal ultrasonography for the diagnosis of ectopic pregnancy prior to surgery. *Human Reproduction*, 20:1404–1409, 2005.

[104] G. Condous, E. Okaro, A. Khalid, D. Timmerman, C. Lu, Y. Zhou, S. Van Huffel, and T. Bourne. The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location. *Human Reproduction*, 19:1900–1910, 2004.

[105] G. Condous, E. Okaro, A. Khalid, Y. Zhou, C. Lu, S. Van Huffel, D. Timmerman, and T. Bourne. Role of biochemical and ultrasonographic indices in the management of pregnancies of unknown location. *Ultrasound in Obstetrics and Gynecology*, 20 Suppl 1:36–37, 2002. Abstract presented at ISUOG2002, New York (USA).

[106] G. Condous, D. Timmerman, S. Goldstein, L. Valentin, D. Jurkovic, and T. Bourne. Pregnancies of unknown location: consensus statement. *Ultrasound in Obstetrics and Gynecology*, 28:121–122, 2006.

[107] G. Condous, B. Van Calster, E. Kirk, D. Timmerman, S. Van Huffel, and T. Bourne. Prospective cross-validation of three methods of predicting failing pregnancies of unknown location. *Human Reproduction*, 22:1156–1160, 2007.

[108] G. Condous, B. Van Calster, S. Van Huffel, and A. Lam. What is the value of preoperative bimanual pelvic examination in women undergoing laparoscopic total hysterectomy? *Journal of Minimally Invasive Gynecology*, 14:334–338, 2007.

[109] G. Condous, Y. Zhou, E. Okaro, A. Khalid, M. Alkatib, S. Rao, C. Lu, S. Van Huffel, D. Timmerman, and T. Bourne. Prediction of ectopic pregnancy with the hCG ratio: A new biochemical marker. *Human Reproduction*, 18 Suppl 1:95–96, 2003. Abstract presented as ESHRE2003, Madrid (Spain).

[110] G.S. Condous. Ultrasound diagnosis of ectopic pregnancy. *Seminars on Reproductive Medicine*, 25:85–91, 2007.

[111] B.C. Cooper, A.K. Sood, C.S. Davis, J.M. Ritchie, J.I. Sorosky, B. Anderson, and R.E. Buller. Preoperative ca125 levels: an independent prognostic factor for epithelial ovarian cancer. *Obstetrics & Gynecology*, 100:59–64, 2002.

[112] J. Coste, N. Job-Spira, B. Aublet-Cuvelier, E. Germain, E. Glowaczower, H. Fernandez, and J.L. Pouly. Incidence of ectopic pregnancy. First results of a population-based register in France. *Human Reproduction*, 9:742–745, 1994.

[113] M.D. Creinin and A.E. Washington. Cost of ectopic pregnancy management: surgery versus methotrexate. *Fertility and Sterility*, 60:963–969, 1993.

[114] N. Cristianini and J. Shawe-Taylor. *Support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, 2000.

[115] R. Dart, L. Dart, M. Segal, C. Page, and J. Brancato. The ability of a single serum progesterone value to identify abnormal pregnancies in patients with beta-human chorionic gonadotropin values less than 1,000 mIU/mL. *Academic Emergency Medicine*, 5:304–309, 1998.

[116] R.G. Dart, L. Dart, P. Mitchell, and C. Berty. The predictive value of endometrial stripe thickness in patients with suspected ectopic pregnancy who have an empty uterus at ultrasonography. *Academic Emergency Medicine*, 6:602–608, 1998.

[117] R.G. Dart, B. Kaplan, and C. Cox. Transvaginal ultrasound in patients with low beta-human chorionic gonadotropin values: how often is the study diagnostic? *Annals of Emergency Medicine*, 30:135–140, 1997.

[118] F. De Smet, J. De Brabanter, M.L. Konstantinovic, N. Pochet, T. Van den Bosch, P. Moerman, B. De Moor, I. Vergote, and D. Timmerman. New models to predict depth of infiltration in endometrial carcinoma based on transvaginal ultrasonography. *Ultrasound in Obstetrics and Gynecology*, 27:664–671, 2006.

[119] F. De Smet, N. Pochet, K. Engelen, T. Van Gorp, P. Van Hummelen, K. Marchal, F. Amant, D. Timmerman, B. De Moor, and I. Vergote. Predicting the clinical behavior of ovarian cancer from gene expression profiles. *International Journal of Gynecological Cancer*, 15:1–5, 2005.

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

[121] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[122] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[123] H.M. den Hartog, N.A. Nicolson, M.M. Derix, A.L. van Bemmel, B. Kremer, and J. Jolles. Salivary cortisol patterns and cognitive speed in major depression: a comparison with allergic rhinitis and healthy control subjects. *Biological Psychology*, 63:1–14, 2003.

[124] P.D. DePriest, D. Shenson, A. Fried, J.E. Hunter, S.J. Andrews, H.H. Gallion, E.J. Pavlik, R.J. Kryscio, and J.R. van Nagell Jr. A morphology based index based on sonographic findings in ovarian cancer. *Gynecologic Oncology*, 51:7–11, 1993.

[125] S. Derksen and H.J. Keselman. Backward, forward and stepwise automated subset algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 42:265–282, 1992.

[126] A. Devos, A.W. Simonetti, M. van der Graaf, L. Lukas, J.A.K. Suykens, L. Vanhamme, L.M.C. Buydens, A. Heerschap, and S. Van Huffel. The use of multivariate MR imaging intensities versus metabolic data from MR spectroscopic imaging for brain tumour classification. *Journal of Magnetic Resonance*, 173:218–228, 2005.

[127] E. D'Hondt, G. Moulin-Romsee, F. Persyn, B. Van Calster, P. Neven, and S. Stroobants. Detection of sentinel lymph nodes in breast cancer: everyday practice and pitfalls. Talk given at the 15th International Meeting of the European Society of Gynaecologic Oncology, Berlin (Germany), Oct. 2007.

[128] T. Dietterich. Machine learning. In L. Nadel, editor, *Encyclopedia of cognitive science*. MacMillan, London, 2003.

[129] T.G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[130] T.G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[131] I. Dimou, B. Van Calster, S. Van Huffel, D. Timmerman, and M. Zervakis. Evaluation of imputation methods in ovarian tumour diagnostic models using generalised linear models and support vector machines. *Medical Decision Making*, submitted.

[132] I. Dimou, B. Van Calster, S. Van Huffel, D. Timmerman, and M. Zervakis. Comparison of imputation approaches in ovarian tumour diagnostic models based on LS-SVMs. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, Plymouth (UK), 2007.

[133] I.N. Dimou, G.C. Manikis, and M.E. Zervakis. Classifier fusion approaches for diagnostic cancer models. In *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS2006)*, pages 5334–5337, New York City (USA), 2006.

[134] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35:352–359, 2002.

[135] S. Dreiseitl, L. Ohno-Machado, and M. Binder. Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making*, 20:323–331, 2000.

[136] C. Drummond and R.C. Holte. Cost curves: an improved method for visualizing classifier performance. *Machine Learning*, 65:95–130, 2006.

[137] K. Duan, S.S. Keerthi, W. Chu, S.K. Shevade, and A.N. Poo. Multi-category classification by soft-max combination of binary classifiers. In T. Windeatt and F. Roli, editors, *Multiple classifier systems*, volume 2709 of *Lecture Notes in Computer Science*, pages 125–134, Berlin, 2003. Springer.

[138] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, New York, 2nd edition, 2001.

[139] D.E. Duffy and T.J. Santner. On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistics - Theory and Methods*, 18:959–980, 1989.

[140] B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

[141] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 2nd edition, 1993.

[142] B. Efron and R. Tibshirani. Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association*, 92:548–560, 1997.

[143] M. Egger, N. Low, G.D. Smith, B. Lindblom, and B. Herrmann. Screening for chlamydial infections and the risk of ectopic pregnancy in a county in Sweden: ecological analysis. *British Medical Journal*, 316:1776–1780, 1998.

[144] N. Einhorn, R.C. Bast Jr, R.C. Knapp, B. Tjernberg, and V.R. Zurawski Jr. Preoperative evaluation of serum ca125 levels in patients with primary epithelial ovarian cancer. *Obstetrics & Gynecology*, 67:414–416, 1986.

[145] M.J. Engelen, H.W. de Bruijn, H. Hollema, K.A. ten Hoor, P.H. Willemse, J.G. Aalders, and A.G. van der Zee. Serum CA 125, carcinoembryonic antigen, and CA 19-9 as tumor markers in borderline ovarian tumors. *Gynecologic Oncology*, 78:16–20, 2000.

[146] M.J. Engelen, H.E. Kos, P.H. Willemse, J.G. Aalders, E.G. de Vries, M. Schaapveld, R. Otter, and A.G. van der Zee. Surgery by consultant gynecologic oncologists improves survival in patients with ovarian carcinoma. *Cancer*, 106:589–598, 2006.

[147] M. Espinoza, J.A.K. Suykens, R. Belmans, and B. De Moor. Electric load forecasting: using kernel-based modeling for nonlinear system identification. *IEEE Control Systems Magazine*, 27:43–56, 2007.

[148] F.J. Esteva and G.N. Hortobagyi. Prognostic molecular markers in early breast cancer. *Breast Cancer Research*, 6:109–118, 2004.

[149] T.A. Etchells and P.J.G. Lisboa. Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach. *IEEE Transactions on Neural Networks*, 17:374–384, 2006.

[150] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in knowledge discovery and data mining*, pages 1–30. MIT Press, Cambridge, 1996.

[151] R.J. Feise. Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology*, 2:paper 8, 2002.

[152] J. Ferlay, F. Bray, P. Pisani, and D.M. Parki. *GLOBOCAN 2002: cancer incidence, mortality and prevalence worldwide (IARC CancerBase No 5 version 2.0)*. IARCPress, Lyon, 2004.

[153] E. Ferrazzi, G. Zanetta, D. Dordoni, N. Berlanda, R. Mezzopane, and G. Lissoni. Transvaginal ultrasonographic characterization of ovarian masses: comparison of five scoring systems in a multicenter study. *Ultrasound in Obstetrics and Gynecology*, 10:192–197, 1997.

[154] C. Ferri, P. Flach, J. Hernández-Orallo, and A. Senad. Modifying ROC curves to incorporate predicted probabilities. In *Proceedings of the 2nd Workshop on ROC Analysis in Machine Learning (ROCML2005)*, Bonn (Germany), 2005.

[155] M. Fieschi, J.-C. Dufour, P. Staccini, J. Gouvernet, and O. Bouhaddou. Medical decision support systems: old dilemmas and new paradigms? *Methods of Information in Medicine*, 42:190–198, 2003.

[156] B.L. Ford. An overview of hot-deck procedures. In W.G. Madow, I. Olkin, and D.B. Rubin, editors, *Incomplete data in sample surveys*, pages 185–207. Academic Press, New York, 1983.

[157] E. Frank and S. Kramer. Ensembles of nested dichotomies for multi-class problems. In *Proceedings of the 21st International Conference on Machine Learning (ICML2004)*, Madison, 2004. Omnipress.

[158] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 38:337–374, 2000.

[159] J.H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.

[160] J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

[161] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman & Hall, Boca Raton, 2nd edition, 2004.

[162] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, and G. Condous. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Human Reproduction*, 21:1824–1831, 2006.

[163] P.D. Gluckman and M.A. Hanson. Living with the past: evolution, development, and patterns of disease. *Science*, 305:1733–1736, 2004.

[164] T.E. Goldner, H.W. Lawson, Z. Xia, and H.K. Atrash. Surveillance for ectopic pregnancy - United States, 1970-1989. *MMWR CDC Surveillance Summaries*, 42(SS6):73–85, 1993.

[165] S. Goldstein, J.R. Snyder, C. Watson, and M. Danon. Very early pregnancy detection with endovaginal ultrasound. *Obstetrics & Gynecology*, 72:200–204, 1988.

[166] S.R. Goldstein. Early detection of pathologic pregnancy by transvaginal sonography. *Journal of Clinical Ultrasound*, 18:262–273, 1990.

[167] S.N. Goodman. Toward evidence-based medical statistics. 2: the Bayes factor. *Annals of Internal Medicine*, 130:1005–1013, 1999.

[168] C.R. Gracia and K.T. Barnhart. Diagnosing ectopic pregnancy: decision analysis comparing six strategies. *Obstetrics & Gynecology*, 97:464–470, 2001.

[169] S. Granberg, A. Norström, and M. Wikland. Tumours in the lower pelvis as imaged by vaginal sonography. *Gynecologic Oncology*, 37:224–229, 1990.

[170] S. Granberg, M. Wikland, and I. Jansson. Macroscopic characterization of ovarian tumours and the relation to the histological diagnosis: criteria to be used for ultrasound evaluation. *Gynecologic Oncology*, 35:139–144, 1989.

[171] Y. Grandvalet and Y. Bengio. Hypothesis testing for cross-validation. Technical report 1285, Département d'Informatique et Recherche Opérationnelle, Université de Montréal, USA, 2006.

[172] M. Green and M. Ohlsson. Comparison of standard resampling methods for performance estimation of artificial neural network ensembles. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, Plymouth (UK), 2007.

[173] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[174] M. Hahlin, J. Thorburn, and I. Bryman. The expectant management of early pregnancies of uncertain site. *Human Reproduction*, 10:1223–1227, 1995.

[175] Z. Haider, G. Condous, A. Khalid, E. Kirk, F. Mukri, B. Van Calster, D. Timmerman, and T. Bourne. The impact of the availability of ultrasonography in the acute gynaecology unit. *Ultrasound in Obstetrics and Gynecology*, 28:207–213, 2006.

[176] Z. Haider, G. Condous, E. Kirk, A. Syed, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Development of logistic regression models to predict chlamydial infection in women attending an acute gynaecology unit (AGU). *BJOG - An International Journal of Obstetrics and Gynaecology*, 112:1451–1452, 2005. Abstract presented at BBRS2005, Liverpool (UK).

[177] Z. Haider, E. Kirk, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Transvaginal ultrasonography and saline infusion sonography compared with outpatient hysteroscopy for the diagnosis of endometrial polyps and submucous fibroids. *Ultrasound in Obstetrics and Gynecology*, 28:607, 2006. Abstract presented at ISUOG2006, London (UK).

[178] D.J. Hand. Deconstructing statistical questions (with discussion). *Journal of the Royal Statistical Society Series A*, 157:317–356, 1994.

[179] D.J. Hand. Measuring diagnostic accuracy of statistical prediction rules. *Statistica Neerlandica*, 55:3–16, 2001.

[180] D.J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21:1–15, 2006.

[181] D.J. Hand and R.J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45:171–186, 2001.

[182] J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[183] J.A. Hanley and B.J. McNeil. A method of comparing the area under two ROC curves derived from the same cases. *Radiology*, 148:839–843, 1983.

[184] F.E. Harrell Jr, K.L. Lee, and D.B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

[185] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *Annals of Statistics*, 26:451–471, 1998.

[186] T. Hastie, R. Tibshirani, and J. Friedman. *Data mining, inference, and prediction*. Springer, New York, 2001.

[187] T.J. Hastie and R.J. Tibshirani. *Generalized additive models*. Chapman & Hall/CRC, Boca Raton, 1990.

[188] C. Heim, P.M. Plotsky, and C.B. Nemeroff. Importance of studying the contributions of early adverse experience to neurobiological findings in depression. *Neuropsychopharmacology*, 29:641–648, 2004.

[189] A.P.M. Heintz, F. Odicino, P. Maisonneuve, U. Beller, J.L. Benedet, W.T. Creasman, H.Y. Ngan, and S. Pecorelli. Carcinoma of the ovary. *International Journal of Gynecology and Obstetrics*, 83, Suppl. 1:135–166, 2003.

[190] T.H. Hollingsworth. Using an electronic computer in a problem of medical diagnosis. *Journal of the Royal Statistical Society Series A*, 122:221–231, 1959.

[191] V. Honavar. Artificial intelligence: an overview. Unpublished manuscript, Iowa State University, USA, 2006. Available at http://www.cs.iastate.edu/∼cs572/handout1.pdf (accessed on Sep. 7, 2007).

[192] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[193] D.W. Hosmer and S. Lemeshow. *Applied logistic regression*. Wiley, New York, 2nd edition, 2000.

[194] H.J. Huang, P. Neven, M. Drijkoningen, R. Paridaens, H. Wildiers, E. Van Limbergen, P. Berteloot, F. Amant, M.-R. Christiaens, and I. Vergote. Association between HER-2/neu and the progesterone receptor in oestrogen-dependent breast cancer is age-related. *Breast Cancer Research and Treatment*, 91:81–87, 2005.

[195] J. Huang and C.X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17:299–310, 2005.

[196] T.-K. Huang, R.C. Weng, and C.-J. Lin. Generalized Bradley-Terry models and multi-class probability estimates. *Journal of Machine Learning Research*, 7:85–115, 2006.

[197] D.L. Hunt, R.B. Haynes, S.E. Hanna, and K. Smith. Effects of computer-based clinical decision support systems on physician performance and patient outcomes. *Journal of the American Medical Association*, 280:1339–1346, 1998.

[198] C.M. Hurvich and C.-L. Tsai. The impact of model selection on inference in linear regression. *The American Statistician*, 44:214–217, 1990.

[199] D. Husmeier, W. Penny, and S.J. Roberts. An empirical evaluation of Bayesian sampling with hybrid Monte Carlo for training neural network classifiers. *Neural Networks*, 12:677–705, 1999.

[200] J.P.A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2:696–701, 2005.

[201] L.M. Irvine and M.E. Setchell. Serum progesterone testing to predict ectopic pregnancy in symptomatic first-trimester patients. *Human Reproduction*, 16:2230–2234, 2001.

[202] T. Jaakkola and M.I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

[203] I. Jacobs, D. Oram, J. Fairbanks, J. Turner, C. Frost, and J. Grudzinskas. A risk of malignancy index incorporating ca 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *British Journal of Obstetrics and Gynaecology*, 97:922–929, 1990.

[204] I.J. Jacobs, S. Skates, A.P. Davies, R.P. Woolas, A. Jeyerajah, P. Weidemann, K. Sibley, and D.H. Oram. Risk of diagnosis of ovarian cancer after raised serum CA 125 concentration: a prospective cohort study. *British Medical Journal*, 313:1355–1358, 1996.

[205] R. Jaeschke, G. Guyatt, and D.L. Sackett. Users' guides to the medical literature. III. how to use an article about a diagnostic test. A. are the results of the study valid? evidence-based medicine working group. *Journal of the American Medical Association*, 271:389–391, 1994.

[206] R. Jaeschke, G.H. Guyatt, and D.L. Sackett. Users' guides to the medical literature. III. how to use an article about a diagnostic test. B. what are the results and will they help me in caring for my patients? evidence-based medicine working group. *Journal of the American Medical Association*, 271:703–707, 1994.

[207] A. Jain, R.P. Dubes, and C. Chen. Bootstrap techniques for error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:628–633, 1987.

[208] D. Jameson. *Mind-body health and stress tolerance*. iUniverse, Lincoln, 2003.

[209] B.J. Jefferis, C. Power, and C. Hertzman. Birth weight, childhood socioeconomic environment, and cognitive development in the 1958 British birth cohort study. *British Medical Journal*, 325:305, 2002.

[210] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, and M.J. Thun. Cancer statistics, 2007. *CA - A Cancer Journal for Clinicians*, 57:43–66, 2007.

[211] N. Job-Spira, H. Fernandez, J. Bouyer, J.-L. Pouly, E. Germain, and J. Coste J. Ruptured tubal ectopic pregnancy: risk factors and reproductive outcome: results of a population-based study in France. *American Journal of Obstetrics and Gynecology*, 180:938–944, 1999.

[212] L. Jokubkiene, P. Sladkevicius, and L. Valentin. Does three-dimensional power Doppler ultrasound help in discrimination between benign and malignant ovarian masses? *Ultrasound in Obstetrics and Gynecology*, 29:215–225, 2007.

[213] A.C. Justice, K.E. Covinsky, and J.A. Berlin. Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, 130:515–524, 1999.

[214] N. Kadar, B.V. Caldwell, and R. Romero. A method of screening for ectopic pregnancy and its indications. *Obstetrics & Gynecology*, 58:162–166, 1981.

[215] N. Kadar, G. DeVore, and R. Romero. Discriminatory hCG zone: its use in the sonographic evaluation for ectopic pregnancy. *Obstetrics & Gynecology*, 58:156–161, 1981.

[216] E. Kajantie. Fetal origins of stress-related adult disease. *Annals of the New York Academy of Sciences*, 1083:11–27, 2006.

[217] J.D. Kalbfleisch and R.L. Prentice. *The statistical analysis of failure time data*. Wiley, New York, 2nd edition, 2002.

[218] F. Kamwendo, L. Forslin, L. Bodin, and D. Danielsson. Epidemiology of ectopic pregnancy during a 28 year period and the role of pelvic inflammatory disease. *Sexually transmitted infections*, 76:28–32, 2000.

[219] B.C. Kaplan, R.G. Dart, M. Moskos, E. Kuligowska, B. Chun, M. Adel Hamid, K. Northern, J. Schmidt, and A. Kharwadkar. Ectopic pregnancy: prospective study with improved diagnostic accuracy. *Annals of Emergency Medicine*, 28:10–17, 1996.

[220] P. Karsmakers, K. Pelckmans, and J.A.K. Suykens. Multi-class kernel logistic regression: a fixed size implementation. In *Proceedings of the 20th International Joint Conference on Neural Networks (IJCNN2007)*, pages 1756–1761, Orlando (USA), 2007.

[221] S.S. Keerthi, K. Duan, S.K. Shevade, and A.N. Poo. A fast dual algorithm for kernel logistic regression. In *Proceedings of the 19th International Conference on Machine Learning (ICML2002)*, pages 299–306, Sydney (Australia), 2002.

[222] S. Kehoe, J. Powell, S. Wilson, and C. Woodman. The influence of the operating surgeon's specialisation on patient survival in ovarian carcinoma. *British Journal of Cancer*, 70:1014–1017, 1994.

[223] K.S. Khan, P.F.W. Chien, and L.S. Dwarakanath. Logistic regression models in obstetrics and gynecology literature. *Obstetrics and Gynecology*, 93:1014–1020, 1999.

[224] K.S. Khan, D. Wojdyla, L. Say, A.M. Glmezoglu, and P.F.A. Van Look. WHO analysis of causes of maternal death: a systematic review. *Lancet*, 367:1066–1074, 2006.

[225] E. Kirk, P. Alnaes-Katjavivi, G. Condous, B. Van Calster, S. Van Huffel, O. Istre, D. Timmerman, and T. Bourne. Validation of a logistic regression model to predict the outcome of pregnancies of unknown location in a new patient population. *Ultrasound in Obstetrics and Gynecology*, 28:365, 2006. Abstract presented at ISUOG2006, London (UK).

[226] E. Kirk, S. Bora, B. Van Calster, G. Condous, S. Van Huffel, D. Timmerman, and T. Bourne. Chlamydia trachomatis infection in patients attending an early pregnancy unit: prevalence, symptoms, pregnancy location and viability. *Acta Obstetricia et Gynecologica Scandinavica*, submitted.

[227] E. Kirk and T. Bourne. The nonsurgical management of ectopic pregnancy. *Current Opinion in Obstetrics and Gynecology*, 18:587–593, 2006.

[228] E. Kirk, G. Condous, C. Bottomley, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Using a mathematical model in the management of pregnancies of unknown location: a prospective interventional study. *Ultrasound in Obstetrics and Gynecology*, 28:364, 2006. Abstract presented at ISUOG2006, London (UK).

[229] E. Kirk, G. Condous, C. Bottomley, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. What is the optimal approach to classifying viable intra-uterine pregnancies in the PUL population? *Ultrasound in Obstetrics and Gynecology*, 28:364, 2006. Abstract presented at ISUOG2006, London (UK).

[230] E. Kirk, G. Condous, C. Bottomley, B. Van Calster, D. Timmerman, S. Van Huffel, and T. Bourne. Ultrasound and not biochemistry determines the ultimate viability of intra-uterine pregnancies in the PUL population. *Ultrasound in Obstetrics and Gynecology*, 28:364, 2006. Abstract presented at ISUOG2006, London (UK).

[231] E. Kirk, G. Condous, and T. Bourne. Ectopic pregnancy deaths: What should we be doing? *Hospital Medicine*, 65:657–660, 2004.

[232] E. Kirk, G. Condous, and T. Bourne. The non-surgical management of ectopic pregnancy. *Ultrasound in Obstetrics and Gynecology*, 27:91–100, 2006.

[233] E. Kirk, G. Condous, B. Van Calster, Z. Haider, and T. Bourne. Is it possible to predict successful treatment response to methotrexate earlier than seven days? *Ultrasound in Obstetrics and Gynecology*, 26:456, 2005. Abstract presented at ISUOG2005, Vancouver (Canada).

[234] E. Kirk, G. Condous, B. Van Calster, Z. Haider, S. Van Huffel, D. Timmerman, and T. Bourne. A validation of the most commonly used protocol to predict the success of single dose methotrexate in the treatment of ectopic pregnancy. *Human Reproduction*, 22:858–863, 2007.

[235] E. Kirk, G. Condous, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Rationalizing the follow-up of pregnancies of unknown location. Talk given at the 31st British International Conference on Obstetrics and Gynaecology, London (UK), Jul. 2007.

[236] E. Kirk, G. Condous, B. Van Calster, S. Van Huffel, D. Timmerman, and T. Bourne. Rationalizing the follow-up of pregnancies of unknown location. *Human Reproduction*, 22:1744–1750, 2007.

[237] E. Kirk, G. Condous, Z. Haider, B. Van Calster, D. Timmerman, and T. Bourne. Is the ultrasonographic appearance of an ectopic pregnancy a predictor of success for conservative management strategies? *BJOG - An International Journal of Obstetrics and Gynaecology*, 113:864, 2006. Abstract presented at BBRS Competition Meeting 2005, London (UK).

[238] E. Kirk, G. Condous, Z. Haider, T. Chang, B. Van Calster, and T. Bourne. Can the ultrasonographic appearance of an ectopic pregnancy determine the likelihood of success for the conservative management? *Ultrasound in Obstetrics and Gynecology*, 26:361–362, 2005. Abstract presented at ISUOG2005, Vancouver (Canada).

[239] E. Kirk, L. Tan, G. Condous, B. Van Calster, C. Bottomley, S. Van Huffel, D. Timmerman, and T. Bourne. Selecting failing pregnancies in the pregnancy of unknown location (PUL) population to reduce follow-up. *BJOG - An International Journal of Obstetrics and Gynaecology*, 114:1037, 2007. Abstract presented at BBRS2006, Leicester (UK).

[240] E. Kirk, L. Tan, G. Condous, B. Van Calster, N. Rosello, D. Gould, P. Alnaes-Katjavivi, O. Istre, S. Van Huffel, D. Timmerman, and T. Bourne. External validation of a mathematical model to predict the outcome of pregnancies of unknown location (PULs). manuscript in preparation.

[241] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1145. San Francisco, Morgan Kaufmann, 1995.

[242] R. Kohavi and G.H. John. Wrappers for feature selection. *Artificial Intelligence*, 97:273–324, 1997.

[243] R. Kohavi and G.H. John. The wrapper approach. In H. Liu and H. Motoda, editors, *Feature selection for knowledge discovery and data mining*, pages 33–50. Kluwer Academic Publishers, Norwell, 1998.

[244] R. Kohavi and F. Provost. Glossary of terms. *Machine Learning*, 30:271–274, 1998.

[245] M.A. Kohn, K. Kerr, D. Malkevich, N. O'Neil, M.J. Kerr, and B.C. Kaplan. Beta-human chorionic gonadotropin levels and the likelihood of ectopic pregnancy in emergency department patients with abdominal pain or vaginal bleeding. *Academic Emergency Medicine*, 10:119–126, 2003.

[246] E.B. Kong and T.G. Dietterich. Probability estimation via error-correcting output coding. In M.H. Hamza, editor, *Proceedings of the IASTED International Conference on Artificial Intelligence and Soft Computing (ASC1997)*, Calgary, 1997. ACTA Press.

[247] L.I. Kuncheva. *Combining pattern classifiers: methods and algorithms*. Wiley, New York, 2004.

[248] P. Lachenbruch. An almost unbiased method for the probability of misclassification in discriminant analysis. *Biometrics*, 23:639–645, 1967.

[249] G. Landström, J. Thorburn, and I. Bryman. Treatmrent failures and complications of ectopic pregnancy: changes over a 20 year period. *Human Reproduction*, 13:203–207, 1998.

[250] T.A. Lasko, J.G. Bhagwat, K.H. Zou, and L. Ohno-Machado. The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, 38:404–415, 2005.

[251] G. Latchaw, P. Takacs, L. Gaitan, S. Geren, and J. Burzawa. Risk factors associated with the rupture of tubal ectopic pregnancy. *Gynecologic and Obstetric Investigation*, 60:177–180, 2005.

[252] S. Le Cessie and J.C. Van Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society C*, 41:191–201, 1992.

[253] D.B. Leake. Artificial intelligence. In G. Considine and P. Kulik, editors, *Van Nostrand scientific encyclopedia*. Wiley, New York, 9th edition, 2002.

[254] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99:67–81, 2004.

[255] S. Lemeshow and D.W. Hosmer. The use of goodness-of-fit statistics in the development of logistic regression models. *American Journal of Epidemiology*, 115:92–106, 1982.

[256] J.P. Lerner, I.E. Timor-Trisch, A. Federman, and G. Abramovich. Transvaginal ultrasonographic characterization of ovarian masses with an improved, weighted scoring syste. *American Journal of Obstetrics and Gynecology*, 170:81–85, 1994.

[257] G. Lewis and J. Drife, editors. *Why mothers die 2000-2002. The sixth report of the confidential inquiries into maternal deaths in the United Kingdom*. RCOG Press, London, 2004.

[258] R.J. Lewis and R.L. Wears. An introduction to the Bayesian analysis of clinical trials. *Annals of Emergency Medicine*, 22:1328–1336, 1993.

[259] R.J. Lilford and D. Braunholtz. For debate: the statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal*, 313:603–607, 1996.

[260] H.-T. Lin, C.-J. C.-J. Lin, and R.C. Weng. A note on Platt's probabilistic outputs for support vector machines. Technical report, National Taiwan University, Taiwan, 2003.

[261] G.H. Lipscomb, K.J. Puckett, D. Bran, and F.W. Ling. Management of separation pain after single-dose methotrexate therapy for ectopic pregnancy. *Obstetrics & Gynecology*, 93:590–593, 1999.

[262] G.H. Lipscomb, T.G. Stovall, and F.W. Ling. Nonsurgical treatment of ectopic pregnancy. *New England Journal of Medicine*, 343:1325–1329, 2000.

[263] P.J. Lisboa and A.F.G. Taktak. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Networks*, 19:408–415, 2006.

[264] P.J.G. Lisboa, A. Vellido, and H. Wong. Bias reduction in skewed binary classification with Bayesian neural networks. *Neural Networks*, 13:407–410, 2000.

[265] J.R. Little and D. Rubin. *Statistical analysis with missing data*. Wiley, New York, 2nd edition, 2002.

[266] C. Lu. *Probabilistic machine learning approaches to medical classification problems*. PhD thesis, Department of Electrical Engineering (ESAT-SISTA), Katholieke Universiteit Leuven, Belgium, January 2005.

[267] C. Lu, T. Van Gestel, J.A.K. Suykens, S. Van Huffel, I. Vergote, and D. Timmerman. Preoperative prediction of malignancy of ovarian tumours using least squares support vector machines. *Artificial Intelligence in Medicine*, 28:281–306, 2003.

[268] C. Lu, J.A.K. Suykens, D. Timmerman, I. Vergote, and S. Van Huffel. Linear and nonlinear preoperative classification of ovarian tumours. In T. Ichimura and K. Yoshida, editors, *Knowledge-based intelligent systems for health care*, pages 343–382. Advanced Knowledge International, Magill, 2004.

[269] D.J.C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4:698–714, 1992.

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

[271] D.P. MacKinnon, C.M. Lockwood, and J. Williams. Confidence limits for the indirect effect: distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39:99–128, 2004.

[272] J.I. Makinen, R.U. Erkkola, and P.J. Laippala. Causes of the increase in the incidence of ectopic pregnancy. A study on 1017 patients from 1966 to 1985 in Turku, Finland. *American Journal of Obstetrics and Gynecology*, 160:642–646, 1989.

[273] G. Marion and D. Saad. A statistical mechanical analysis of a Bayesian inference scheme for an unrealizable rule. *Journal of Physics A: Mathematical and General*, 28:2159–2171, 1995.

[274] D.K. McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9:190–195, 1989.

[275] M.L. McCord, D. Muram, J.E. Buster, K.L Arheart, T.G. Stovall, and S.A. Carson. Single serum progesterone as a screen for ectopic pregnancy: exchanging specificity and sensitivity to obtain optimal test performance. *Fertility and Sterility*, 66:513–516, 1996.

[276] S. Menard. Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54:17–24, 2000.

[277] H.L. Mertz and T.M. Yalcinkaya. Early diagnosis of ectopic pregnancy. Does use of a strict algorithm decrease the incidence of tubal rupture? *The Journal of Reproductive Medicine*, 46:29–33, 2001.

[278] C.E. Metz, B.A. Herman, and J.-H. Shen. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, 17:1033–1053, 1998.

[279] D. Michie, D.J. Spiegelhalter, and C.C. Taylor. Introduction. In D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors, *Machine learning, neural and statistical classification*, pages 1–5. Ellis Horwood, New York, 1994.

[280] Y. Miki, J. Swensen, D. Shattuck-Eidens, P.A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L.M. Bennett, W. Ding, et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266:66–71, 1994.

[281] R.A. Miller. Medical diagnostic decision support systems - past, present, and future. *Journal of the American Medical Informatics Association*, 1:8–27, 1994.

[282] D. Minaretzis, C. Tsionou, D. Tziortziotis, S. Michalas, and D. Aravantinos. Ovarian tumours: prediction of the probability of malignancy by using patient's age and tumour morphologic features with a logistic model. *Gynecologic and Obstetric Investigation*, 38:140–144, 1994.

[283] T. Mitchell. Does machine learning really work? *AI Magazine*, 18:11–20, 1997.

[284] T. Mitchell. The discipline of machine learning. Technical report CMU-ML-06-108, Carnegie Mellon University, USA, 2006.

[285] M. Mittlböck and M. Schemper. Explained variation for logistic regression. *Statistics in Medicine*, 15:1987–1997, 1996.

[286] B.W. Mol, J.G. Lijmer, W.M. Ankum, F. van der Veen, and P.M. Bossuyt. The accuracy of single serum progesterone measurement in the diagnosis of ectopic pregnancy: a meta-analysis. *Human Reproduction*, 13:3220–3227, 1998.

[287] B.W.J. Mol, D. Boll, M. De Kanter, A.P.M. Heintz, E.A. Sijmons, S.G. Oei, H. Bal, and H.A. Brölmann. Distinguishing the benign and malignant adnexal mass: an external validation of prognostic models. *Gynecologic Oncology*, 80:162–167, 2001.

[288] A.M. Molinaro, R. Simon, and R.M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21:3301–3307, 2005.

[289] L. Morales, S. Pans, K. Verschueren, B. Van Calster, R. Paridaens, R. Westhovens, D. Timmerman, L. De Smet, I. Vergote, M.-R. Christiaens, and P. Neven. A prospective study to assess the short term intra-articular and tenosynovial changes in the aromatase inhibitor-associated arthralgia syndrome. *Journal of Clinical Oncology*, in press.

[290] D. Mossman. Three-way ROCs. *Medical Decision Making*, 19:78–89, 1999.

[291] G. Moulin-Romsee, E. D'Hondt, F. Persyn, B. Van Calster, S. Ceyssens, P. Neven, and S. Stroobants. Detection of sentinel lymph nodes in breast cancer: everyday practice and pitfalls. volume 34s2, page S387, 2007. Abstract presented at EANM2007, Copenhagen (Denmark).

[292] I.T. Nabney. *Netlab: algorithms for pattern recognition*. Springer, London, 2002.

[293] I.T. Nabney, D.J. Evans, Y. Brulé, and C. Gordon. Assessing the effectiveness of Bayesian feature selection. In D. Husmeier, R. Dybowski, and S. Roberts, editors, *Probabilistic modelling in medical informatics and bioinformatics*, pages 371–390. Springer, London, 2005.

[294] C. Nadeau and Y. Bengio. Inference for the generalization error. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in neural information processing systems*, volume 12, pages 307–313. Cambridge, MIT Press, 2000.

[295] N.J.D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692, 1991.

[296] F.A. Nash. Differential diagnosis: an apparatus to assist the logical faculties. *Lancet*, 1:874, 1954.

[297] R.M. Neal. *Bayesian learning for neural networks*. Springer, New York, 1996.

[298] J. Neter, M.H. Kutner, C.J. Nachtsheim CJ, and W. Wasserman. *Applied linear statistical models*. McGraw-Hill, Boston, 4th edition, 1996.

[299] P. Neven, B. Van Calster, I. Van den Bempt, S. Van Huffel, V. Van Belle, W. Hendrickx, J. Decock, H. Wildiers, R. Paridaens, F. Amant, K. Leunen, P. Berteloot, D. Timmerman, E. Van Limbergen, C. Weltens, W. Van den Bogaert, A. Smeets, I. Vergote, M.-R. Christiaens, and M. Drijkoningen. Age interacts with the expression of steroid and HER-2 receptors in operable invasive breast cancer. *Breast Cancer Research and Treatment*, in press.

[300] P. Neven, I. Van den Bempt, B. Van Calster, N. Pochet, J. Cheng, S. Van Huffel, W. Hendrickx, J. Decock, K. Leunen, F. Amant, P. Berteloot, H. Wildiers, R. Paridaens, E. Van Limbergen, C. Weltens, D. Timmerman, T. Van Gorp, A. Smeets, W. Van den Bogaert, I. Vergote, M.-R. Christiaens, and M. Drijkoningen. Age-related prevalence of breast cancers by ER, PR and HER-2. Poster presented at the 10th Biennial Conference on Primary Therapy of Early Breast Cancer, St Gallen (Switzerland), Mar. 2007.

[301] R.G. Newcombe. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17:2635–2650, 1998.

[302] R.G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17:857–872, 1998.

[303] R.G. Newcombe. A deficiency of the odds ratio as a measure of effect size. *Statistics in Medicine*, 25:4235–4240, 2006.

[304] A.Y. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *Proceedings of the 21st International Conference on Machine Learning (ICML2004)*. Madison, Omnipress, 2004.

[305] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML2005)*. Madison, Omnipress, 2005.

[306] S.A. Obed. Diagnosis of unruptured ectopic pregnancy is still uncommon in Ghana. *Ghana Medical Journal*, 40:3–7, 2006.

[307] N.A. Obuchowski and M.L. Lieber. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Academic Radiology*, 5:561–571, 1998.

[308] T.G. O'Connor, Y. Ben-Shlomo, J. Heron, H. Golding, D. Adams, and V. Glover. Prenatal anxiety predicts individual differences in cortisol in pre-adolescent children. *Biological Psychiatry*, 58:211–217, 2005.

[309] Office for National Statistics. Cancer registration statistics 2004 (series mb1 no. 35). London, Office for National Statistics, Crown copyright, 2006.

[310] F. Ojeda, J.A.K. Suykens, and B. De Moor. Variable selection by rank-one updates for least squares support vector machines. In *Proceedings of the 20th International Joint Conference on Neural Networks (IJCNN2007)*, pages 2283–2288, Orlando (USA), 2007.

[311] S. Ong and M. Wingfield. Increasing incidence of ectopic pregnancy: is it iatrogenic? *Irish Medical Journal*, 92:364–365, 1999.

[312] D. Opitz and R. Maclin. Popular ensemble methods: an empirical study. *Machine Learning*, 11:169–198, 1999.

[313] S.J. Ory. New options for diagnosis and treatment of ectopic pregnancy. *Journal of the American Medical Association*, 267:534–537, 1992.

[314] C.K. Osborne and R. Schiff. Growth factor receptor cross-talk with estrogen receptor as a mechanism for tamoxifen resistance in breast cancer. *The Breast*, 12:362–367, 2003.

[315] M.Y. Park and T. Hastie. l_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society B*, 69:659–677, 2007.

[316] P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, and A.R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49:1373–1379, 1996.

[317] K. Pelckmans, I. Goethals, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Componentwise least squares support vector machines. In L. Wang, editor, *Support vector machines: theory and applications*, pages 77–98. Springer, Berline, 2005.

[318] K. Pelckmans, J.A.K. Suykens, T. Van Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor, and J. Vandewalle. LS-SVMlab: a Matlab/C toolbox for least squares support vector machines. Internal report 02-44, ESAT, Katholieke Universiteit Leuven, Belgium, 2002.

[319] W.D. Penny and S.J. Roberts. Bayesian neural networks for classification: how useful is the evidence framework? *Neural Networks*, 12:877–892, 1999.

[320] M.S. Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, New York, 2003.

[321] M.S. Pepe, G. Longton, G.L. Anderson, and M. Schummer. Selecting differentially expressed genes from microarray experiments. *Biometrics*, 59:133–142, 2003.

[322] T.V. Perneger. What's wrong with Bonferroni adjustments. *British Medical Journal*, 316:1236–1238, 1998.

[323] J. Pizarro, E. Guerrero, and P.L. Galindo. Multiple comparison procedures applied to model selection. *Neurocomputing*, 48:155–173, 2002.

[324] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A.J. Smola, P.L. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in large margin classifiers*, pages 61–74. MIT Press, Cambridge, 2000.

[325] J.C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin dags for multiclass classification. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in neural information processing systems*, volume 12, pages 547–553. Cambridge, MIT Press, 2000.

[326] N.L.M.M. Pochet and J.A.K. Suykens. Support vector machines versus logistic regression: improving prospective performance in clinical decision-making. *Ultrasound in Obstetrics and Gynecology*, 27:607–608, 2006.

[327] C. Power and B.J. Jefferis. Fetal environment and subsequent obesity: a study of maternal smoking. *International Journal of Epidemiology*, 31:413–419, 2002.

[328] D. Price, S. Knerr, L. Personnaz, and G. Dreyfus. Pairwise neural network classifiers with probabilistic outputs. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Advances in neural information processing systems*, volume 7, pages 1109–1116. Cambridge, MIT Press, 1995.

[329] H.J. Prömpeler, H. Madjar, W. Sauerbrei, U. Lattermann, and A. Pfleiderer. Diagnostic formula for the differentiation of adnexal tumours by transvaginal sonographyg. *Obstetrics & Gynecology*, 89:428–433, 1997.

[330] F. Provost and P. Domingos. Tree induction for probability-based ranking. *Machine Learning*, 52:199–215, 2003.

[331] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42:203–231, 2001.

[332] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th International Conference on Machine Learning (ICML2001)*, pages 445–453, Madison (USA), 1998.

[333] F. Provost and R. Kohavi. On applied research in machine learning. *Machine Learning*, 30:127–132, 1998.

[334] C.E. Rasmussen. Evaluation of Gaussian processes and other methods for non-linear regression. PhD thesis, University of Toronto, Canada, 1996.

[335] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for AdaBoost. *Machine Learning*, 42:287–320, 2001.

[336] P. Refregier and F. Vallet. Probabilistic approach for multiclass classification with neural networks. In *Proceedings of the International Conference on Artificial Networks*, pages 1003–1007, 1991.

[337] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

[338] V. Roth. Probabilistic discriminative kernel classifiers for multi-class problems. In B. Radig and S. Florczyk, editors, *Pattern recognition - DAGM'01*, volume 2191 of *Lecture Notes in Computer Science*, pages 246–253. Berlin, Springer, 2001.

[339] K.J. Rothman. No adjustments are needed for multiple comparisons. *Epidemiology*, 1:43–46, 1990.

[340] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[341] P.J. Rousseeuw and A. Christmann. Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, 43:315–332, 2003.

[342] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[343] D.L. Sackett, J.J. Deeks, and D.G. Altman. Down with odds ratios! *Evidence-Based Medicine*, 1:164–166, 1996.

[344] D.L. Sackett, S.E. Strauss, W.S. Richardson, W. Rosenberg, and R.B. Haynes. *Evidence-based medicine: how to practice and teach* EBM. Churchill Livingstone, Edinburgh, 2nd edition, 2000.

[345] L. Saitta and F. Neri. Learning in the "real world". *Machine Learning*, 30:133–163, 1998.

[346] D.S. Salsburg. The religion of statistics as practiced in medical journals. *The American Statistician*, 39:220–223, 1985.

[347] S.L. Salzberg. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317–328, 1997.

[348] A.M. Sassone, I.E. Timor-Tritsch, A. Artner, C. Westhoff, and W.B. Warren. Transvaginal sonographic characterization of ovarian disease: evaluation of a new scoring system to predict ovarian malignancyh. *Obstetrics & Gynecology*, 78:70–76, 1991.

[349] L. Savage and A. Widener. Memo to the media: ultrasound may better classify ovarian tumors. *Journal of the National Cancer Institute*, 99:1653, 2007.

[350] D.J. Saville. Multiple comparison procedures: the practical solution. *The American Statistician*, 44:174–180, 1990.

[351] J.L. Schafer. *Analysis of incomplete multivariate data*. Chapman & Hall, London, 1997.

[352] J.L. Schafer and M.K. Olsen. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, 35:545–571, 1998.

[353] R.E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.

[354] T. Schneider. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14:853–871, 2001.

[355] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[356] B.E. Seeber and K.T. Barnhart. Suspected ectopic pregnancy. *Obstetrics & Gynecology*, 107:399–413, 2006.

[357] M. Sewell. Ensemble methods. Unpublished manuscript, University College London, UK, 2007. Available at http://www.cs.ucl.ac.uk/staff/M.Sewell/ (accessed on August 31, 2007).

[358] E.S. Shtatland, K. Kleinman, and E.M. Cain. Stepwise methods in using SAS PROC LOGISTIC and SAS ENTERPRISE MINER for prediction. In *Proceedings of the 28th Conference of the SAS Users Group International (SUGI28)*, Seattle (USA), 2003.

[359] S. Siegel and N.J. Castellan Jr. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York, 2nd edition, 1988.

[360] I. Sim, P. Gorman, R.A. Greenes, R.B. Haynes, B. Kaplan, H. Lehmann, and P.C. Tang. Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association*, 8:527–534, 2001.

[361] R.L. Somorjai, B. Dolenko, and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, 19:1484–1491, 2003.

[362] D.J. Spiegelhalter, J.P. Myles, D.R. Jones, and K.R. Abrams. An introduction to Bayesian methods in health technology assessment. *British Medical Journal*, 319:508–512, 1999.

[363] V. Stalbovskaya, E.C. Ifeachor, S. Van Huffel, and D. Timmerman. A new method for modeling preoperative diagnosis of ovarian tumors. *IEEE Transactions on Biomedical Engineering*, 54:2064–2072, 2007.

[364] J.A. Sterne and G. Davey Smith. Sifting the evidence - what's wrong with significance tests. *British Medical Journal*, 322:226–231, 2001.

[365] E.W. Steyerberg, M.J.C. Eijkmans, and J.D.F. Habbema. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, 52:935–942, 1999.

[366] E.W. Steyerberg, M.J.C. Eijkmans, and J.D.F. Habbema. Application of shrinkage techniques in logistic regression analysis: a case study. *Statistica Neerlandica*, 55:76–88, 2001.

[367] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36:111–147, 1974.

[368] M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64:29–35, 1977.

[369] T.G. Stovall, F.W. Ling, L.A. Gray, S.A. Carson, and J.E. Buster. Methotrexate treatment of unruptured ectopic pregnancy: a report of 100 cases. *Obstetrics & Gynecology*, 77:749–753, 1991.

[370] Y. Sun, S. Todorovic, and J. Li. Unifying multi-class AdaBoost algorithms with binary base learners under the margin framework. *Pattern Recognition Letters*, 28:631–643, 2007.

[371] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least squares support vector machines*. World Scientific, Singapore, 2002.

[372] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. Support vector machines: least squares approaches and extenstions. In J. Suykens, G. Horváth, S. Basu, C. Micchelli, and J. Vandewalle, editors, *Advances in learning theory: methods, models and applications*, pages 155–177. IOS Press, Amsterdam, 2003.

[373] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.

[374] J.A.K. Suykens and J. Vandewalle. Multiclass least squares support vector machine classifiers. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN1999)*, Washington DC (USA), 1999.

[375] J.A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1988.

[376] A. Tailor, D. Jurkovic, T. Bourne, W.P. Collins, and S. Campbell. Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis. *Ultrasound in Obstetrics and Gynecology*, 10:41–47, 1997.

[377] A. Tailor, D. Jurkovic, T.H. Bourne, W.P. Collins, and S. Campbell. Sonographic prediction of malignancy in adnexal masses using an artificial neural network. *British Journal of Obstetrics and Gynaecology*, 106:21–30, 1999.

[378] A.F.G. Taktak, A. Eleuteri, S.P. Lake, and A.C. Fisher. Evaluation of prognostic models: discrimination and calibration performance. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, Plymouth (UK), 2007.

[379] T.Z. Tan, C. Quek, and G.S. Ng. Ovarian cancer diagnosis by hippocampus and neocortex-inspired learning memory structures. *Neural Networks*, 18:818–825, 2005.

[380] D.M.J. Tax and R.P.W. Duin. Using two-class classifiers for multiclass classification. In *Proceedings of the sixteenth International Conference on Pattern Recognition*, pages 124–127. Los Alamitos, IEEE Computer Society Press, 2002.

[381] H.H. Thodberg. A review of Bayesian neural networks with an application to near infrared spectroscopy. *IEEE Transactions on Neural Networks*, 7:56–72, 1996.

[382] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.

[383] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistical Society B*, 63:411–423, 2001.

[384] D. Timmerman. Editorial: lack of standardization in gynaecological ultrasonography. *Ultrasound in Obstetrics and Gynecology*, 16:395–398, 2000.

[385] D. Timmerman. The use of mathematical models to evaluate pelvic masses; can they beat an expert operator? *Best Practice & Research Clinical Obstetrics & Gynaecology*, 18:91–104, 2004.

[386] D. Timmerman, T.H. Bourne, A. Tailor, W.P. Collins, H. Verrelst, K. Vandenberghe, and I. Vergote. A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: the development of a new logistic regression model. *American Journal of Obstetrics and Gynecology*, 181:57–65, 1999.

[387] D. Timmerman, B. Van Calster, T. Bourne, A. Testa, C. Van Holsbeke, E. Domali, D. Jurkovic, P. Neven, S. Van Huffel, and L. Valentin. Subjective assessment based on ultrasonography and the use of serum CA-125 to predict malignancy in ovarian tumors. *Journal of Clinical Oncology*, 25:4194–4200, 2007.

[388] D. Timmerman, B. Van Calster, D. Jurkovic, L. Valentin, A. Testa, J.P. Bernard, C. Van Holsbeke, S. Van Huffel, I. Vergote, and T. Bourne. Shortcomings of serum CA125 in diagnosing ovarian cancer. *Ultrasound in Obstetrics and Gynecology*, 28:408, 2006. Abstract presented at ISUOG2006, London (UK).

[389] D. Timmerman, B. Van Calster, D. Jurkovic, L. Valentin, A.C. Testa, J.P. Bernard, C. Van Holsbeke, S. Van Huffel, I. Vergote, and T. Bourne. Is serum CA125 essential to distinguish between malignant and benign adnexal tumours? a multicentre study by the International Ovarian Tumour Analysis (IOTA) group. *International Journal of Gynecological Cancer*, 15s2:178–179, 2005. Abstract presented at ESGO2005, Istanbul (Turkey).

[390] D. Timmerman, E. Domali, B. Van Calster, A. Testa, C. Van Holsbeke, M. Konstantinovic, S. Van Huffel, and L. Valentin. Prospective internal evaluation of IOTA logistic regression model. *Ultrasound in Obstetrics and Gynecology*, 28:506, 2006. Abstract presented at ISUOG2006, London (UK).

[391] D. Timmerman, P. Schwärzler, W.P. Collins, F. Claerhout, M. Coenen, F. Amant, I. Vergote, and T.H. Bourne. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. *Ultrasound in Obstetrics and Gynecology*, 13:11–16, 1999.

[392] D. Timmerman, A. Testa, T. Bourne, E. Ferrazzi, L. Ameye, C. Van Holsbeke, B. Van Calster, W. Collins, I. Vergote, S. Van Huffel, and L. Valentin. Mathematical models to distinguish between malignant and benign adnexal masses: results from a multicenter study bt the International Ovarian Tumor Analysis (IOTA) group. Talk given at ASUM2006, Melbourne (Australia).

[393] D. Timmerman, A.C. Testa, T. Bourne, L. Ameye, D. Jurkovic, C. Van Holsbeke, D. Paladini, B. Van Calster, I. Vergote, S. Van Huffel, and L. Valentin. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound in Obstetrics and Gynecology*, in press.

[394] D. Timmerman, A.C. Testa, T. Bourne, E. Ferrazzi, L. Ameye, M.L. Konstantinovic, B. Van Calster, W.P. Collins, I. Vergote, S. Van Huffel, and L. Valentin. 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. *Journal of Clinical Oncology*, 23:8794–8801, 2005.

[395] D. Timmerman, L. Valentin, T.H. Bourne, W.P. Collins, H. Verrelst, and I. Vergote. Terms, definitions and measurements to describe the sonographic features of adnexal tumours: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) group. *Ultrasound in Obstetrics and Gynecology*, 16:500–505, 2000.

[396] D. Timmerman, H. Verrelst, T.H. Bourne, B. De Moor, W.P. Collins, and I. Vergote. Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. *Ultrasound in Obstetrics and Gynecology*, 13:17–25, 1999.

[397] J.F. Timms, E. Arslan-Low, A. Gentry-Maharaj, Z. Luo, D. T'Jampens, V.N. Podust, J. Ford, E.T. Fung, A. Gammerman, I. Jacobs, and U. Menon. Preanalytic influence of sample handling on SELDI-TOF serum protein profiles. *Clinical Chemistry*, 53:645–656, 2007.

[398] I.E. Timor-Tritsch, M.N. Yeh, D.B. Peisner, K.B. Lesser, and B.S. Slavik. The use of transvaginal ultrasound in the diagnosis of ectopic pregnancy. *American Journal of Obstetrics and Gynecology*, 161:157–161, 1989.

[399] S. Tingulstad, B. Hagen, F.E. Skjeldestad, M. Onsrud, T. Kiserud, and T. Halvorsen. Evaluation of a risk of malignancy index based on serum ca 125, ultrasound findings and menopausal status in the preoperative diagnosis of pelvic masses. *British Journal of Obstetrics and Gynaecology*, 103:826–831, 1996.

[400] M.E. Tipping. The relevance vector machine. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in neural information processing systems*, volume 12, pages 652–658. Cambridge, MIT Press, 2000.

[401] M.E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[402] T. Tongsong and S. Pongsatha. Transvaginal sonographic features in diagnosis of ectopic pregnancy. *International Journal of Gynaecology and Obstetrics*, 43:277–283, 1993.

[403] D. Trafimow. Hypothesis testing and theory evaluation at the boundaries: suprising insights from Bayes's theorem. *Psychological Review*, 110:526–535, 2003.

[404] M.K. Tuxen. Tumour marker CA125 in ovarian cancer. *Journal of Tumor Marker Oncology*, 16:49–68, 2001.

[405] D.M. Twickler, T.B. Forte, R. Santos-Ramos, D. McIntire, P. Harris, and D.S. Miller. The ovarian tumour index predicts risk for malignancy. *Cancer*, 86:2280–2290, 1999.

[406] D.G. Tzikas, L. Wei, A. Likas, Y. Yang, and N.P. Galatsanos. A tutorial on relevance vector machines for regression and classification with applications. *EURASIP News Letter*, 17:4–23, 2006.

[407] K. Vach, W. Sauerbrei, and M. Schumacher. Variable selection and shrinkage: comparison of some approaches. *Statistica Neerlandica*, 55:53–75, 2001.

[408] L. Valentin, B. Hagen, S. Tingulstad, and S. Eik-Nes. Comparison of 'pattern recognition' and logistic regression models for discrimination between benign and malignant pelvic masses: a prospective cross validation. *Ultrasound in Obstetrics and Gynecology*, 18:357–365, 2001.

[409] L. Valentin, D. Jurkovic, B. Van Calster, A. Testa, C. Van Holsbeke, T. Bourne, I. Vergote, S. Van Huffel, and D. Timmerman. CA-125 does not add to and is not superior to pattern recognition for discrimination between benign and malignant adnexal masses. Abstract presented at ISUOG2007, Florence (Italy), 2007.

[410] L. Valentin, D. Jurkovic, B. Van Calster, A. Testa, C. Van Holsbeke, T. Bourne, I. Vergote, S. Van Huffel, and D. Timmerman. Discrimination between benign and malignant adnexal masses by specialist ultrasound examination: adding CA-125 to ultrasound does not improve diagnostic performance. submitted.

[411] V. Van Belle, B. Van Calster, S. Van Huffel, O. Brouckaert, I. Van den Bempt, S. Pintens, R. Paridaens, M.-R. Christiaens, and P. Neven. Improved npi for breast cancer prognosis. Submitted to the 6th European Breast Cancer Conference, 2008.

[412] B. Van Calster. Does CA-125 help us to assess to risk of malignancy? Invited talk given at the 17th World Congress on Ultrasound in Obstetrics and Gynecology, Florence (Italy), Oct. 2007.

[413] B. Van Calster. Does CA-125 help us to assess to risk of ovarian tumour malignancy? Invited talk given at the 16th World Congress on Ultrasound in Obstetrics and Gynecology, London (UK), Sep. 2006.

[414] B. Van Calster, V. Van Belle, and S. Van Huffel. Multi-class AUC metrics and weighted alternatives. Accepted for publication in the Proceedings of the 21st International Joint Conference on Artificial Neural Networks, 2008.

[415] B. Van Calster, G. Condous, E. Kirk, T. Bourne, D. Timmerman, and S. Van Huffel. Probabilistic multi-class classification of pregnancies of unknown location: comparison of methods with extensive performance evaluation. *Artificial Intelligence in Medicine*, submitted.

[416] B. Van Calster, G. Condous, S. Van Huffel, T. Bourne, and D. Timmerman. Prediction of pregnancy of unknown location outcome: the difficult yet important case of ectopic pregnancy. In J.M. Fonseca, editor, *Proceedings of the Second International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2005)*, pages 98–105. London, IEE, 2005.

[417] B. Van Calster, I. Van den Bempt, M. Drijkoningen, N. Pochet, J. Cheng, S. Van Huffel, W. Hendrickx, J. Decock, H.-J. Huang, K. Leunen, F. Amant, P. Berteloot, R. Paridaens, H. Wildiers, E. Van Limbergen, C. Weltens, D. Timmerman, T. Van Gorp, A. Smeets, W. Van den Bogaert, I. Vergote, M.-R. Christiaens, and P. Neven. Lymph node status of operable breast cancers by combined er, pr and her-2 status: triple positive tumours are more likely lymph node positive. *Breast Cancer Research and Treatment*, in press.

[418] B. Van Calster, W. Hendrickx, P. Neven, H. Wildiers, D. Timmerman, and S. Van Huffel. Piecewise logistic regression models can be a practical solution to nonlinearity in the logit: examples on breast cancer. Poster presented at the 28th Annual Conference of the International Society for Clinical Biostatistics, Alexandroupolis (Greece), Jul. 2007.

[419] B. Van Calster, C. Lu, D. Timmerman, and S. Van Huffel. Applying mackay's evidence procedure to predict malignancy of ovarian tumours. Talk given at the 2nd Workshop on Mathematics and Medical Diagonsis, Erice (Italy), Jul. 2006.

[420] B. Van Calster, J. Luts, J.A.K. Suykens, G. Condous, T. Bourne, D. Timmerman, and S. Van Huffel. Comparing methods for multi-class probabilities in medical decision making using LS-SVMs and kernel logistic regression. In J. Marques de Sá, L.A. Alexandre, W. Duch, and D. Mandic, editors, *Artificial neural networks: ICANN2007*, volume 4669 of *Lecture Notes in Computer Science*, pages 139–148. Berlin, Springer, 2007.

[421] B. Van Calster, I. Nabney, D. Timmerman, and S. Van Huffel. The Bayesian approach: a natural framework for statistical modeling. *Ultrasound in Obstetrics and Gynecology*, 29:485–488, 2007.

[422] B. Van Calster, P. Neven, S. Van Huffel, and J. Røislien. Piecewise logistic regression and the effect of age on breast cancer diagnostics. *Statistics in Medicine*, submitted.

[423] B. Van Calster, T. Smits, and S. Van Huffel. The curse of scoreless draws in soccer: the relationship with a team's offensive, defensive, and overall performance. *Journal of the Quantitative Analysis in Sports*, 4(1):article 4, 2008.

[424] B. Van Calster, D. Timmerman, T. Bourne, A. Testa, C. Van Holsbeke, E. Domali, D. Jurkovic, P. Neven, S. Van Huffel, and L. Valentin. Discrimination between benign and malignant adnexal masses by specialist ultrasound examination versus serum CA-125. *Journal of the National Cancer Institute*, 99:1706–1714, 2007.

[425] B. Van Calster, D. Timmerman, T. Bourne, A. Testa, L. Valentin, and S. Van Huffel. The pre-operative characterization of ovarian pathology using serum ca125 and ultrasonography. *Ultrasound in Obstetrics and Gynecology*, 28:408, 2006. Abstract presented at ISUOG2006, London (UK).

[426] B. Van Calster, D. Timmerman, C. Lu, J.A.K. Suykens, L. Valentin, C. Van Holsbeke, F. Amant, I. Vergote, and S. Van Huffel. Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. *Ultrasound in Obstetrics and Gynecology*, 29:496–504, 2007.

[427] B. Van Calster, D. Timmerman, I.T. Nabney, L. Valentin, A.C. Testa, C. Van Holsbeke, I. Vergote, and S. Van Huffel. Using Bayesian neural networks with ARD input selection to detect malignant ovarian masses prior to surgery. *Neural Computing and Applications*, in press.

[428] B. Van Calster, D. Timmerman, I.T. Nabney, L. Valentin, C. Van Holsbeke, and S. Van Huffel. Classifying ovarian tumors using Bayesian multi-layer perceptrons and automatic relevance determination: a multi-center study. In *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS2006)*, pages 5342–5345, New York City (USA), 2006.

[429] B. Van Calster, D. Timmerman, A.C. Testa, L. Valentin, and S. Van Huffel. Multi-class classification of ovarian tumors. Accepted for publication in the Proceedings of the 11th European Symposium on Artificial Neural Networks, 2008.

[430] B. Van Calster, D. Timmerman, L. Valentin, A. Testa, C. Van Holsbeke, and S. Van Huffel. Pre-operative prediction of ovarian tumours as benign, primary invasive, borderline, or metastatic. Talk given at the LKI Oncoforum, Leuven (Belgium), Feb. 2008.

[431] B. Van Calster, D. Timmerman, L. Valentin, A. Testa, C. Van Holsbeke, and S. Van Huffel. Logistic regression models to distinguish between benign, primary invasive, borderline malignant and metastatic ovarian tumors. *Ultrasound in Obstetrics and Gynecology*, 30:414, 2007. Abstract presented at ISUOG2007, Florence (Italy).

[432] B. Van den Bergh, B. Van Calster, S. Pinna Puissant, and S. Van Huffel. Adrenocortical activity is related to physical complaints, behavioral and emotional disorders. Poster presented at the 4th International Conference on the (Non)Expression of Emotions in Health and Disease, Tilburg (the Netherlands), Oct. 2007.

[433] B. Van den Bergh, B. Van Calster, T. Smits, S. Van Huffel, and L. Lagae. Prenatal maternal anxiety predicts flattening of the day-time cortisol profile and depression in adolescence. *Early Human Development*, 82:541, 2006. Abstract presented at DOHaD2006, Utrecht (Netherlands).

[434] B. Van den Bergh, B. Van Calster, T. Smits, L. Lagae, and S. Van Huffel. Cortisol and depression in 14-15 year olds are linked to maternal anxiety during pregnancy. Paper presented at the 19th Biennial Meeting of the International Society for the Study of Behavioural Development, Melbourne (Australia), Jul. 2006.

[435] B. Van den Bergh, B. Van Calster, T. Smits, L. Lagae, and S. Van Huffel. Hpa axis dysregulation induced by antenatal maternal anxiety enhances neurobiological vulnerability to depression in adolescents. Paper presented at the 33rd Biennial Meeting of the Society for Research in Child Development, Boston (USA), Mar. 2007.

[436] B.R.H. Van den Bergh, B. Van Calster, S. Pinna Puissant, and S. Van Huffel. Bidirectional associations between diurnal cortisol profiles and self- and other-reported symptoms of depressed mood, trait anxiety and aggressive behavior in post-puberty. *Hormones and Behavior*, submitted.

[437] B.R.H. Van den Bergh, B. Van Calster, T. Smits, S. Van Huffel, and L. Lagae. Antenatal maternal anxiety is related to HPA-axis dysregulation and self-reported depressive symptoms in adolescence: a prospective study on the fetal origins of depressed mood. *Neuropsychopharmacology*, 33:536–545, 2008.

[438] B.R.H. Van den Bergh, B. Van Calster, T. Smits, L. Lagae, and S. Van Huffel. A flattened diurnal cortisol profile and depression in adolescence are related to antenatal maternal anxiety. *Cell Stress & Chaperones*, 12:e1 (abstract 6K_04_P), 2007. Abstract presented at CSSi2007, Budapest (Hungary).

[439] T. Van Gestel, J. Suykens, D. Baestaens, A. Lambrechts, G. Lanckriet, B. Vandaele, B. De Moor, and J. Vandewalle. Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Transactions on Neural Networks*, 12:809–821, 2001.

[440] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle. Multiclass LS-SVMs: moderated outputs and colding-decoding schemes. *Neural Processing Letters*, 15:45–58, 2002.

[441] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle. A Bayesian framework for least squares support vector machine classifiers. *Neural Computation*, 15:1115–1148, 2002.

[442] T. Van Gestel, J.A.K. Suykens, B. De Moor, and J. Vandewalle. Automatic relevance determination for least squares support vector machines classifiers. In *Proceedings of the 9th Eureopean Symposium on Artificial Neural Networks (ESANN2001)*, pages 13–18, Bruges (Belgium), 2001.

[443] C. Van Holsbeke, B. Van Calster, A.C. Testa, E. Domali, C. Lu, S. Van Huffel, L. Valentin, and D. Timmerman. Predicting malignancy of ovarian tumors: prospective evaluation of models from the IOTA study. submitted.

[444] C. Van Holsbeke, B. Van Calster, L. Valentin, A. Testa, E. Domali, C. Lu, S. Van Huffel, and D. Timmerman. The first prospective evaluation of the IOTA phase I mathematical models to predict the character of an adnexal mass. *Ultrasound in Obstetrics and Gynecology*, 30:414–415, 2007. Abstract presented at ISUOG2007, Florence (Italy).

[445] C. Van Holsbeke, B. Van Calster, L. Valentin, A. Testa, E. Ferrazzi, Ph. Moerman, S. Van Huffel, I. Vergote, and D. Timmerman. External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by

the International Ovarian Tumor Analysis (IOTA) group. *Ultrasound in Obstetrics and Gynecology*, 28:405, 2006. Abstract presented at ISUOG2006, London (UK).

[446] C. Van Holsbeke, B. Van Calster, L. Valentin, A.C. Testa, E. Ferrazzi, I. Dimou, C. Lu, P. Moerman, S. Van Huffel, I. Vergote, and D. Timmerman. External validation of mathematical models to distinguish between benign and malignant adnexal tumors. Poster presented at the 15th International Meeting of the European Society of Gynaecological Oncology, Berlin (Germany), Oct. 2007.

[447] C. Van Holsbeke, B. Van Calster, L. Valentin, A.C. Testa, E. Ferrazzi, I. Dimou, C. Lu, P. Moerman, S. Van Huffel, I. Vergote, and D. Timmerman. External validation of mathematical models to distinguish between benign and malignant adnexal tumors. *Clinical Cancer Research*, 13:4440–4447, 2007.

[448] J.C. Van Houwelingen and S. Le Cessie. Predictive value of statistical models. *Statistics in Medicine*, 9:1303–1325, 1990.

[449] V. Vapnik. *The nature of statistical learning theory*. Springer, New York, 1995.

[450] V. Vapnik. *Statistical learning theory*. Wiley, New York, 1998.

[451] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:paper 91, 2006.

[452] A. Vellido and P.J.G. Lisboa. An electronic commerce application of the Bayesian framework for MLPs: the effect of marginalisation and ARD. *Neural Computing and Applications*, 10:3–11, 2001.

[453] G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer, New York, 2000.

[454] I. Vergote, J. De Brabanter, A. Fyles, K. Bertelsen, N. Einhorn, P. Sevelda, M.E. Gore, J. Kærn, H. Verrelst, K. Sjövall, D. Timmerman, J. Vandewalle, M. Van Gramberen, and C.G. Tropé. Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. *Lancet*, 357:176–182, 2001.

[455] S. Viaene, B. Baesens, T. Van Gestel, J. Suykens, D. Van den Poel, J. Vanthienen, B. De Moor, and G. Dedene. Knowledge discovery in a direct marketing case using least squares support vector machines. *International Journal of Intelligent Systems*, 16:1023–1036, 2001.

[456] A.J. Vickers and E.B. Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26:565–574, 2006.

[457] F. Vivarelli and C.K.I. Williams. Comparing Bayesian neural network algorithms for classifying segmented outdoor images. *Neural Networks*, 14:427–437, 2001.

[458] G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

[459] W.B. Warren, I.E. Timor-Tritsch, D.B. Peisner, S. Raju, and M.G. Rosen. Dating the early pregnancy by sequential appearance of embryonic structures. *American Journal of Obstetrics and Gynecology*, 161:747–753, 1989.

[460] A.E. Washington and P. Katz. Ectopic pregnancy in the United States: economic consequences and payment source trends. *Obstetrics & Gynecology*, 81:287–292, 1993.

[461] L.A.M. Welberg and J.R. Seckl. Prenatal stress, glucocorticoids and the programming of the brain. *Journal of Neuroendocrinology*, 13:113–128, 2001.

[462] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the Seventh Eureopean Symposium on Artificial Neural Networks (ESANN1999)*, pages 219–224, Bruges (Belgium), 1999.

[463] H. Wildiers, B. Van Calster, W. Hendrickx, S. Van Huffel, A. Smeets, R. Paridaens, D. Timmerman, M. Drijkoningen, K. Leunen, C. Weltens, M.-R. Christiaens, and P. Neven. Axillary lymph node involvement in breast cancer decreases with age but increases again after age 70 mainly in smaller tumors. Poster presented at the 10th Biennial Conference on Primary Therapy of Early Breast Cancer, St Gallen (Switzerland), Mar. 2007.

[464] H. Wildiers, B. Van Calster, L.V. van de Poll-Franse, W. Hendrickx, S. Van Huffel, A. Smeets, R. Paridaens, K. Deraedt, M.-R. Christiaens, and P. Neven. External validation of a piecewise effect model of axillary lymph node involvement in elderly breast cancer patients. Submitted to the 6th European Breast Cancer Conference, 2008.

[465] H. Wildiers, B. Van Calster, L.V. van de Poll-Franse, W. Hendrickx, S. Van Huffel, A. Smeets, R. Paridaens, K. Deraedt, K. Leunen, C. Weltens, M.-R. Christiaens, and P. Neven. Axillary lymph node involvement in breast cancer decreases with age but increases again after age 70 only in smaller tumors. *Journal of Clinical Oncology*, submitted.

[466] J.C. Wojak, M.J. Clayton, and T.E. Nolan. Outcomes of ultrasound diagnosis of ectopic pregnancy. Dependence on observer experience. *Investigative Radiology*, 30:115–117, 1995.

[467] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1:67–82, 1997.

[468] R. Wooster, G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, and G. Micklem. Identification of the breast cancer susceptibility gene brca2. *Nature*, 378:789–792, 1995.

[469] R. Wooster and B.L. Weber. Breast and ovarian cancer. *New England Journal of Medicine*, 348:2339–2347, 2003.

[470] S. Wu and P. Flach. A scored AUC metric for classifier evaluation and selection. In *Proceedings of the 2nd Workshop on ROC Analysis in Machine Learning (ROCML2005)*, Bonn (Germany), 2005.

[471] S. Wu, P. Flach, and C. Ferri. An improved model selection heuristic for AUC. In J.N. Kok, J. Koronacki, R. López de Mántaras, S. Matwin, D. Mladenic, and A. Skowron, editors, *Machine Learning: ECML2007*, volume 4701 of *Lecture Notes in Computer Science*, pages 478–489. Berlin, Springer, 2007.

[472] T.-F. Wu, C.-J. Lin, and R.C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.

[473] J. Wyatt. Computer-based knowledge systems. *Lancet*, 338:1431–1436, 1991.

[474] J.C. Wyatt and D.G. Altman. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal*, 311:1539–1541, 1995.

[475] J.C. Wyatt and J.L.Y. Liu. Basic concepts in medical informatics. *Journal of Epidemiology and Community Health*, 56:808–812, 2002.

[476] N. Yukinawa, S. Oba, K. Kato, and S. Ishii. Multi-class pattern classification based on a probabilistic model of combining binary classifiers. In W. Duch, J. Kacprzyk, E. Oja, and S. Zadrozny, editors, *Artificial neural networks: formal models and their applications - ICANN2005*, volume 3697 of *Lecture Notes in Computer Science*, pages 337–342. Berlin, Springer, 2005.

[477] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2002)*, pages 694–699, Edmonton (Canada), 2002.

[478] S.B. Zane, B.A. Kieke Jr, J.S. Kendrick, and C. Bruce. Surveillance in a time of changing health care practices: estimating ectopic pregnancy incidence in the United States. *Maternal and Child Health Journal*, 6:227–236, 2002.

[479] D.D. Zhang, X.H. Zhou, D.H. Freeman Jr, and J.L. Freeman. A non-parametric approach for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statistics in Medicine*, 21:701–715, 2002.

[480] J. Zhu and T. Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5:427–443, 2004.

[481] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14:185–205, 2005.

[482] K.H. Zou, W.J. Hall, and D.E. Shapiro. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16:2143–2156, 1997.

Publication list

Papers in international journals

G. Condous, E. Kirk, A. Syed, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Do levels of serum cancer antigen 125 and creatine kinase predict the outcome in pregnancies of unknown location? *Human Reproduction*, 20:3348-3354, 2005.

D. Timmerman, A.C. Testa, T. Bourne, E. Ferrazzi, L. Ameye, M.L. Konstantinovic, B. Van Calster, W.P. Collins, I. Vergote, S. Van Huffel, L. Valentin. 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. *Journal of Clinical Oncology*, 23:8794-8801, 2005.

Z. Haider, G. Condous, A. Khalid, E. Kirk, F. Mukri, B. Van Calster, D. Timmerman, T. Bourne. The impact of the availability of ultrasonography in the acute gynaecology unit. *Ultrasound in Obstetrics and Gynecology*, 28:207-213, 2006.

O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, G. Condous. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Human Reproduction*, 21:1824-1831, 2006.

G. Condous, E. Kirk, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Failing pregnancies of unknown location: A prospective evaluation of the human chorionic gonadotrophin ratio. *BJOG - An International Journal of Obstetrics and Gynaecology*, 113:521-527, 2006.

G. Condous, E. Kirk, C. Lu, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. There is no role for uterine curettage in the contemporary diagnostic work-up of women with a pregnancy of unknown location. *Human Reproduction*, 21:2706-2710, 2006.

E. Kirk, G. Condous, B. Van Calster, Z. Haider, S. Van Huffel, D. Timmerman, T. Bourne. A validation of the most commonly used protocol to predict the success of single dose methotrexate in the treatment of ectopic pregnancy. *Human*

Reproduction, 22:858-863, 2007.

G. Condous, B. Van Calster, E. Kirk, D. Timmerman, S. Van Huffel, T. Bourne. Prospective cross-validation of three methods of predicting failing pregnancies of unknown location. *Human Reproduction*, 22:1156-1160, 2007.

B. Van Calster, D. Timmerman, C. Lu, J.A.K. Suykens, L. Valentin, C. Van Holsbeke, F. Amant, I. Vergote, S. Van Huffel. Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. *Ultrasound in Obstetrics and Gynecology*, 29:496-504, 2007.

B. Van Calster, I. Nabney, D. Timmerman, S. Van Huffel. The Bayesian approach: a natural framework for statistical modeling. *Ultrasound in Obstetrics and Gynecology*, 29:485-488, 2007.

G. Condous, B. Van Calster, S. Van Huffel, A. Lam. What is the value of preoperative bimanual pelvic examination in women undergoing laparoscopic total hysterectomy? *Journal of Minimally Invasive Gynecology*, 14:334-338, 2007.

G. Condous, B. Van Calster, E. Kirk, Z. Haider, D. Timmerman, S. Van Huffel, T. Bourne. Prediction of ectopic pregnancy in women with a pregnancy of unknown location. *Ultrasound in Obstetrics and Gynecology*, 29:680-687, 2007.

E. Kirk, G. Condous, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Rationalizing the follow-up of pregnancies of unknown location. *Human Reproduction*, 22:1744-1750, 2007.

C. Van Holsbeke, B. Van Calster, L. Valentin, A.C. Testa, E. Ferrazzi, I. Dimou, C. Lu, P. Moerman, S. Van Huffel, I. Vergote, D. Timmerman. External validation of mathematical models to distinguish between benign and malignant adnexal tumors. *Clinical Cancer Research*, 13:4440-4447, 2007.

G. Condous, B. Van Calster, E. Kirk, Z. Haider, D. Timmerman, S. Van Huffel, T. Bourne. Clinical information does not improve the performance of mathematical models in predicting the outcome of pregnancies of unknown location. *Fertility and Sterility*, 88:572-580, 2007.

D. Timmerman, B. Van Calster, T. Bourne, A. Testa, C. Van Holsbeke, E. Domali, D. Jurkovic, P. Neven, S. Van Huffel, L. Valentin. Subjective assessment based on ultrasonography and the use of serum CA-125 to predict malignancy in ovarian tumors. *Journal of Clinical Oncology*, 25:4194-4200, 2007.

F. Amant, A. De Knijf, B. Van Calster, K. Leunen, P. Neven, P. Berteloot, I. Vergote, S. Van Huffel, P. Moerman. Clinical study investigating the role

of lymphadenectomy, surgical castration and adjuvant hormonal treatment in endometrial stromal sarcoma. *British Journal of Cancer*, 97:1194-1199, 2007.

B. Van Calster, D. Timmerman, T. Bourne, A. Testa, C. Van Holsbeke, E. Domali, D. Jurkovic, P. Neven, S. Van Huffel, L. Valentin. Discrimination between benign and malignant adnexal masses by specialist ultrasound examination versus serum CA-125. *Journal of the National Cancer Institute*, 99:1706-1714, 2007.

B.R.H. Van den Bergh, B. Van Calster, T. Smits, S. Van Huffel, L. Lagae. Antenatal maternal anxiety is related to HPA-axis dysregulation and self-reported depressive symptoms in adolescence: a prospective study on the fetal origins of depressed mood. *Neuropsychopharmacology*, 33:536-545, 2008.

B. Van Calster, T. Smits, S. Van Huffel. The curse of scoreless draws in soccer: the relationship with a team's offensive, defensive, and overall performance. *Journal of Quantitative Analysis in Sports*, 4(1):article 4, 2008.

P. Neven, B. Van Calster, I. Van den Bempt, S. Van Huffel, V. Van Belle, W. Hendrickx, J. Decock, H. Wildiers, R. Paridaens, F. Amant, K. Leunen, P. Berteloot, D. Timmerman, E. Van Limbergen, C. Weltens, W. Van den Bogaert, A. Smeets, I. Vergote, M.-R. Christiaens, M. Drijkoningen. Age interacts with the expression of steroid and HER-2 receptors in operable invasive breast cancer. *Breast Cancer Research and Treatment*, in press.

B. Van Calster, D. Timmerman, I.T. Nabney, L. Valentin, A.C. Testa, C. Van Holsbeke, I. Vergote, S. Van Huffel. Using Bayesian Neural Networks with ARD input selection to detect malignant ovarian masses prior to surgery. *Neural Computing and Applications*, in press.

D. Timmerman, A.C. Testa, T. Bourne, L. Ameye, D. Jurkovic, C. Van Holsbeke, D. Paladini, B. Van Calster, I. Vergote, S. Van Huffel, L. Valentin. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound in Obstetrics and Gynecology*, in press.

B. Van Calster, I. Van den Bempt, M. Drijkoningen, N. Pochet, J. Cheng, S. Van Huffel, W. Hendrickx, J. Decock, H.-J. Huang, K. Leunen, F. Amant, P. Berteloot, R. Paridaens, H. Wildiers, E. Van Limbergen, C. Weltens, D. Timmerman, T. Van Gorp, A. Smeets, W. Van den Bogaert, I. Vergote, M.-R. Christiaens, P. Neven. Lymph node status of operable breast cancers by combined ER, PR and HER-2 status: Triple positive tumours are more likely lymph node positive. *Breast Cancer Research and Treatment*, in press.

L. Morales, S. Pans, K. Verschueren, B. Van Calster, R. Paridaens, R. Westhovens, D. Timmerman, L. De Smet, I. Vergote, M.-R. Christiaens, P. Neven.

A prospective study to assess the short term intra-articular and tenosynovial changes in the aromatase inhibitor-associated arthralgia syndrome. *Journal of Clinical Oncology*, in press.

Letters, comments, and replies

D. Timmerman, C. Van Holsbeke, T. Van den Bosch, B. Van Calster, S. Van Huffel, I. Vergote. Letter to the editor. *International Journal of Gynecological Cancer*, 17:543, 2007.

D. Timmerman, T. Bourne, A.C. Testa, F. Amant, B. Van Calster, C. Van Holsbeke, S. Van Huffel, I. Vergote, L. Valentin. Reply to "Pitfalls of CA-125 levels in the preoperative workup of ovarian masses" by Denis Querleu et al. *Journal of Clinical Oncology*, 26:513, 2008.

P. Neven, O. Brouckaert, V. Van Belle, I. Van den Bempt, W. Hendrickx, H. Cho, K. Deraedt, B. Van Calster, S. Van Huffel, P. Moerman, F. Amant, K. Leunen, A. Smeets, H. Wildiers, R. Paridaens, I. Vergote, M.-R. Christiaens. In early-stage breast cancer, the estrogen receptor interacts with the correlation between the HER-2 status and age at diagnosis, tumour grade and lymph node involvement. *Journal of Clinical Oncology*, in press.

Papers submitted to international journals

B. Van Calster, G. Condous, E. Kirk, T. Bourne, D. Timmerman, S. Van Huffel. Probabilistic multi-class classification of pregnancies of unknown location: comparison of methods with extensive performance evaluation. Submitted to *Artificial Intelligence in Medicine*, 2007.

B. Van Calster, S. Van Huffel, P. Neven, J. Røislien. Piecewise logistic regression and the effect of age on breast cancer diagnostics. Submitted to *Statistics in Medicine*, 2007.

E. Kirk, S. Bora, B. Van Calster, G. Condous, S. Van Huffel, D. Timmerman, T. Bourne. Chlamydia trachomatis infection in patients attending an early pregnancy unit: prevalence, symptoms, pregnancy location and viability. Submitted to *Acta Obstetricia et Gynecologica Scandinavica*, 2007.

B. Van den Bergh, B. Van Calster, S. Pinna Puissant, S. Van Huffel. Bidirectional associations between diurnal cortisol profiles and self- and other-reported symptoms of depressed mood, trait anxiety and aggressive behavior in post-puberty. Submitted to *Hormones and Behavior*, 2007.

I. Dimou, B. Van Calster, S. Van Huffel, D. Timmerman, M. Zervakis. Evaluation of imputation methods in ovarian tumour diagnostic models using generalised linear models and support vector machines. Submitted to *Medical Decision Making*, 2008.

C. Van Holsbeke, B. Van Calster, A.C. Testa, E. Domali, C. Lu, S. Van Huffel, L. Valentin, D. Timmerman. Predicting malignancy of ovarian tumors: prospective evaluation of models from the IOTA study. Submitted to *Clinical Cancer Research*, 2008.

L. Valentin, D. Jurkovic, B. Van Calster, A. Testa, C. Van Holsbeke, T. Bourne, I. Vergote, S. Van Huffel, D. Timmerman. Discrimination between benign and malignant adnexal masses by specialist ultrasound examination: adding CA-125 to ultrasound does not improve diagnostic performance. Submitted to *British Medical Journal*, 2008.

H.P. Wildiers, B. Van Calster, L.V. van de Poll-Franse, W. Hendrickx, S. Van Huffel, A. Smeets, R. Paridaens, K. Deraedt, K. Leunen, C. Weltens, M.-R. Christiaens, P. Neven. Axillary lymph node involvement in breast cancer decreases with age but increases again after age 70 only in smaller tumors. Submitted to *Journal of Clinical Oncology*, 2008.

Papers in preparation

E. Kirk, L. Tan, G. Condous, B. Van Calster, N. Rosello, D. Gould, P. Alnaes-Katjavivi, O. Istre, S. Van Huffel, D. Timmerman, T. Bourne. External validation of a mathematical model to predict the outcome of pregnancies of unknown location (PULs).

T. Smits, B. Van Calster, S. Van Huffel. Scoring personality: soccer game outcomes as behavioral correlates of cross-cultural differences in personality and value orientation.

T. Bignardi, G. Condous, E. Kirk, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Ultrasound and not biochemistry determines the ultimate viability of intra-uterine pregnancies in the pregnancy of unknown location population.

A. Khalid, G. Condous, E. Kirk, B. Van Calster, D. Timmerman, S. Van Huffel, T. Bourne. What is the value of ultrasound in the contemporary diagnostic work-up of women with possible underlying polycystic ovarian syndrome?

B. Van Calster, A. Khalid, T. Smits, G. Condous, E. Kirk, T. Bourne, D. Timmerman, S. Van Huffel. Examining the relationship between ultrasound features

and clinical symptoms of PCOS using Bayesian networks.

B. Van Calster, G. Condous, E. Kirk, Z. Haider, D. Timmerman, S. Van Huffel, T. Bourne. Discovering the outcome of pregnancies of unknown location using mathematical models: overview and visualization of predictions.

Z. Haider, G. Condous, E. Kirk, A. Syed, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. The development of a logistic regression model to predict chlamydia trachomatis infection in women attending an acute gynaecology unit.

E. D'Hondt, G. Moulin-Romsee, F. Persyn, B. Van Calster, S. Ceyssens, P. Neven, S. Stroobants. Detection of sentinel lymph nodes in breast cancer: impact of methodology on clinical performance.

Papers in international conference proceedings

B. Van Calster, G. Condous, S. Van Huffel, T. Bourne, D. Timmerman. Prediction of pregnancy of unknown location outcome: The difficult yet important case of ectopic pregnancy. In J.M. Fonseca (ed.), Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare (Costa da Caparica, Portugal, Jun. 2005), p. 98-105. London, IEE, 2005.

B. Van Calster, D. Timmerman, I.T. Nabney, L. Valentin, C. Van Holsbeke, S. Van Huffel. Classifying ovarian tumors using Bayesian Multi-Layer Perceptrons and Automatic Relevance Determination: A multi-center study. Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (New York City, USA, Sep. 2006), pp. 5342-5345. Los Alamitos, IEEE Computer Society Press, 2006.

M.S.H. Aung, P.J.G. Lisboa, T.A.E. Etchells, A.C. Testa, B. Van Calster, S. Van Huffel, L. Valentin, D. Timmerman. Comparing analytical decision support models through Boolean rule extraction: a case study of ovarian tumor malignancy. In D. Liu, S. Fei, Z.G. Hou, H. Zhang, C. Sun (eds.), Advances in Neural Networks - ISNN2007 (Nanjing, China, Jun. 2007), Lecture Notes in Computer Science Volume 4492, pp. 1177-1186. Berlin, Springer, 2007.

I. Dimou, B. Van Calster, S. Van Huffel, D. Timmerman, M. Zervakis. Comparison of imputation approaches in ovarian tumour diagnostic models based on LS-SVMs. In E. Ifeachor E, A. Anastasiou A (eds.), Third International Conference on Computational Intelligence in Medicine and Healthcare (Plymouth, UK, Jul. 2007), p. 46. Plymouth, University of Plymouth, 2007. (full paper on CD-ROM)

B. Van Calster, J. Luts, J.A.K. Suykens, G. Condous, T. Bourne, D. Timmerman, S. Van Huffel. Comparing methods for multi-class probabilities in medical

decision making using LS-SVMs and kernel logistic regression. In J. Marques de Sá, L.A. Alexandre, W. Duch, D. Mandic (eds.), Artificial Neural Networks - ICANN 2007 (Porto, Portugal, Sep. 2007), Lecture Notes in Computer Science Volume 4669, pp. 139-148. Berlin, Springer, 2007.

B. Van Calster, D. Timmerman, A.C. Testa, L. Valentin, S. Van Huffel. Multi-class classification of ovarian tumours. Paper to be published in the Proceedings of the 11th European Symposium on Artificial Neural Networks (Bruges, Belgium, Apr. 2008), 2008.

B. Van Calster, V. Van Belle, S. Van Huffel. Multi-class AUC metrics and weighted alternatives. Paper to be published in the Proceedings of the 21st International Joint Conference on Neural Networks (Hong Kong, China, Jun. 2008), 2008.

B. Van Calster, S. Van Huffel, D. Timmerman, E. Kirk, T. Bourne, G. Condous. Towards a clinical decision support system for pregnancies of unknown location. Paper to be published in the Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems (Jyväskylä, Finland, Jun. 2008), 2008.

Conference abstracts in international journals

Blair Bell Research Society, Jun. 2005, Liverpool, UK

Z. Haider, G. Condous, E. Kirk, A. Syed, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Development of logistic regression models to predict chlamydial infection in women attending an acute gynaecology unit (AGU). *BJOG - An International Journal of Obstetrics and Gynaecology*, 112:1451-1452, 2005.

15th World Congress on Ultrasound in Obstetrics and Gynecology, Sep. 2005, Vancouver, Canada

E. Kirk, G. Condous, Z. Haider, T. Chang, B. Van Calster, T. Bourne. Can the ultrasonographic appearance of an ectopic pregnancy determine the likelihood of success for the conservative management? *Ultrasound in Obstetrics and Gynecology*, 26:361-362, 2005.

G. Condous, E. Kirk, Z. Haider, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. What is the optimal approach to accurately classifying failing pregnancies of unknown location (PULs)? *Ultrasound in Obstetrics and Gynecology*, 26:362, 2005.

G. Condous, E. Kirk, Z. Haider, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Can we improve the performance of logistic regression analysis for predicting the outcome of pregnancies of unknown location (PULs)? *Ultrasound*

in Obstetrics and Gynecology, 26:455, 2005.

E. Kirk, G. Condous, B. Van Calster, Z. Haider, T. Bourne. Is it possible to predict successful treatment response to methotrexate earlier than seven days? *Ultrasound in Obstetrics and Gynecology*, 26:456, 2005.

14th International Meeting of the European Society of Gynaecological Oncology, Sep. 2005, Istanbul, Turkey

D. Timmerman, B. Van Calster, D. Jurkovic, L. Valentin, A.C. Testa, J.P. Bernard, C. Van Holsbeke, S. Van Huffel, I. Vergote, T. Bourne. Is serum CA125 essential to distinguish between malignant and benign adnexal tumours? A multicentre study by the International Ovarian Tumour Analysis (IOTA) Group. *International Journal of Gynecological Cancer*, 15s2:178-179, 2005.

Blair Bell Research Society's Biannual Competition Meeting, Nov. 2005, London, UK

E. Kirk, G. Condous, Z. Haider, B. Van Calster, D. Timmerman, T. Bourne. Is the ultrasonographic appearance of an ectopic pregnancy a predictor of success for conservative management strategies? *BJOG - An International Journal of Obstetrics and Gynaecology*, 113:864, 2006.

16th World Congress on Ultrasound in Obstetrics and Gynecology, Sep. 2006, London, UK

E. Kirk, G. Condous, C. Bottomley, B. Van Calster, D. Timmerman, S. Van Huffel, T. Bourne. Ultrasound and not biochemistry determines the ultimate viability of intra-uterine pregnancies in the PUL population. *Ultrasound in Obstetrics and Gynecology*, 28:364, 2006.

E. Kirk, G. Condous, C. Bottomley, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. What is the optimal approach to classifying viable intra-uterine pregnancies in the PUL population? *Ultrasound in Obstetrics and Gynecology*, 28:364, 2006.

E. Kirk, G. Condous, C. Bottomley, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Using a mathematical model in the management of pregnancies of unknown location: a prospective interventional study. *Ultrasound in Obstetrics and Gynecology*, 28:364, 2006.

E. Kirk, P. Alnaes-Katjavivi, G. Condous, B. Van Calster, S. Van Huffel, O. Istre, D. Timmerman, T. Bourne. Validation of a logistic regression model to predict the outcome of pregnancies of unknown location in a new patient population. *Ultrasound in Obstetrics and Gynecology*, 28:365, 2006.

C. Van Holsbeke, B. Van Calster, L. Valentin, A. Testa, E. Ferrazzi, P. Moerman, S. Van Huffel, I. Vergote, D. Timmerman. External validation of mathematical models to distinguish benign and malignant adnexal tumors: a multicenter study by the international ovarian tumor analysis (IOTA) group. *Ultrasound in Obstetrics and Gynecology*, 28:405, 2006.

D. Timmerman, B. Van Calster, D. Jurkovic, L. Valentin, A. Testa, J.P. Bernard, C. Van Holsbeke, S. Van Huffel, I. Vergote, T. Bourne. Shortcomings of serum CA125 in diagnosing ovarian cancer. *Ultrasound in Obstetrics and Gynecology*, 28:408, 2006.

B. Van Calster, D. Timmerman, T. Bourne, A. Testa, L. Valentin, S. Van Huffel. The pre-operative characterization of ovarian pathology using serum CA125 and ultrasonography. *Ultrasound in Obstetrics and Gynecology*, 28:408, 2006.

D. Timmerman, E. Domali, B. Van Calster, A. Testa, C. Van Holsbeke, M. Konstantinovic, S. Van Huffel, L. Valentin. Prospective internal evaluation of IOTA logistic regression model. *Ultrasound in Obstetrics and Gynecology*, 28:506, 2006.

Z. Haider, E. Kirk, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Transvaginal ultrasonography and saline infusion sonography compared with outpatient hysteroscopy for the diagnosis of endometrial polyps and submucous fibroids. *Ultrasound in Obstetrics and Gynecology*, 28:607, 2006.

4th World Congress on Developmental Origins of Health and Disease, Sep. 2006, Utrecht, the Netherlands

B. Van den Bergh, B. Van Calster, T. Smits, S. Van Huffel, L. Lagae. Prenatal maternal anxiety predicts flattening of the day-time cortisol profile and depression in adolescence. *Early Human Development*, 82:541, 2006.

Blair Bell Research Society, Oct. 2006, Leicester, UK

E. Kirk, L. Tan, G. Condous, B. Van Calster, C. Bottomley, S. Van Huffel, D. Timmerman, T. Bourne. Selecting failing pregnancies in the pregnancy of unknown location (PUL) population to reduce follow-up. *BJOG - An International Journal of Obstetrics and Gynaecology*, 114:1037, 2007.

3rd International Congress on Stress Responses in Biology and Medicine, Aug. 2007, Budapest, Hungary

B.R.H. Van den Bergh, B. Van Calster, T. Smits, L. Lagae, S. Van Huffel. A flattened diurnal cortisol profile and depression in adolescence are related to antenatal maternal anxiety. *Cell Stress & Chaperones*, 12:e1 (Abstract 6K_04_P),

2007.

17th World Congress on Ultrasound in Obstetrics and Gynecology, Oct. 2007, Florence, Italy

B. Van Calster, D. Timmerman, L. Valentin, A. Testa, C. Van Holsbeke, S. Van Huffel. Logistic regression models to distinguish between benign, primary invasive, borderline malignant and metastatic ovarian tumors. *Ultrasound in Obstetrics and Gynecology*, 30:414, 2007.

C. Van Holsbeke, B. Van Calster, L. Valentin, A. Testa, E. Domali, C. Lu, S. Van Huffel, D. Timmerman. The first prospective evaluation of the IOTA phase 1 mathematical models to predict the character of an adnexal mass. *Ultrasound in Obstetrics and Gynecology*, 30:414-415, 2007.

L. Valentin, D. Jurkovic, B. Van Calster, A. Testa, C. Van Holsbeke, T. Bourne, I. Vergote, S. Van Huffel, D. Timmerman. CA-125 does not add to and is not superior to pattern recognition for discrimination between benign and malignant adnexal masses. *Ultrasound in Obstetrics and Gynecology*, 30:415, 2007.

20th Annual Congress of the European Association of Nuclear Medicine, Oct. 2007, Copenhagen, Denmark

G. Moulin-Romsee, E. D'Hondt, F. Persyn, B. Van Calster, S. Ceyssens, P. Neven, S. Stroobants. Detection of sentinel lymph nodes in breast cancer: every day practice and pitfalls. *European Journal of Nuclear Medicine and Molecular Imaging*, 34(Suppl. 2):S387.

Unpublished conference contributions

B. Van Calster, G. Condous, S. Van Huffel, T. Bourne, D. Timmerman. The classification of pregnancies of unknown location: recent results and future directions. Poster presented at the IAP V/22 study day, Liege, Belgium, May 2005.

G. Condous, E. Kirk, Z. Haider, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Can we improve the performance of diagnostic tests to predict the outcome of pregnancies of unknown location? Presented at the 35th Annual Scientific Meeting of the Australasian Society for Ultrasound in Medicine, Adelaide, Australia, Oct. 2005.

G. Condous, E. Kirk, Z. Haider, B. Van Calster, T. Bourne. Can we use the ultrasonographic appearance of an ectopic pregnancy to predict the likelihood of success for expectant and medical management? Presented at the 35th Annual Scientific Meeting of the Australasian Society for Ultrasound in Medicine, Adelaide,

Australia, Oct. 2005.

G. Condous, E. Kirk, Z. Haider, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Can we predict the outcome of medical management of ectopic pregnancies earlier than one week? Presented at the 35th Annual Scientific Meeting of the Australasian Society for Ultrasound in Medicine, Adelaide, Australia, Oct. 2005.

B. Van Calster, C. Lu, D. Timmerman, S. Van Huffel. Applying MacKay's evidence procedure to predict malignancy of ovarian tumours. Presented at the 2nd Workshop on Mathematics and Medical Diagnosis, Erice, Italy, Jul. 2006.

B. Van den Bergh, B. Van Calster, T. Smits, L. Lagae, S. Van Huffel. Cortisol and depression in 14-15 year olds are linked to maternal anxiety during pregnancy. Presented at the 19th Biennial Meeting of the International Society for the Study of Behavioural Development, Melbourne, Australia, Jul. 2006.

B. Van Calster. Does CA-125 help us to assess the risk of ovarian tumour malignancy? Invited talk at the 16th World Congress on Ultrasound in Obstetrics and Gynecology, London, UK, Sep. 2006.

D. Timmerman, A. Testa, T. Bourne, E. Ferrazzi, L. Ameye, C. Van Holsbeke, B. Van Calster, W. Collins, I. Vergote, S. Van Huffel, L. Valentin. Mathematical models to distinguish between malignant and benign adnexal masses: results from a multicenter study by the International Ovarian Tumor Analysis (IOTA) group. Presented at the 36th Annual Scientific Meeting of the Australasian Society for Ultrasound in Medicine, Melbourne, Australia, Sep. 2006.

D. Timmerman, A. Testa, T. Bourne, E. Ferrazzi, L. Ameye, C. Van Holsbeke, B. Van Calster, W.P. Collins, I. Vergote, S. Van Huffel, L. Valentin. Lessons from the IOTA study. Invited keynote lecture at the 15th National Congress of the Italian Society of Ultrasound in Obstetrics and Gynecology, Brescia, Italy, Sep. 2006.

B.R.H. Van den Bergh, B. Van Calster, T. Smits, L. Lagae, S. Van Huffel. HPA axis dysregulation induced by antenatal maternal anxiety enhances neurobiological vulnerability to depression in adolescents. Presented at the 33rd Biennial Meeting of the Society for Research in Child Development, Boston, USA, Mar. 2007.

P. Neven, I. Van den Bempt, B. Van Calster, N. Pochet, J. Cheng, S. Van Huffel, W. Hendrickx, J. Decock, K. Leunen, F. Amant, P. Berteloot, H. Wildiers, R. Paridaens, E. Van Limbergen, C. Weltens, D. Timmerman, T. Van Gorp, A. Smeets, W. Van den Bogaert, I. Vergote, M.-R. Christiaens, M. Drijkoningen. Age-related prevalence of breast cancers by ER, PR and HER-2. Poster presented at the 10th Biennial Conference on Primary Therapy of Early Breast Cancer, St Gallen,

Switzerland, Mar. 2007.

H. Wildiers, B. Van Calster, W. Hendrickx, S. Van Huffel, A. Smeets, R. Paridaens, D. Timmerman, M. Drijkoningen, K. Leunen, C. Weltens, M.-R. Christiaens, P. Neven. Axillary lymph node involvement in breast cancer decreases with age but increases again after age 70 mainly in smaller tumors. Poster presented at the 10th Biennial Conference on Primary Therapy of Early Breast Cancer, St Gallen, Switzerland, Mar. 2007.

E. D'Hondt, G. Moulin-Romsee, F. Persyn, B. Van Calster, P. Neven, S. Stroobants. Detection of sentinel lymph nodes in breast cancer: every day practice and pitfalls. Poster presented at the 13th Symposium of the Belgian Society of Nuclear Medicine, Spa, Belgium, May 2007.

B. Van Calster, W. Hendrickx, P. Neven, H. Wildiers, D. Timmerman, S. Van Huffel. Piecewise logistic regression models can be a practical solution to nonlinearity in the logit: examples on breast cancer. Poster presented at the 28th Annual Conference of the International Society for Clinical Biostatistics, Alexandroupolis, Greece, Jul. 2007.

E. Kirk, G. Condous, B. Van Calster, S. Van Huffel, D. Timmerman, T. Bourne. Rationalizing the follow-up of pregnancies of unknown location. Presented at the 31st British International Conference on Obstetrics and Gynaecology, London, UK, Jul. 2007.

B. Van Calster. Does CA 125 help us to assess the risk of malignancy? Invited talk at the 17th World Congress on Ultrasound in Obstetrics and Gynecology, Florence, Italy, Oct. 2007.

B. Van den Bergh, B. Van Calster, S. Pinna Puissant, S. Van Huffel. Adrenocortical activity is related to physical complaints, behavioral and emotional disorders. Poster presented at the 4th International Conference on the (Non)Expression of Emotions in Health and Disease, Tilburg, the Netherlands, Oct. 2007.

F. Amant, A. De Knijf, B. Van Calster, K. Leunen, P. Neven, P. Berteloot, I. Vergote, S. Van Huffel, P. Moerman. In contrast to surgical castration and lymphadenectomy, adjuvant hormonal treatment should be considered in the treatment for endometrial stromal sarcoma. Poster presented at the 15th International Meeting of the European Society of Gynaecological Oncology, Berlin, Germany, Oct. 2007.

C. Van Holsbeke, B. Van Calster, L. Valentin, A.C. Testa, E. Ferrazzi, I. Dimou, C. Lu, P. Moerman, S. Van Huffel, I. Vergote, D. Timmerman. External validation of mathematical models to distinguish between benign and malignant adnexal tumors.

Poster presented at the 15th International Meeting of the European Society of Gynaecological Oncology, Berlin, Germany, Oct. 2007.

E. D'Hondt, G. Moulin-Romsee, F. Persyn, B. Van Calster, P. Neven, S. Stroobants. Detection of sentinel lymph nodes in breast cancer: everyday practice and pitfalls. Presented at the 15th International Meeting of the European Society of Gynaecological Oncology, Berlin, Germany, Oct. 2007.

B. Van Calster, D. Timmerman, L. Valentin, A. Testa, C. Van Holsbeke, S. Van Huffel. Pre-operative prediction of ovarian tumours as benign, primary invasive, borderline, or metastatic. Presented at the 5th LKI Oncoforum, Leuven, Belgium, Feb. 2008.

V. Van Belle, B. Van Calster, S. Van Huffel, O. Brouckaert, I. Van den Bempt, S. Pintens, P. Moerman, R. Paridaens, M.-R. Christiaens, P. Neven. Improved NPI for breast cancer prognosis. Poster presented at the 5th LKI Oncoforum, Leuven, Belgium, Feb. 2008.

Abstracts submitted to international conferences

V. Van Belle, B. Van Calster, S. Van Huffel, O. Brouckaert, I. Van den Bempt, S. Pintens, R. Paridaens, M.-R. Christiaens, P. Neven. Improved NPI for breast cancer prognosis. Abstract submitted to the 6th European Breast Cancer Conference, 2008.

H. Wildiers, B. Van Calster, L.V. van de Poll-Franse, W. Hendrickx, S. Van Huffel, A. Smeets, R. Paridaens, K. Deraedt, M.-R. Christiaens, P. Neven. External validation of a piecewise effect model of axillary lymph node involvement in elderly breast cancer patients. Abstract submitted to the 6th European Breast Cancer Conference, 2008.

E. Kirk, A.T. Papageorghiou, B. Van Calster, G. Condous, C. Bottomley, D. Timmerman, K. Spencer, T. Bourne. Can serial measurements of inhibin A and activin A be used to predict the outcome of pregnancies of unknown location? Abstract submitted to the 24th Annual Meeting of the European Society Human Reproduction and Embryology, 2008.

Curriculum Vitae

Ben Van Calster was born in Leuven on December 13, 1977. In 1997 he received the Bachelor's degree ('Kandidaat'), and in 2000 the Master's degree ('Licentiaat') in Psychology at the Katholieke Universiteit Leuven. His main topic was developmental psychology, and his Master's thesis dealt with the differential impact of extinction on expectancy learning and evaluative learning. He fulfilled the first part of his research training at the Center for Developmental Psychology of the Katholieke Universiteit Leuven, where he worked on the factorial structure of the emotional autonomy scale under supervision of professor Luc Goossens and Wim Beyers [1, 2]. The second part was fulfilled at the Research Center for Health Promotion of the University of Bergen (Norway) under the supervision of professor Oddrun Samdal, with research on the school environment as a risk or resource for students' well-being.

In 2003, he received the M.Sc. degree in Statistics at the Katholieke Universiteit Leuven, with a thesis on mathematical models for predicting the evolution of a pregnancy of unknown location under the supervision of professors Sabine Van Huffel and Dirk Timmerman.

Between September 2000 and February 2001, he worked as a teaching assistant for the first introductory course on statistics for psychologists. From February 2001 until June 2003, he worked as a research and teaching assistant at the Laboratory for Experimental Social Psychology. His research, under the guidance of professor Vera Hoorens, focused on trait transference [3]. Then, before starting the present PhD work, he was a biostatistician at Medisearch International (now SGS-Medisearch), a full service contract research organization for the execution and analysis of clinical trials.

He is married to Ilse De Croock, with whom he has a daughter Maren (March 3, 2006).

Bibliography:

[1] W. Beyers, L. Goossens, B. Van Calster, and B. Duriez. An alternative substantive factor structure of the Emotional Autonomy Scale. *European Journal of Psychological Measurement*, 21:147-155, 2005.

[2] W. Beyers, L. Goossens, and B. Van Calster. Emotional autonomy in adolescence: a multidimensional construct. *Paper presented at the 7th Biennial Conference of the European Association for Research on Adolescence*, Jena (Germany), 2000.

[3] B. Van Calster and V. Hoorens. Trait transference and the logic of conversation. *Poster presented at the 13th Annual Meeting of the European Association for Experimental Social Psychology*, San Sebastian (Spain), 2002.