

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

LEAST SQUARES SUPPORT VECTOR MACHINES CLASSIFICATION APPLIED TO BRAIN TUMOUR RECOGNITION USING MAGNETIC RESONANCE SPECTROSCOPY

Promotors:
Prof. dr. ir. J. Vandewalle
Prof. dr. ir. S. Van Huffel

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Lukas

December 2003

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

LEAST SQUARES SUPPORT VECTOR MACHINES CLASSIFICATION APPLIED TO BRAIN TUMOUR RECOGNITION USING MAGNETIC RESONANCE SPECTROSCOPY

Jury:
Prof. dr. ir. J. Berlamont, voorzitter
Prof. dr. ir. J. Vandewalle, promotor
Prof. dr. ir. S. Van Huffel, promotor
Prof. dr. ir. A. Barbé
Prof. dr. L.M.C. Buydens (KU Nijmegen)
Prof. dr. ir. J.A.K. Suykens
Prof. dr. ir. S. Vandewalle
Prof. dr. ir. M. Van Hulle (Faculteit Geneeskunde)

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Lukas

U.D.C. 681.3*I4:616.831-006 December 2003

© Katholieke Universiteit Leuven – Faculteit Toegepaste Wetenschappen
Arenbergkasteel, B-3001 Heverlee (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/2003/7515/71

ISBN 90-5682-460-0

Acknowledgement

I consider myself very fortunate to have had the opportunity to study and live in Leuven. With its ancient beauty, the city hosts an excellent academic environment. It is indeed one of the best learning places, not to mention meeting charming peoples. During my study years, many individuals and organisations eventually contributed to this thesis, and I would like to send out my thanks to all of them. I do apologize for any unintended omissions.

First and foremost, I would like to express my gratitude to Prof. Joos Vandewalle for giving me the opportunity to start my Ph.D. under his guidance. He always puts a confidence in me to reach this long-awaited dream. His great experience, patience and understanding have continuously encouraged me.

Prof. Sabine Van Huffel, for giving me not only opportunity to work within the biomedical engineering application domain, but valuable research suggestions as well. Her unique personality in guiding her students shown genuine care and attention towards each of us, and are always encouraging and relieving.

Prof. Johan Suykens who welcomed and brought me in the journey of developing kernel machines methods. He has always been my inspiring daily supervisor from the very beginning. His eagerness to share his knowledge without reserve gives me the best example of the important "support values" in real world.

I would also like to thank Prof. André Barbé, Prof. Stefan Vandewalle, for their time and expertise in reading and giving valuables comments on the earlier versions of this thesis. I thank Prof. Lutgarde M.C. Buydens and Prof. Marc Van Hulle for agreeing to become members of the jury. Prof. Jean Berlamont is gratefully acknowledged for chairing the jury.

Andy Devos and Dr. Leentje Vanhamme are my closest collaborators in this thesis. This work would not have been possible without their presence and working with them during these years was a great pleasure.

This work has been realized within the framework of the EU-funded research project INTERPRET. In particular I would like to thank Prof. C. Arús, who coordinated the project, and his partners Dr. A.R. Tate, Dr. F.A. Howe, Dr. C. Majós, Dr. A. Moreno-Torres, Dr. M. van der Graaf for their support and

intensive discussions on the experiments and the joint publications of brain tumor classification using short and long echo time MRS spectra.

I also thank the Chemometrics Group from the department of Analytical Chemistry at the Katholieke Universiteit Nijmegen, headed by Prof. Lutgarde M.C. Buydens, and the MRS group of the department of Radiology St. Radboud hospital Nijmegen, headed by Prof. Arend Heerschap. Special thanks go to Arjan W. Simonetti, Uwe Thissen and Dr. M. van der Graaf for our research collaborations in brain tumour classification using MRI and MRSI data.

A hearty is sent out to all members of SCD/SISTA research group for generating a superb research atmosphere. I thank Lu Chuan, Lieveke Ameye, Andy Devos, and my previous roommates, Mustak Yalçin, Jos De Brabanter and Fritz von Haeseler for their daily charms in sharing our office space. I greatly appreciate the efforts of Ida Tassens, Bart Motmans, and Ilse Pardon, for their help with the administrative matters.

The family Handaja, Fr. Kees Bertens, Sr. Luciana, Lena, Gigis, Koen-Ili and Jenni Handayani have been my personal supports. Distance and time have never separated us, they might have heard me more in difficult times rather than in happy ones. Their supports and prayers have kept me on track in pursuing my dream.

I am deeply grateful to Joske Vettenburg from the housing service, for assisting me in finding convenient accommodation for handicapped student. Peter Merkx, Hans Wennink and Lanny Suhendra from the Stichting Porticus in the Netherland kindly provided me with a three-wheeled vehicle for my better mobility.

My sincere appreciation to the Universitas Katolik Atma Jaya, Jakarta, and the Katholieke Universiteit Leuven for giving me the opportunity to pursue my study. I gratefully acknowledge the K.U. Leuven and the F.W.O. for granting financial support during my master and doctorate study.

I also wish to thank all my friends in Leuven for the help, support, and for having share many experiences. In particular, I would like to thank Andy-Inneke-Peter-Thomas, Hartanto-Bonnie-Janice, Andre-Sari, Eko-Santhi-Jujud-Tirta, Hadeli, and Andang for opening their hearts and hands.

During my stay in Leuven I also enjoyed the friendship atmosphere in Sint-Janshuis, Van Dalecollege and Romerohuis I, as well as in ILT and CLT language classes.

I would not be standing here before you all without the love, care and continuous support from my family: Papa-Mama, Edy-HongLan-Rani, Lusia-Yopi-Agung, and Bowo-Lenny-Esa. Nothing can sufficiently speak my word of thanks to them all, especially to my dearest wife, Tanti Johana. *Terima kasih, Nana.*

Abstract

Magnetic Resonance Spectroscopy (MRS) is a technique which has evolved rapidly over the past 15 years. It has been used specifically in the context of brain tumours and has shown very encouraging correlations between brain tumour type and spectral pattern. *In vivo* MRS enables the quantification of metabolite concentrations non-invasively, thereby avoiding serious risks to brain damage.

While Magnetic Resonance Imaging (MRI) is commonly used for identifying the location and size of brain tumours, MRS complements it with the potential to provide detailed chemical information about metabolites present in the brain tissue and enable an early detection of abnormality.

However, the introduction of MRS in clinical medicine has been difficult due to problems associated with the acquisition of *in vivo* MRS signals from living tissues at low magnetic fields acceptable for patients. The low signal-to-noise ratio makes accurate analysis of the metabolites represented by the spectra very difficult. Furthermore, for clinicians, it is time-consuming to analyze and interpret the MR spectra. This task requires considerable experience from an expert spectroscopist. Automated processing, analysis and interpretation of the MRS spectra is therefore highly valuable.

The high dimensionality of the spectra, the presence of noise and artefacts, and the low amount of data of specific pathologies (e.g. specific brain tumour types) available complicate the classification.

In this thesis, advanced classification techniques based on Least Squares Support Vector Machines (LS-SVM) are developed and applied to brain tumour classification. LS-SVM classifiers using linear as well as nonlinear Radial Basis Function (RBF) kernels are compared with the classical techniques.

Given the problem of classifying four major types of brain tumours: *glioblastomas, meningiomas, metastases and astrocytomas*, LS-SVM classifiers perform very well regarding discrimination and prediction of these tumour types, except for the discrimination of glioblastomas vs. metastases, due to the high similarities among the available training spectra.

Dimensionality reduction, by selecting frequency regions or peak integration could possibly reduce the influence of disturbing noise and artefacts in the spectra. Using some prior knowledge for selecting resonance peaks allows to reach a similar performance as using complete spectra. Analogously, a rough quantification method using peak integration yields a relatively high classification performance.

Without the need of prior knowledge and/or dimensionality reduction, LS-SVM classifiers with linear or nonlinear kernels can be applied directly, and automatically give a reliable result. A good generalization performance could be expected from the maximal margin separation between the classes.

Glossary

This section lists symbols and acronyms that occur frequently in this thesis.

Remark. The notation used in this thesis allows one to distinguish among scalars, vectors and matrices. Lower case characters represent scalar values. Boldface lower case characters are used for vectors. Boldface capitals represent matrices.

Mathematical Notation

a	scalar a
$\mathbf{v}$	vector $\mathbf{v}$
$\mathbf{M}$	matrix $\mathbf{M}$
$\mathbf{M}^T$	transpose of matrix $\mathbf{M}$
$\mathbf{M}^{-1}$	inverse of matrix $\mathbf{M}$
$\det \mathbf{M}$	determinant of matrix $\mathbf{M}$
$\|\mathbf{M}\|$	Frobenius norm of matrix $\mathbf{M}$
$\mathbf{v}(m)$	m–th element of vector $\mathbf{v}$
$\mathbf{M}(m,n)$	element on the m–th row and n–th column of matrix $\mathbf{M}$
$\mathbb{N}$	the set of natural numbers
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^n$	set of real n–dimensional vectors
$\hat{x}$	estimate of x
$\|x\|$	absolute value of x
$\|\|\mathbf{x}\|\|_2$	2–norm of $\mathbf{x}$
σ_x^2	variance of x
$a \ll b$	a is much smaller than b
$a \approx b$	a is approximately equal to b

Fixed Symbols

f	frequency
e	error
$\mathbf{0}$	zero vector
$\mathbf{0}_n$	$N \times N$ zero matrix
$\mathbf{I}_n$	$n \times n$ identity matrix
j	$\sqrt{-1}$

Acronyms and Abbreviations

AMARES	Advanced Method for Accurate, Robust, and Efficient Spectral fitting
AUC	area under ROC curve
CDVC	clinical data validation committee
Cho	choline
CNS	Central Nervous System
Cr	creatine
CSF	cerebrospinal fluid
CSI	Chemical Shift Imaging
FDA	Fisher discriminant analysis
FFT	Fast Fourier Transform
FID	Free Induction Decay
FOV	field of view
GD	Gadolinium enhanced
Glc	glucose
Glu	glutamate
Gln	glutamine
Glx	glutamine and glutamate
Gly	glycine
HLSVD	Hankel Lanczos Singular Value Decomposition
HSVD	Hankel Singular Value Decomposition
INTERPRET	International Network for Pattern Recognition of Tumours using Magnetic Resonance (IST-1999-10310)
Lac	lactate
LDA	Linear Discriminant Analysis
Lin	linear
LS	least squares
LS-SVM	Least Squares Support Vector Machine
mI	myo-inositol
MR	magnetic resonance
MRI	magnetic resonance imaging

MRS	magnetic resonance spectroscopy
MRSI	magnetic resonance spectroscopy imaging
NAA	N-acetyl-aspartate
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PCr	phosphocreatine
PD	proton density
PRESS	point resolved spectroscopy
QP	Quadratic Programming
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
SNR	signal–to–noise ratio
STEAM	stimulated-echo acquisition mode
SVD	Singular Value Decomposition
SVM	Support Vector Machine
T	Tesla
TE	echo time
TLS	total least squares
TR	repetition time
T_1	spin-lattice relaxation time
T_2	spin-spin relaxation time
VOI	volume of interest
WHO	World Health Organisation

Medical and Magnetic Resonance Terminologies

This section is compiled from [8, 32, 37, 70, 76]

- **Anaplastic**: cancerous cells that divide rapidly and bear little or no resemblance to normal cells.
- **Angiography**: x-ray examination of blood vessels.
- **Astrocytoma**: a type of brain tumor that begins in the brain or spinal chord in small, star-shaped cells called astrocytes.
- **Benign**: a tumour that is not cancerous
- **Biopsy**: removal of a small piece of living tissue from a part of the body for microscopic examination.
- **Brain herniation**: the protrusion of the brain through the foramen magnum, or the opening at the base of the skull. Critical brain tissue is compressed as this occurs, which is often fatal.

- **Burr hole**: a 1 cm-diameter hole cut in the skull using a drill.
- **Cancer**: a term for diseases in which abnormal cells divide without control.
- **Chemotherapy**: the prevention or treatment of a tumour by the use of chemical substances.
- **Craniotomy**: surgical opening of a portion of the skull performed to expose a lesion of the brain.
- **CT (Computerised Tomography) scan**: a computer aided x-ray used to provide a picture of the inside of the body.
- **Edema**: an accumulation of an excessive amount of watery fluid in cells, tissues or serous cavities.
- **EEG (Electroencephalography)**: a technique for recording the electrical activity of the brain.
- **Epilepsy**: disorder of brain function usually characterised by recurrent attacks of unconsciousness (seizures).
- **Gamma rays**: electromagnetic radiation with a wave length shorter than x-rays which are given off by certain radioactive substances.
- **Glial tissue**: the special connective tissue of the central nervous system.
- **Glioblastoma**: a general term that refers to malignant astrocytoma, a type of brain tumor.
- **Glioblastoma multiforme (GBM)**: a type of brain tumor that forms from glial (supportive) tissue of the brain. It grows very quickly and has cells that look very different from normal cells. Also called grade IV astrocytoma.
- **Glioma**: a cancer of the brain that begins in glial cells (cells that surround and support nerve cells).
- **Haemorrhage**: the escape of blood from a ruptured blood vessel, externally or internally.
- **High grade tumour**: a tumour that grows quickly over a period of a few months.
- **Histology**: the study of tissues and cells under a microscope.
- **Histopathology**: the study of the microscopic structure, composition and function of diseased tissue

- **Inflammation**: the reaction of living tissues to injury, infection or irritation.
- **Lesion**: area of tissue with impaired function as a result of damage by disease or wounding.
- **Low grade tumour**: a tumour that develops slowly over a number of years.
- **Malignant tumour**: a tumour that invades and destroys the tissue where it originates and which can spread to other sites in the body.
- **Membrane**: a thin layer of tissue surrounding an organ, lining a cavity or separating adjacent structures or cavities.
- **Meninges**: the three connective tissue membranes that line the skull and vertebral canal and enclose the brain and spinal cord.
- **Meningioma**: a type of tumor that occurs in the meninges, the membranes that cover and protect the brain and spinal cord. Meningiomas usually grow slowly.
- **Metabolism**: the chemical changes in living cells by which energy is provided for vital processes and activities and new material is assimilated.
- **Metabolite**: a substance essential to the metabolism of a particular organism.
- **Metastasis**: the spread of cancer from one part of the body to another. A tumor formed from cells that have spread is called a secondary tumor, a metastatic tumor, or a metastasis. The secondary tumor contains cells that are like those in the original (primary) tumor. The plural form of metastasis is metastases
- **Mitosis**: the type of cell division that makes new body cells.
- **MRI (Magnetic Resonance Imaging)**: a procedure which shows a picture, in any plane, of the inside of the body. This procedure is noninvasive as it uses magnetic waves.
- **MRI/CT contrast**: a dye injected into the veins to improve MRI and CT scan pictures.
- **Nausea**: a feeling of sickness associated with the urge to vomit.
- **Necrosis**: pathologic death of cells, tissue, or an organ from irreversible damage.
- **Nerve cell**: the basic functional unit of the nervous system. These specialized cells transmit chemicals and electrical impulses and so carry information from one part of the body to another.

- **Nervous system**: the vast network of nerve cells which carries information to and from all parts of the body in order to bring about bodily activity. It is classically divided into the brain, spine and peripheral nervous system.
- **Neurological**: refers to conditions occuring in the nervous system, including the brain, spine and all the peripheral nerves.
- **Neurologist**: a doctor who specialises in all conditions affecting the nervous system.
- **Neuron**: another term for nerve cells.
- **Neurosurgeon**: a doctor who specialises in the surgical treatment of disorders of the nervous system.
- **Oncologist**: a specialist in the treatment of tumours.
- **Pathologist**: a doctor who specialises in the study of disease, its causes, mechanisms and effects on the body.
- **Primary tumour**: tumour that only occurs in the organ in which it started.
- **Prognosis**: the assessment of the future course and outcome of a patient's condition.
- **Quantitation**: the mathematical processing performed on the NRM signal to obtain numerical values for the quantities which characterize each of the components of a NMR spectroscopy signal.
- **Radiotherapy**: a therapeutic procedure in which conditions are treated with energy in the form of waves or particles.
- **Scan**: the examination of the body or part of the body using CT or MRI.
- **Secondary tumour**: the spread of cancer from one part of the body to another.
- **Shunt**: a device which is inserted to remove fluid from the brain or a tumour.
- **Stereotactic biopsy**: a surgical procedure guided by scans and a frame.
- **Stereotactic local irradiation**: radiotherapy directed at tissue by finely targeted beams using scans and a frame.
- **Tumour**: any abnormal swelling in or on any part of the body. This term is usually used when the swelling is as a result of an overgrowth of cells.

Contents

Acknowledgement **i**

Abstract **iii**

Glossary **v**

Contents **xi**

1 Introduction **1**

1.1 Magnetic Resonance Spectroscopy 1

1.2 MRS for brain tumour evaluation 2

1.3 Brain tumour classification 4

1.4 Support Vector Machine classifiers 5

1.5 Brain tumour datasets 5

1.6 Challenges 6

1.7 Goals of the thesis 7

1.8 Outline of the thesis 7

2 Pattern Recognition Systems **11**

2.1 Pattern Recognition 11

2.1.1 Patterns 11

2.1.2 Features 11
2.1.3 Feature selection 12
2.1.4 Pattern recognition 12
2.1.5 Typical pattern recognition systems 12
2.2 Performance Measures 14
2.2.1 Correct classification rate 14
2.2.2 Receiver Operating Characteristic 15
2.2.3 Statistical analysis 17
2.3 Conclusions 19

3 Linear Classification **21**
3.1 Binary classification problems 21
3.2 Linear classifiers 22
3.3 Linear Discriminant Analysis (LDA) 26
3.4 Principal Component Analysis (PCA) 27
3.5 Conclusions 28

4 Kernel-based Classification **29**
4.1 Introduction 29
4.2 Maximal margin classifiers 29
4.3 Construction of the optimal hyperplane 32
4.4 Optimal hyperplane for linearly nonseparable case 34
4.5 Kernel-based classifiers 36
4.6 Nonlinear SVM classifiers 37
4.7 Least Squares SVM classifiers 41
4.8 Conclusions 43

5 Magnetic Resonance Spectroscopy **45**

5.1 Introduction 45
5.2 Brain tumours 45
5.3 Brain tumour diagnosis 47
5.4 Brain tumour treatment 48
5.5 Nuclear Magnetic Resonance (NMR) 49
5.6 Magnetic Resonance Imaging (MRI) 50
5.7 Magnetic Resonance Spectroscopy (MRS) 51
5.7.1 Basic principles 51
5.7.2 Relaxation 53
5.7.3 Sequences 53
5.7.4 Mathematical model of MRS signal 54
5.8 The role of MRS in brain tumour diagnosis 55
5.9 Conclusions 56

6 Classification of brain tumours using long echo time MRS data 59

6.1 Introduction 59
6.2 Material 61
6.3 Methods 64
6.3.1 Selected frequency regions 64
6.3.2 Peak integration 65
6.3.3 Training and test data 65
6.3.4 Statistical analysis 67
6.4 Results 68
6.4.1 Classification using complete spectra 68
6.4.2 Classification using selected frequency regions 68
6.4.3 Classification using peak integration 72
6.4.4 Multiclass approach 72

6.5 Discussion 81
6.5.1 Limitations 81
6.5.2 Glioblastomas versus metastases 81
6.5.3 Classification techniques 84
6.5.4 Influence of dimensionality reduction 85
6.5.5 Multiclass classification 86
6.6 Conclusions 86

7 Classification of brain tumours using short echo time MRS data 89

7.1 Introduction 89
7.2 Material 90
7.3 Methods 92
7.3.1 Classification techniques 92
7.3.2 Input features 93
7.3.3 Dimensionality reduction by prior knowledge 95
7.3.4 Experimental setting 95
7.3.5 Multiclass classification 97
7.3.6 Statistical analysis 97
7.4 Results 97
7.4.1 Binary classification 97
7.4.2 Multiclass approach 100
7.5 Discussion 102
7.5.1 Classification techniques 102
7.5.2 Influence of input features 102
7.5.3 Influence of dimensionality reduction 103
7.5.4 Multiclass classification 104

7.6 Conclusions 104

8 Brain tumour classification using MRSI and MRI data 109

8.1 Introduction 109

8.2 Material 111

8.3 Methods 112

8.3.1 Input features 112

8.3.2 Classification techniques 113

8.3.3 Experimental setting 114

8.4 Results 115

8.4.1 Healthy vs. tumour tissue 115

8.4.2 Normal tissue of brain tumour patients vs. normal tissue of volunteers 116

8.4.3 Benign vs. malignant tumours 116

8.4.4 Low grade astrocytomas vs. high grade astrocytomas . . 116

8.4.5 Astrocytic tumours vs. meningiomas 117

8.4.6 Grade II vs. grade III astrocytic tumours 118

8.5 Discussion 119

8.5.1 Classification techniques 119

8.5.2 MRSI versus MRI 119

8.5.3 Several classification problems 120

8.6 Conclusions 121

9 Conclusions 131

9.1 Summary 131

9.2 Open problems 133

Appendix 135

A Articles in international journals 135
B Articles in conference proceedings 135
C Abstracts for conferences 136

Bibliography **137**

Biography **147**

Chapter 1

Introduction

1.1 Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance (NMR) constitutes an important breakthrough in medicine. This has been internationally recognized in 2003 by awarding the Nobel prize for Medicine to two pioneers in the field of Magnetic Resonance Imaging (MRI), Paul Lauterbur and Peter Mansfield. MRI allows the visualization of the inner body without inserting any obstructive instrument or carrying out an invasive surgery to the patient. It has been used in addition to other medical imaging techniques, such as ultrasound and x-rays. Before the introduction of MRI in the clinical arena in 1982, the only way to get a 3-D representation of the human body was by using Computerized Axial Tomography, also known as CAT or CT scans. Although CT works well in certain contexts, it has limitations. It exposes patients to radiation and only shows the body on its axial (top to bottom) plane. In contrast, MRI does not rely on the absorption of x-rays, but is based on the magnetic resonance technique.

While MRI [58] is able to visualize anatomical structure, Magnetic Resonance Spectroscopy (MRS) helps the investigation of physiological function, providing a snapshot of metabolism at the molecular level.

MRI uses strong magnetic fields to generate an energy exchange between external magnetic fields and protons within the tissues of the patient. This energy exchange is then emitted back to the machine as a radiofrequency signal which is decoded by sophisticated mathematical software. Using conventional MRI techniques the software displays gray-scale anatomical images based on the spatial distribution and composition of water protons.

MRS utilizes the principle that nuclei in different chemical structures have different characteristic resonance patterns, depending on the neighboring chemical structures and the interactions between these structures. MRS detects the chemical composition of the tissue being studied, and after some calculation, it displays a waveform with peaks corresponding to the various chemicals detected. MRS relies on a phenomenon called the *chemical shift*. Different chemical components will appear as peaks at different locations along a horizontal axis with shift or displacement from the zero point measured in parts per million (ppm). The peaks in the NMR spectrum represent different chemical components of the tissue, and the waveform reflects the quantities of various components. Substances in higher concentrations in the tissue will be shown with a higher peak than substances present in lower concentrations.

In the last five decades, MRS has been increasingly used as an important analytical tool in organic chemistry, biology and material science. Spectra can be obtained *in vitro* to elucidate the chemical structures of molecules. Recent advances also enable the use of MRS for studying living tissue. *In vivo* MRS measurement allows detailed biochemical information to be obtained non-invasively from specific tissue slices of the patients.

Although MRS is widely used as a research tool and has great potential as a clinical aid, the development of MRS in clinical medicine has been slow. This is due to difficulties associated with acquiring *in vivo* MRS signals from living tissues at the low magnetic fields acceptable for human patients. The low signal-to-noise ratio (SNR) makes accurate quantification of the metabolites represented by the spectra very difficult [18].

Furthermore, it is very difficult and time-consuming to analyze and interpret the spectra. This task requires considerable experience from an expert spectroscopist. Recently acquisition techniques have been improved significantly. Unfortunately not enough attention was paid to the automatic analysis and interpretation of those data. In order to process these large amounts of data the analysis and interpretation of the MRS spectra needs to be automated.

In order to support the application of MRS into clinical practice this thesis aims at exploring and developing some automatic classification techniques to enhance analysis and interpretation of MRS spectra.

1.2 MRS for brain tumour evaluation

Yearly, about 29,000 persons in Europe are diagnosed with a brain tumour (1998 estimates) [12], a rate of 7 per 100,000 people. Accurate diagnosis of these tumours is indispensable. Prognosis of tumour development and life expectation largely depends on the tumour type and grade. Additionally, different

types of tumours respond differently to treatment. For example, some respond better to radiotherapy than others.

The most commonly used MRS technique for diagnosing brain tumours is proton MRS (^{1}H MRS). The hydrogen nucleus consists of a single proton and is therefore its use in MRS commonly referred to as proton MRS. However, MRS can also be performed using other nuclei such as carbon-13 (^{13}C), which has been used to study glycogen metabolism and the pathogenesis of diabetes, or phosphorous-31 (^{31}P) which allows the study of bioenergetics and lipid metabolism.

MRS can be performed with either a single- or multi-voxel technique. A single-voxel technique acquires a signal from a particular volume element, also called voxel, while a multi-voxel technique acquires signals from an array or matrix of voxels of the localized area (one signal per voxel). Information from a multi-voxel matrix can be visualized by means of MRS Imaging (MRSI).

The parameters of the MRS signal yield information about molecules of the examined subject: the MRS spectral frequency characterizes the molecular identity, the damping describes the mobility and the spectral amplitude (intensity) relates to the concentration of the corresponding molecules.

There are several factors which make the MRS signal processing difficult, e.g. signal degradation, related to the sensitivity of the technique, technical limitations when measuring living organisms, as well as time limitation during measurement. In general, *in vivo* MRS signals are characterized by a low signal-to-noise ratio, strongly overlapping spectral components, and the presence of disturbing components (e.g. the residual water peak in ^{1}H MRS).

Different types of signals can be measured by applying different measurement sequences. In the clinical MRS examination of the human brain, it is important to distinguish two sequences: long and short echo time ^{1}H-MRS. Many studies have been devoted to the analysis of these long and short echo time MRS signals. Long echo time MRS signals contain less components and are easier to process. However, they also provide limited information and are therefore restrictive in diagnosing some patient diseases. Hence, there is a growing interest for the processing of short echo time MRS. More metabolite information is contained in short echo time MRS signals, however they are also much more difficult to process.

A reliable, accurate and robust processing of the MRS signals is very important for a trustworthy diagnosis. Indeed, MRS signal processing essentially contributes to the distinction of tumour type and its grade. Non-invasive methods, such as MRS, will therefore provide a significant alternative for the traditional invasive methods, e.g. brain biopsy, which poses serious risks to a patient.

Reproducibility and accuracy of acquired MRS data are crucial for an objective and reliable evaluation of MRS signals. Advances in research of pattern recognition and classification make an automated MRS signal processing viable. These advances are exploited in this thesis, thereby improving brain tumour recognition based on spectral features extracted from the raw MRS spectra.



Figure 1.1: *Brain scanning using Magnetic Resonance (MR) technology.*

1.3 Brain tumour classification

Most current methods for spectral analysis [3, 71, 85, 95, 117, 119] aim at explicit quantification of the metabolites from each MR spectrum of the brain. Quantified peaks depict the biochemical structure of the tissue being examined. However, explicit quantification of metabolites is not an easy task, due to the overlap between peaks, as well as the low SNR and artefacts in the spectra.

Two simple methods can be applied as alternative to metabolite quantification. In the first method a trapezoidal rule [41] is applied to the peaks within the known metabolite ranges. This method gives a rough estimate of the area under the peak, but ignores the tails of the spectra. In the second method several frequency ranges are selected covering the known metabolite peaks of interest. In this approach overlapping peaks might just be regarded as one single lump, instead of separated peaks.

Normally the numbers of required data are at least 3-6 times larger than the number of variables. This amount of data is not easily gathered in medical applications, especially in the diagnosis of brain tumours where acquisition of data is quite sophisticated and different protocols are used. Therefore it is very hard to perform a classification using traditional methods without first reducing the dimensionality of the MRS spectra. This problem of classifying high dimensional data motivates the application of Support Vector Machines (SVM).

1.4 Support Vector Machine classifiers

The Support Vector Machine algorithm was first developed in 1963 by Vapnik and Lerner [121] and Vapnik and Chervonenkis [122] as an extension of the Generalized Portrait algorithm. This algorithm is firmly grounded in the framework of statistical learning theory – Vapnik Chervonenkis (VC) theory, which improves the generalization ability of learning machines to unseen data [97]. The Support Vector Machine stirred up attention only in 1995 with the appearance of Vapnik's book *The Nature of Statistical Learning Theory* [123]. In the last few years Support Vector Machines have shown excellent performance in many real-world applications such as optical character recognition, text categorization, time series prediction or biological sequence analysis.

SVM is a nonlinear classification algorithm based on kernel methods. In contrast to linear classification methods, the kernel methods map the original parameter vectors into a higher (possibly infinite) dimensional feature space through a nonlinear kernel function. Without the need to compute the nonlinear mapping explicitly, dot-products can be computed efficiently in higher dimensional space. The dominant feature which makes SVM very attractive is that classes which are nonlinearly separable in the original space can be linearly separated in the higher dimensional feature space. Thus SVM is capable to solve complex nonlinear classification problems.

Important characteristics of SVM are its ability to solve classification problems by means of convex quadratic programming (QP), and also the sparseness resulting from this QP problem. A least squares version of SVM (LS-SVM) is introduced by Suykens in 1999 [101] with the idea of modifying Vapnik's SVM formulation by adding a least squares term in the cost function. This variant circumvents the need to solve a more difficult QP problem and only requires the solution of a set of linear equations. This approach significantly reduces the complexity and computation in solving the problem.

1.5 Brain tumour datasets

The long echo time and short echo time ^{1}H MRS dataset used in Chapter 6 and Chapter 7 of this thesis have been provided within the framework of the EU funded INTERPRET project (IST-1999-10310) [45]. These datasets were developed by 6 centres : CDP (Centre Diagnòstic Pedralbes), Barcelona, Spain; FLENI (Fundación para la Lucha contra las Enfermedades Neurológicas de la Infancia), Buenos Aires, Argentina; IDI (Institut de Diagnòstic per la Imatge), Barcelona, Spain; MUL (Uniwersytet Medyczny w Łodzi), Łodzi, Poland; SGHMS (St George's Hospital Medical School), London, United Kingdom and UMCN (Universitair Medisch Centrum Nijmegen), Nijmegen, The

Netherlands. All MRS data have passed a quality control and validation procedure, which was regulated by strict rules agreed on by all INTERPRET partners [45].

The MRI and MRSI datasets used in Chapter 8 were provided by the acquisition centre UMCN (Universitair Medisch Centrum Nijmegen), Nijmegen, The Netherlands.

1.6 Challenges

The classification of brain tumours based on MRS spectra faces several problems:

- **High dimensional input space**
  Depending on the acquisition resolution, complete MRS spectra consist of hundreds to thousands of spectral points. While the higher dimension of this input potentially makes MRS a unique and powerful technique for clinical medicine, on the other hand it complicates the computation and the design of an appropriate method to handle it.
- **Few training data**
  To achieve a sufficiently accurate classification performance, most learning algorithms require a substantial amount of training data which scales with the number of inputs. However, in real life, this requires a huge amount of measurement time before this large amount of data becomes available. Statistical techniques such as bootstrap and cross-validation have been developed to permit the use of these limited data sets and still give an empirical measure of the classifier performance and its potential generalization.
- **Noise and artefacts**
  Since *in vivo* MRS is here applied to living organism, the quality of the signals is relatively poor (low SNR, the presence of artefacts). Sophisticated methods might be applied to filter out noise and remove artefacts. However, it is expected that the classifier should be robust against these imperfections.
- **Computational efficiency**
  High dimensional input spaces can be computationally difficult and time consuming for classifiers, e.g. weight adjustment of Artificial Neural Network (ANN). It is often required that the input dimension needs to be reduced. It is desired that with the limited resources (computer memory, computer speed, etc.) a classifier can solve the computation as fast as possible.

1.7 Goals of the thesis

In this thesis, advanced classification techniques based on Least Squares Support Vector Machines (LS-SVM) are developed and applied to a particular biomedical problem, namely the classification of brain tumours using magnetic resonance spectroscopy signals. The work presented here aims to assess the performance of kernel-based methods in comparison with classical methods within this biomedical application.

This thesis also shows how the classification techniques handle the aforementioned challenges inherent in brain tumour classification problems. Although the methods and results presented here are for brain tumour classification, they may give insight for other applications.

1.8 Outline of the thesis

This thesis presents the author's work that has been done over the last four years. It consists of two main parts: the background theory and the applications. The first part starts with the introduction to the pattern recognition systems in Chapter 2, followed by Chapter 3 and 4 which explain the linear as well as the kernel-based classifiers. The second part starts with an introduction to brain tumour classification using Magnetic Resonance Spectroscopy (MRS) in Chapter 5, then followed by classification using long echo time MRS spectra in Chapter 6, using short echo time MRS spectra in Chapter 7, and using MRI and short echo time MRSI in Chapter 8. Chapter 9 concludes the thesis. The flow of this thesis is depicted in Figure 1.2.

This thesis summarizes the research in Least Squares Support Vector Machines and its application to brain tumour recognition based on magnetic resonance spectroscopy. Many attempts have been made to take benefit of advanced classification techniques when applied to real data. At the same time these examples serve as a benchmark of the developed technique.

Chapter 2 gives a brief overview of pattern recognition systems, some performance measure and statistical analysis which will be used further in the thesis.

Chapter 3 introduces classification techniques. It begins with the basic idea of linear classification and mentioning some criteria which commonly used. Afterwards Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are explained.

Chapter 4 focuses on the kernel-based classifiers, starting with the original idea of Support Vector Machines, then Least Squares Support Vector Machines are discussed.

Chapter 5 introduces a prospective biomedical application of automated classification using Magnetic Resonance Spectroscopy (MRS). This field is introduced and it is shown how MRS has become an important tool for an accurate non-invasive brain tumour recognition.

Chapter 6 reports the application of brain tumour classification using long echo time proton MRS data.

Chapter 7 explores short echo time proton MRS data for brain tumour classification. Short echo time MRS spectra contain much more information compared to long echo time MRS spectra. However, we face other difficulties : metabolite peaks are strongly overlapping and not easy to separate, artefacts are also harder to discriminate from the signal.

Chapter 8 explores further the relative characteristic information present in MRSI spectra and MR images, for the discrimination of several brain tumour types and healthy brain tissue.

Chapter 9 concludes this work, summarizes the thesis and gives suggestions for further research.

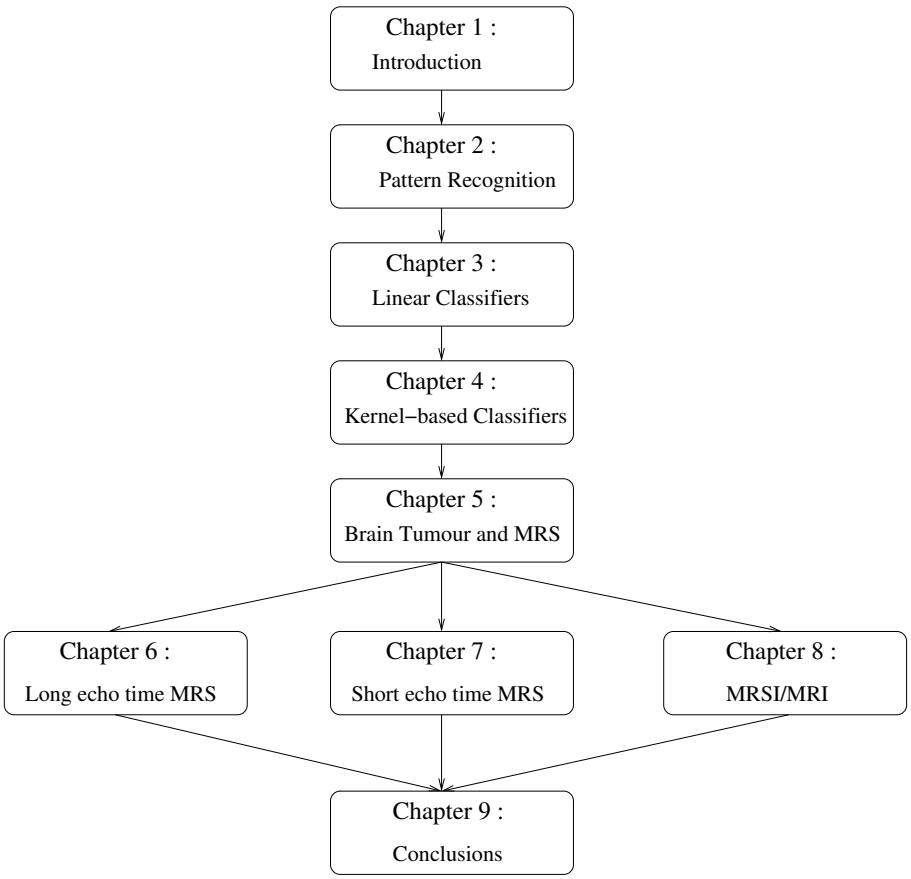


Figure 1.2: *Flow diagram of this thesis.*

Chapter 2

Pattern Recognition Systems

2.1 Pattern Recognition

2.1.1 Patterns

Patterns are the means by which we interpret the world. We are performing the task of pattern recognition at every instant of our lives. The ease with which humans classify and describe patterns often leads to the incorrect assumption that this capability is easy to automate [90].

Pattern recognition is naturally based on patterns. A *pattern* can be as basic as a set of measurements or observations, perhaps represented in a vector or matrix notation. The use of measurements already presupposes some preprocessing and instrumentation system complexity. These measurements could be entities such as blood rate, temperature, volume, age, and the like. Moreover, a pattern can be converted from one representation to another.

2.1.2 Features

In general, features are any extractable measurement used. Examples of low-level features are signal frequency, sample weight. Features can be symbolic (e.g. color), numerical (e.g. length) or a combination of both. Applying *feature extraction* to the input data can result in features as well. Additionally, features may be on a higher level of entities: for example, geometric descriptors, such as aspect ratio of an ellips. Features may be represented by binary variables (e.g. yes/no), discrete (limited values) or continuous variables.

2.1.3 Feature selection

Feature selection is one of the major tasks in any automatic pattern recognition system. The main objective of *feature selection* is to retain the optimum salient characteristic necessary for the recognition process and to reduce the dimensionality of the measurement space, so that effective and easily computable algorithms can be devised for efficient classification. It is important that the extracted features are relevant to the particular task at hand. Also note that errors or noise can be introduced when extracting features. The problem of feature selection has two aspects: the selection of an optimal subset from the available features and the formulation of a suitable criterion to evaluate the goodness of a set of features. In some cases there are mathematical tools that help in feature selection. In other cases, simulation may aid in the choice of appropriate features.

2.1.4 Pattern recognition

Basically, pattern recognition is about *information mapping*, *information labeling* or *information reduction* [90]. An abstract view of the pattern recognition classification problem is depicted in Figure 2.1.

The class-membership space $\mathbf{C}$ denotes an abstract domain for classes c_1, c_2, c_3. Via a relation $\mathbf{G}_i\{i = 1, 2, 3\}$, each class c_i is mapped to a subset of *patterns* in pattern space $\mathbf{P}$, where the i-th pattern is denoted by p_i. These subspaces may overlap, allowing patterns from different classes to share characteristics. Another relation, $\mathbf{F}$, maps patterns from subspaces of $\mathbf{P}$ into features or measured patterns, denoted by m_i. Using this concept, the characterization of pattern recognition problems is simply that, given measurement m_i, we want to invert the mappings $\mathbf{F}$ and $\mathbf{G}$ for all i.

In practice, however, these mappings are rarely invertible, more often not 1:1 (*one-to-one*) functions. As illustrated in Figure 2.1, identical measurements may result from different subsets of pattern p_i, which in turn correspond to different underlying classes. It should also be noted that patterns that are generated by the same class (e.g. p_4 and p_5, from c_3) although *close* in pattern space $\mathbf{P}$, do not necessarily yield *close* measurements (m_1 and m_3 in this case). This is significant when a *clustering* of features is used to measure pattern similarity.

2.1.5 Typical pattern recognition systems

In a general setting the process of pattern recognition is visualized as a sequence of the following steps: (1) data acquisition, (2) preprocessing and enhancement, (3) feature selection and (4) classification. In addition, usually some data that have already been classified or described are assumed to be available for learning

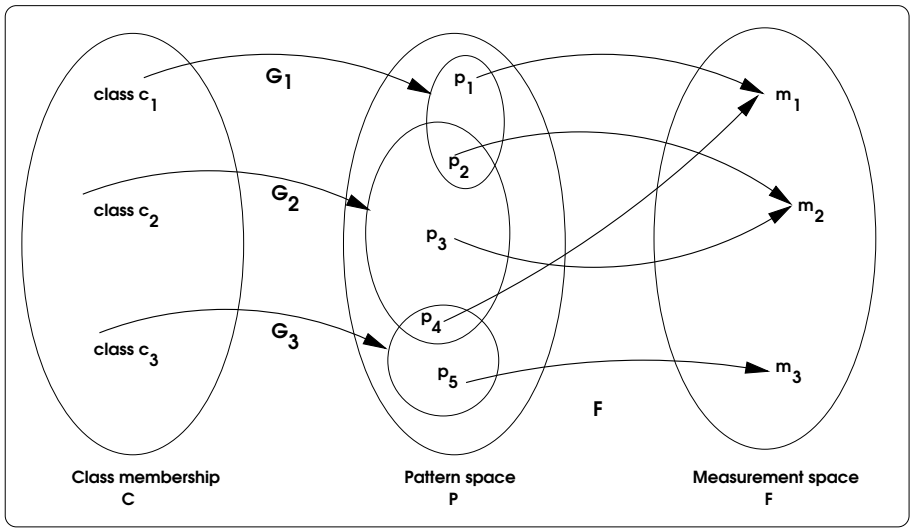


Figure 2.1: *Mapping in an abstract representation of pattern classification systems.*

(the so-called *training set*). Figure 2.2 depicts a typical pattern recognition system [90].

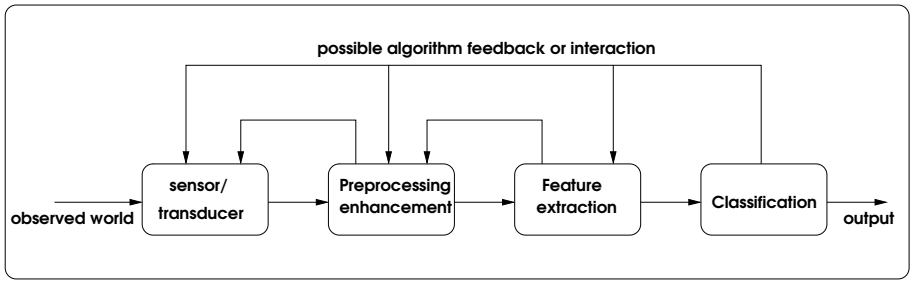


Figure 2.2: *Typical pattern recognition system.*

At the first step, depending on the environment within which the objects are to be classified, data are acquired using a set of sensors or measuring devices. Afterwards, some preprocessing and enhancement can be applied, such as noise reduction, artefact filtering, encoding and enhancement for extracting pattern vectors. Then selected features are extracted, aiming at a manageable amount of information without discarding valuable information, or to make it computationally feasible. Finally, at the fourth step of the scheme the classifier is constructed, or in other words, a transformation relationship is established

between features and classes. This can be, for instance, a linear or nonlinear discriminant function, nearest neighbour rule, a nearest mean (prototype) classification rule or a Bayesian rule for computing *a posteriori* class probability, an artificial neural network or a kernel-based classifier ([23, 89, 113]). This thesis will focus on the development and application of a particular member of kernel-based classifiers, namely the Least Squares Support Vector Machine Classifier (LS-SVM).

2.2 Performance Measures

2.2.1 Correct classification rate

There are many statistical methods for evaluating and estimating the performance of a classifier. Some of the best known methods are *leave-one-out* (loo), *N-fold cross validation* and *bootstrapping* methods [5, 23]. The goal should be to obtain as honest as possible an estimation about the classification accuracy of the system. The *correct classification rate*, or sometimes called *accuracy*, shows the proportion of correct classifications to the total number of classification tests. However this test depends on the assessment setup.

It is always expected that a classifier can reach a performance as high as possible, but overfitting should be avoided. This problem might occur as the classifier fits too much to the training data, but cannot generalize well to new data. A common setup to fairly measure classification performance is by dividing the total number of data into a training set, a validation set and a test set. The classifier is trained with the training set, while the validation set is used to decide when to stop training (i.e. when a minimal error on this validation set is reached). The test data are only used after the training by means of which an independent classification performance can be measured. Figure 2.3 depicts this classification setup.

Unfortunately in real life problems, it is quite often very difficult to have a large dataset. With the limited available number of data, a trade-off should be taken to divide the dataset into a training set and a test set, e.g. 2/3 of the dataset are used as training data while the rest is used later as test data.

Leave-one-out crossvalidation can be used as a method to estimate the classifier performance in unbiased manner. Here, each step one data point is left out and the classifier is trained using the rest and then the classifier is applied to the left out data point. This procedure is repeated such that each data point is left out once. Classification performance is calculated by taking the average of the number of correct classifications.

One should also realize that in real life cases, such as in clinical medicine there

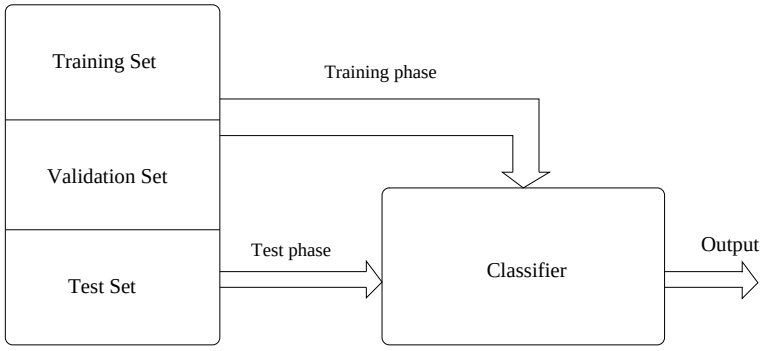


Figure 2.3: *The dataset is divided into a training, validation and test set. The validation set is used in the training process to obtain a good generalization, e.g. to avoid overfitting. The test set is used to measure the performance of the classifier on the independent/unseen data.*

is virtually no perfect test. All tests could have an error rate and on occasion will either fail to identify an abnormality, or identify an abnormality which is not present. It is common to describe this error rate by the terms *true and false positive* and *true and false negative* as follows :

True Positive (TP) : the test result is *positive* in the *presence* of the clinical abnormality.

True Negative (TN) : the test result is *negative* in the *absence* of the clinical abnormality.

False Positive (FP) : the test result is *positive* in the *absence* of the clinical abnormality.

False Negative (FN) : the test result is *negative* in the *presence* of the clinical abnormality.

A tabular format, called contingency table is illustrated in Table 2.1. Table 2.2 defines various terms used to describe the clinical efficiency of a test based on the terms above.

2.2.2 Receiver Operating Characteristic

Receiver Operating Characteristic (ROC) analysis [103] is an objective and highly effective technique for assessing the performance if a measure is used in binary hypothesis testing. This format provides that a data sample be assigned to one of two hypotheses or cases (e.g. normal or pathological) depending on the value of some measured statistic relative to a threshold value. The efficacy

	Predicted Group	
Actual Group	Normal	Abnormal
Normal	TN	FP
Abnormal	FN	TP

Table 2.1: *Contingency table commonly used in clinical medicine to assess the classifier performance in correctly predicting the clinical abnormality to the true diagnosis.*

Sensitivity	$TP/(TP+FN)$
Specificity	$TN/(FP+TN)$
Predictive value (positive)	$TP/(FP+TP)$
Predictive value (negative)	$TN/(FN+TN)$
Relative risk	$\frac{TP/(FP+TP)}{FN/(FN+TN)}$
Likelihood ratio	$\frac{TP/(TP+FN)}{FP/(TN+FP)}$ or $\frac{\text{Sensitivity}}{1-\text{Specificity}}$
Odds	probability$/(1-$probability$)$
Accuracy	$(TP+TN)/total$

Table 2.2: *The definition of various terms used to describe clinical efficiency of a test*

of a measure is then judged on the basis of its sensitivity (the proportion of pathological patients correctly identified) and its specificity (the proportion of control subjects correctly classified). The ROC curve is a graphical presentation of sensitivity versus 1-specificity as a threshold parameter is varied.

Figure 2.4 illustrates the basic idea of how to construct a ROC curve [111]. Suppose there are two overlapping probability density function corresponding to the distribution of two classes. We decide class c_1 for values on the left of the threshold and class c_2 for the values on the right. This decision is associated with an error probability, a, of reaching a wrong decision concering class c_1 (the probability of a correct decision is $1-a$). Similarly, let b be the probability of a wrong decision concerning class c_2. By moving the threshold along the horizontal axis, different areas of a and b can be resulted. It is simple to realize that if the two distributions have complete overlap, then for any position of the threshold $a = 1-b$. This corresponds to the top straight line as depicted in Figure 2.5. On the other hand, we will find that in case the two classes less overlap, the larger area between the curve and the straight line. Furthermore if the two class distributions are completely separated, moving threshold along the whole range of values for a in $[0,1]$ will result in a constant $1-b$ equal to unity. Therefore by calculating the *Area Under the ROC Curve* (AUC) we can measure the class dicrimination capability of a specific classifier. In

practice, the ROC curve can easily be constructed by sweeping the threshold and computing the correct classification percentage over the available training set. The larger the area (the higher AUC value) means the higher the classifier performance.

ROC analysis can be used to choose the best of a host of different candidate diagnostic measures by comparing their ROC areas or to establish for a single measure the tradeoff between misses and false positives by examining the ROC area.

In this thesis ROC analysis and Accuracy are used to measure the performance of the classifiers.

2.2.3 Statistical analysis

This section describes the statistical comparison of the areas under the ROC curves derived from the same set of patients by taking into account the correlation between the areas that is induced by the paired nature of the data. This is used to compare the performance of different classifiers which will be applied on M runs of randomized splittings of the data set.

From $M = 100$ runs, the mean AUCs ($\overline{\text{AUC}}$) are listed in the tables, as well as the standard error SE on the AUC. For each binary classifier C, the mean and standard error of the AUC are calculated.

Consider two classifiers C_1 and C_2 that handle the same input data; e.g. C_1 is *Classifier 1* and C_2 is *Classifier 2* applied to separate input into class 1 or 2. Let the AUC of each classifier $C_i, i = 1, 2$ be $A_{i,l}$ with standard error $SE_{i,l}, i = 1, 2; l = 1, ..., M$, with M the number of stratified randomizations ($M = 100$). The pooled statistics are then given by $\overline{A_i}$ and SE_i ($i = 1, 2$), where n_l is the amount of samples for the stratified randomization $l = 1, ..., M$ [88, 73]:

$$\overline{A_i} = \frac{1}{M}\sum_{l=1}^{M} A_{i,l}, \tag{2.1}$$

$$SE_i = \sqrt{\frac{1}{N-M}\sum_{l=1}^{M}(n_l - 1)SE_{i,l}^2} \tag{2.2}$$

$$= \sqrt{\frac{1}{M}\sum_{l=1}^{M} SE_{i,l}^2} \tag{2.3}$$

The last line is only satisfied if the test set contains an equal amount of samples for each stratified randomization, i.e. $\forall\, l\ n_l = n, N = \sum_{l=1}^{M} n_l$.

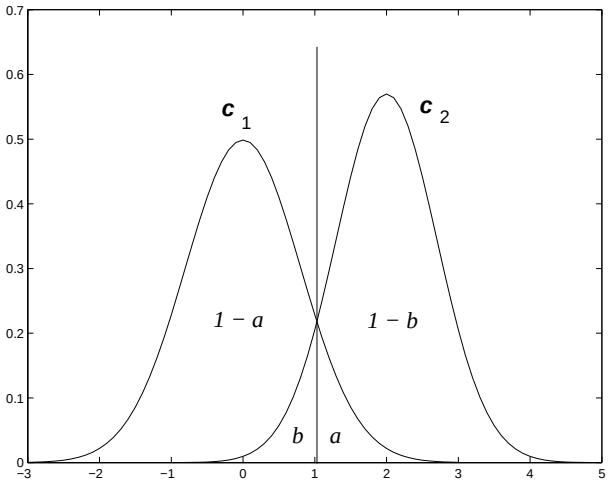


Figure 2.4: *Two overlapping probability density function (pdf) class distributions. Values on the left of the threshold line is classified as class 1, or class 2 otherwise. Moving the threshold line results in different sensitivity and specificity.*

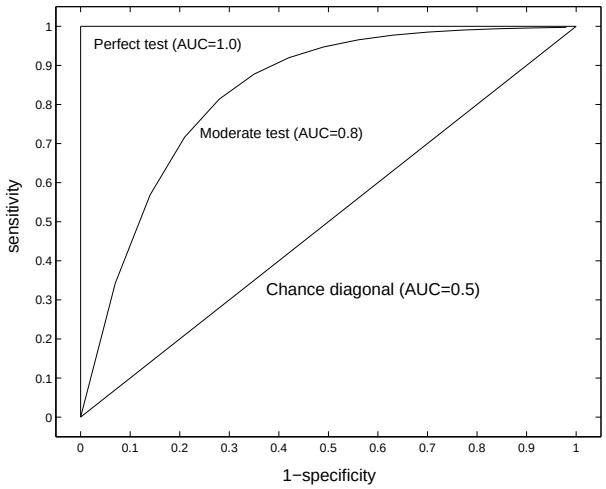


Figure 2.5: *ROC curve for different cases. Area under ROC curve (AUC) serves as a well-established index of diagnostic accuracy. ROC following diagonal line results in AUC=0.5 (pure chance), whereas the maximum value of 1.0 corresponds to perfect assignment (unity sensitivity for all values of specificity).*

A general approach to statistically test whether the areas under two ROC curves derived from the same samples differ significantly from each other is then given by the critical ratio z, defined as [36]:

$$z = \frac{\overline{A_1} - \overline{A_2}}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}},$$

in which r is a quantity representing the correlation introduced between the two areas by studying the same samples. In our study we calculate the z-value based on the pooled statistics $\overline{A_i}, SE_i, i = 1, 2$ from 100 runs as calculated using Eqs. (2.1)-(2.3). If the resulting z-value satisfies $z \geq 1.96$, then $\overline{A_1}$ and $\overline{A_2}$ are statistically different. The cutoff-value 1.96 is taken as the quantity for which, under the hypothesis of equal AUCs ($\overline{A_1} = \overline{A_2}$), $z \geq 1.96$ occurs with a probability of $\alpha = 0.05$ assuming a normal distribution.

This ROC analysis is performed for binary classification. Although ROC analysis has been extended to multiclass classification [99], the result is generally non-intuitive and computationally expensive. This motivates the use of the correct classification rate as performance measure for multiclass classification.

2.3 Conclusions

This chapter introduced the terminologies in a typical pattern recognition system. The flow of information from the observed object, features are extracted and results in classification. The significance of classification performance measures are briefly explained. Some basic terms commonly used in clinical medicine are mentioned as they will be used in this thesis.

Chapter 3

Linear Classification

3.1 Binary classification problems

A simple classification task is to categorize a set of objects, classifying them into two groups on the basis of whether they have some properties or not. This is also called *binary classification*. Some typical examples are:

- medical testing to determine if a patient has a certain disease or not (the classification property is the disease)
- deciding whether an email is a spam or not (the classification property is having (un-)wanted content, in the simplest case)
- quality control in factories; i.e. deciding whether a new product is good enough to be sold, or if it should be discarded (the classification property is passing certain criteria)

Suppose we are given a task of classifying a group of 100 people into two classes: healthy people and unhealthy people. Surely this task is not straightforward as healthiness can be defined in different ways. A very simple approach will be taken for this example.

A first common measurement is the body temperature. Normal body temperature is considered to be 37 degrees Celsius [77]. Body temperature can vary a half degree Celsius above or below 37°C. and still be considered "normal."

If we take as a criterion for an unhealthy person as having a higher body temperature (to have fever), and assume that starting from 37.5 degrees Celcius a

human is considered as unhealthy, then a line can be drawn to roughly separate these two classes.

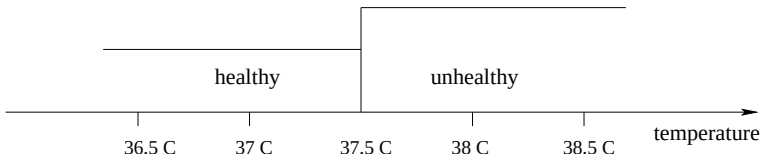


Figure 3.1: *Example of separating healthy and unhealthy persons based on the body temperature.*

However, body temperature varies with many factors including level of activity. For example, food, extra clothing, excitement, and anxiety can all raise the body temperature. A woman's menstrual cycle can also elevate temperature by one degree or more. Therefore rather than a crisp border, a more natural way to describe the problem is by drawing a distribution. Using the measurement on the population a mean temperature of 37 degree for healthy persons is taken, with half degree of variation, while for unhealthy persons a mean of 38.5 degree Celcius is taken with 1 degree of variation.

A histogram as shown in Figure 3.2 gives a better view of the classification, but it is still very difficult to correctly classify a particular person, whose body temperature is located in the range covered by both histograms.

Additional measurements can be added to improve this. Although there are many possible signs of illness, e.g. pale face, red eyes, they should be quantifiable, such as body weight changes, blood pressure or a measure of concentration time. One can take the blood pressure as additional measurement, and assume a person is healthy for blood pressure up to 120 mmHg, otherwise he/she is considered abnormal (blood pressure above 120 mmHg). Now, a pair of values (or a vector) [bodytemp bloodpress]^T can be used to represent each person. The additional element of the vector permits us to draw the classification problem in two-dimensions (2D). Extending the dimensionality adds a new degree of freedom. Separation can be done not only by a single threshold point, but by a line. In higher dimensions, it will be called separating hyperplane.

3.2 Linear classifiers

The simplest form of discriminant function is linear. Linear discriminant function $f(\mathbf{x})$ can be written as

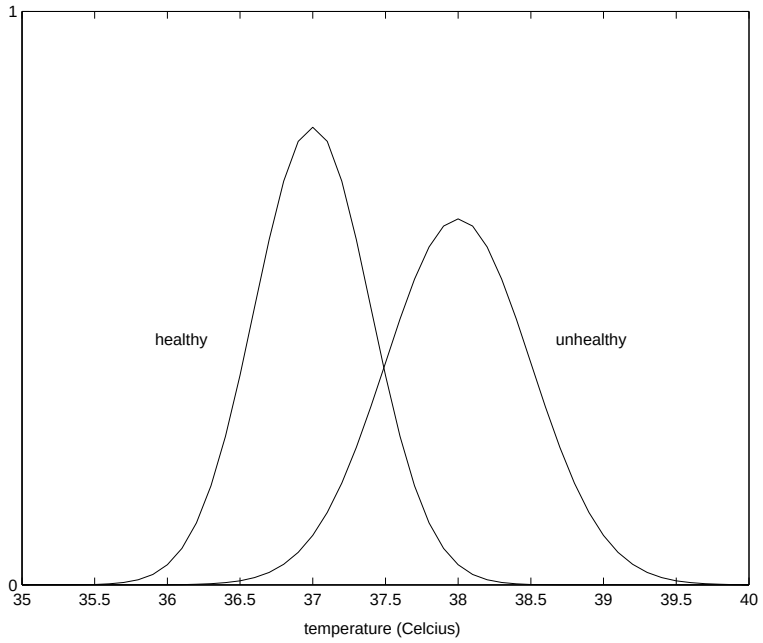


Figure 3.2: *Example of separating healthy and unhealthy persons based on the body temperature distribution.*

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{3.1}$$

where $\mathbf{w} = [w_1, w_2, ...w_l]^T$ is known as the *weight vector* and b as the *bias*. If $\mathbf{x}_1, \mathbf{x}_2$ are two points on the decision hyperplane, then the following is valid

$$0 = \mathbf{w}^T\mathbf{x}_1 + b = \mathbf{w}^T\mathbf{x}_2 + b \Rightarrow \mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0 \tag{3.2}$$

Since the difference vector $\mathbf{x}_1 - \mathbf{x}_2$ obviously lies on the decision hyperplane for any $\mathbf{x}_1, \mathbf{x}_2$, it is clear that the vector $\mathbf{w}$ is orthogonal to the decision hyperplane. Figure 3.4 illustrates this linear classification.

The distance d and z can be easily calculated by

$$d = \frac{|b|}{\sqrt{\mathbf{w}_1^2 + \mathbf{w}_2^2}} \qquad z = \frac{|f(\mathbf{x})|}{\sqrt{\mathbf{w}_1^2 + \mathbf{w}_2^2}} \tag{3.3}$$

In other words, $|f(\mathbf{x})|$ is a measure of Euclidean distance of the point $\mathbf{x}$ from

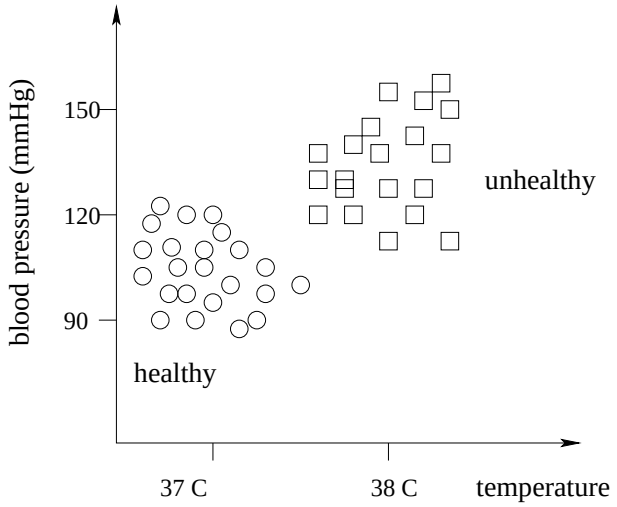


Figure 3.3: *Scatter plot of the binary classification problem*

the decision hyperplane. On one side of the plane $f(\mathbf{x})$ has positive values and on the other negative. In the special case $b = 0$ the hyperplane passes through the origin.

Some criteria commonly used in classification are *distance measure*, *Bayes decision rule*, and *likelihood*. In the following those criteria are explained:

- Distance measure is the simplest and most direct approach to classify data points. Basically, the idea is to classify a data point into the class closest to it. The Euclidean distance and Mahalanobis distance are the two most common definitions. Suppose we have K classes with $(\mu_i, \mathbf{S}_i)$ as the known parameter set of class i, where μ_i is the reference vector of class i and $\mathbf{S}_i$ is the covariance. The Euclidean distance of an observation vector $\mathbf{x}$ from class i is

$$d_i(\mathbf{x}) = \sqrt{||\mathbf{x} - \mu_i||^2} \tag{3.4}$$

  while Mahalanobis distance of vector $\mathbf{x}$ can be written as

$$d_i(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu_i)^T \mathbf{S}_i^{-1} (\mathbf{x} - \mu_i)}. \tag{3.5}$$

  The Euclidean distance is in fact a special case of the Mahalanobis distance.

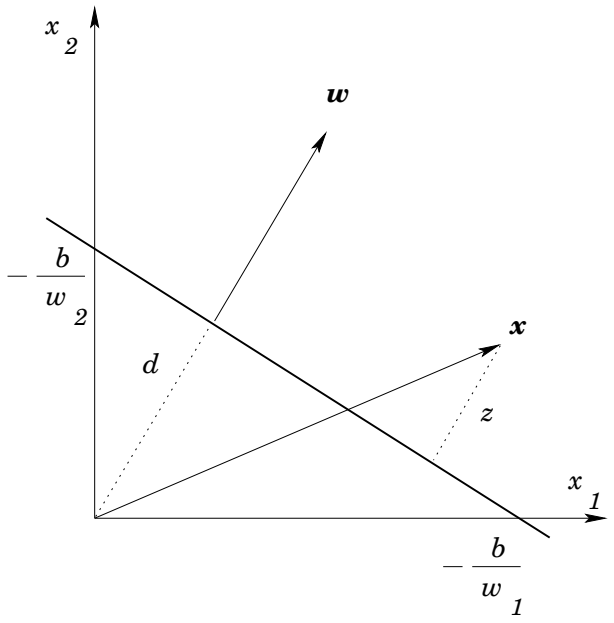


Figure 3.4: *Linear classification.* $f(\mathbf{x}) > 0$ *above the line, and* $f(\mathbf{x}) < 0$ *below the line.*

- Bayes decision rule is based on the assumption that classification problems can be characterized in probabilistic terms and all of the relevant probabilities are known. An observation vector will be assigned to the class that has the largest posterior probability $p(c_j|\mathbf{x})$. Suppose there are k classes $c_1, c_2, ..., c_k$, and the known a priori probability of each class is denoted by $P(c_i), i = 1, 2, ..., k$ and the conditional probability density by $p(\mathbf{x}|c_i), i = 1, 2, ..., k$, then the posteriori probability can be calculated with Bayes rule:

$$p(c_j|\mathbf{x}) = \frac{p(\mathbf{x}|c_j)P(c_j)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|c_j)P(c_j)}{\sum_{i=1}^{k} p(\mathbf{x}|c_i)P(c_i)}. \tag{3.6}$$

- The likelihood criterion is a special case of Bayes decision rule. Assuming that all of the a priori probabilities $P(c_i)$ are equal and the distribution of the classes are normal ($\mathbf{x} \sim \mathcal{N}(\mu_i, \mathbf{S}_i), i = 1, 2, ..., k$. Then $p(c_j|\mathbf{x}) = p(\mathbf{x}|c_j)$ and we have

$$p(\mathbf{x}|c_i) = \frac{1}{2\pi|\mathbf{S}_i|^{1/2}} \exp^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T(\mathbf{x}-\mu_i)}. \tag{3.7}$$

If the parameters of a class are known, the likelihood is in fact the probability density function of the class. The logarithm of the likelihood is usually taken to simplify the calculation. The log-likelihood is written in the following form:

$$P_i(x) = -\frac{1}{2}\ln|\mathbf{S}_i| - \frac{n}{2}\ln 2\pi - \frac{1}{2}(\mathbf{x}-\mu_i)^T\mathbf{S}_i^{-1}(\mathbf{x}-\mu_i). \tag{3.8}$$

However, in practical applications the class probability densities as well as the prior class probabilities may be unknown. One approach is by estimating the class probability densities from the training data and use the estimated class probability densities together with the prior class distributions to construct the decision rule.

Another approach is by assuming that the classes are distributed normally. The major reasons for the popular use of this Gaussian distribution are its computational tractability and the fact that it models adequately a large number of cases. The mean and the covariance matrices can then be estimated from the given data set. If both classes are assumed to have also equal covariance matrices, a linear discriminant analysis can be applied for this problem.

3.3 Linear Discriminant Analysis (LDA)

The Linear Discriminant Analysis originally designed by Fisher [28] for taxonomic classification. Suppose $\mathbf{L}$ represents the linear transformation that maps the original n-dimensional space onto a m-dimensional feature subspace where normally $m \ll n$. The new feature vectors $\mathbf{z}_k \in \mathbb{R}^m$ are defined by $\mathbf{z}_k = \mathbf{L}^T\mathbf{x}$, $k = 1, ..., N$.

LDA searches for those vectors in the underlying space that best discriminate among classes (rather than those that best describe the data). More formally, given a number of independent features relative to which the data is described, LDA creates a linear combination of these which yields the largest mean differences between the desired classes.

For all samples of the classes, we define two measures [68]. The first one is called *within-class* scatter matrix, as given by

$$\mathbf{S}_w = \sum_{j=1}^{n_C}\sum_{k=1}^{N_j}(\mathbf{x}_k^j - \mu_j)(\mathbf{x}_k^j - \mu_j)^T, \tag{3.9}$$

where $\mathbf{x}_k^j$ is the kth sample of class j, μ_j is the mean of class j, n_C is the

number of classes, and N_j is the number of samples in class j.

The second measure is called *between-class* scatter matrix

$$\mathbf{S}_b = \sum_{j=1}^{n_C} (\mu_j - \mu)(\mu_j - \mu)^T, \tag{3.10}$$

where μ represents the mean of all classes.

The goal is to maximize the between-class measure while minimizing the within-class measure. A common way to do this is by maximizing the ratio $\frac{\det(\mathbf{S}_b)}{\det(\mathbf{S}_w)}$. It has been proven [28] that if $\mathbf{S}_w$ is a nonsingular matrix then this ratio is maximized when the column vectors of the projection matrix, $\mathbf{L}$, are the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$. However, it should be noted that there are at most $n_C - 1$ nonzero generalized eigenvectors and, so, an upper bound on m is $n_C - 1$. We also requre at least $n + n_C$ samples to guarantee that $\mathbf{S}_w$ does not become singular. Problems arise when dealing with high dimensional data. The main difficulty in this case lies in the fact that the within-class scatter matrix is almost always singular, therefore the standard algorithm cannot be used. Another disadvantage is a high computational complexity of solving $\mathbf{S}_w^{-1}\mathbf{S}_b$ when working with the high dimensional input space. To solve this [4, 104] propose the use of an intermediate space. Principal Component Analysis is performed to reduce the dimensionality of the data, followed by applying LDA to this data in the reduced dimension. This method was shown efficient by empirical testing but had no theoretical background upon which it was proven correct [27].

3.4 Principal Component Analysis (PCA)

Principal Component Analysis is one of the simplest and common statistical methods to reduce high-dimensional data. It consists of a transformation from the original variables into a new set of uncorrelated variables, called principal components (PCs). These new variables are linear combinations of the original ones ranked in decreasing order of importance and the number of PCs is determined by the cumulative variance explained by the first components. If the original variables are highly correlated, then the first few PCs will account for most of the variation and the remaining PCs can be discarded with slight loss of information. The more PCs are used the better the reconstructed data fits the original noisy data.

Basically PCA maps data into an orthogonal space, where the axes of this new coordinate system lie along the direction of maximum variance of the original

data. By this transformation, it allows the mapping of vectors $\mathbf{x} \in \mathbb{R}^n$ into a lower dimensional vectors $\mathbf{z} \in \mathbb{R}^m$ with $m < n$. The covariance matrix can be estimated by

$$\hat{\mathbf{S}} = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{x}_k - \mu)(\mathbf{x}_k - \mu)^T \tag{3.11}$$

where $\mu = \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k$ and computes the eigenvalue decomposition

$$\hat{\mathbf{S}}\mathbf{u}_i = \lambda_i \mathbf{u}_i. \tag{3.12}$$

Taking the m largest eigenvalues and their corresponding eivenvectors, the score or transformed variables can be calculated using

$$z_i = \mathbf{u}_i^T (\mathbf{x} - \mu), \ i = 1, ..., m. \tag{3.13}$$

In signal processing, PCA is expected to eliminate a significant number of dimensions by partitioning the feature space into subspaces of signal and noise. This is assuming that most of the variance in the data set will be explained by the signal and small contribution by the noise. In this case, taking some PCs which cover a certain amount of variance, e.g. 75%, will significantly reduce the dimension while keeping most information and filtering out noise.

PCA is a useful method for reducing the dimensionality of spectral data because if often works well when the underlying dimensionality of the input space is small. However, it is only appropriate if the underlying relationship between the variables are linear, and most of the variance in the data corresponds to the signal rather than the noise.

3.5 Conclusions

A common way to categorize data into different classes is by creating a linear classification. However, in certain practical applications the dimensionality of the data in the problem space may be very high and it is computationally very expensive. If the underlying relationship between the variables are linear, PCA can be applied to reduce the dimensionality. Otherwise, for a more complex problem, a nonlinear mapping is needed to achieve a better classification.

Chapter 4

Kernel-based Classification

4.1 Introduction

This chapter presents the kernel-based classification. The first section introduces the idea of maximal margin classification, optimal separating hyperplane, followed by kernel methods as the basis for the extension towards nonlinear classification as introduced by Vapnik, which is called Support Vector Machines (SVM). Afterwards, a derivation of the least-squares version from the nonlinear SVM is described.

4.2 Maximal margin classifiers

Consider the class of hyperplanes $\mathbf{w}^T\mathbf{x} + b = 0, \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}$, corresponding to a decision function

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b). \tag{4.1}$$

First, we consider the case of linearly separable data. A hyperplane can separate two classes of data in many possible ways (see Figure 4.1 left). There is no unique separating hyperplane, unless we add a criterion to decide which is the best or the *optimal* separating hyperplane.

Basically the idea of learning from examples is to recognize the pattern of a class by examining the training points corresponding to that class. New data points are assumed to lie somewhere around the known training data. Therefore, a

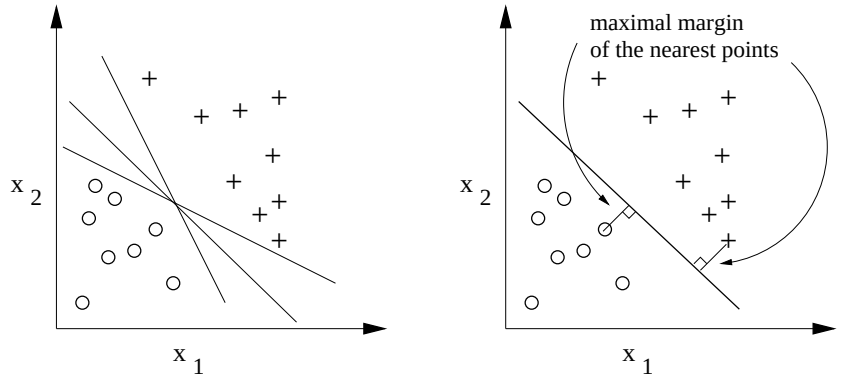


Figure 4.1: *Linear classification: (***left***) example of classification problem where the separating hyperplane is not unique, (***right***) example of a unique hyperplane which corresponds to a maximal margin of the closest points to the separating hyperplane.*

hyperplane should be chosen such that a small shift of the data does not result in prediction change. If the distance between the separating hyperplane and the training points becomes too small, even test examples very close to the training samples may be classified incorrectly.

Based on this idea, Vapnik and Chervonenkis presumed that the generalisation ability depends on the distance between the hyperplane and the training points. They introduced the *Generalized Portrait*, a learning algorithm for separable problems, by constructing a hyperplane which maximally separates the classes (maximum margin):

$$\max_{\mathbf{w},b} \min\{\|\mathbf{x}-\mathbf{x}_k\| : \mathbf{x} \in \mathbb{R}^n, \mathbf{w}^T\mathbf{x}+b=0, k=1,...,N\}. \tag{4.2}$$

To show how this hyperplane can be constructed in an efficient way, we need to start with some definitions.

Definition 4.1 (**Separability**)

A training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) : \mathbf{x}_k \in \mathbb{R}^n, y_k \in \{-1, +1\}\}$ is called separable by a hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ if there exist both a unit vector $\mathbf{w}$ ($||\mathbf{w}|| = 1$) and a constant b such that the following equalities hold:

$$\mathbf{w}^T\mathbf{x}_k + b \geq +1 \quad \text{for} \quad y_k = +1 \tag{4.3}$$

$$\mathbf{w}^T\mathbf{x}_k + b \leq -1 \quad \text{for} \quad y_k = -1. \tag{4.4}$$

The hyperplane defined by $\mathbf{w}$ and b is called a separating hyperplane. □

Definition 4.2 (**Margin**)

Consider separating hyperplane H as defined by $\mathbf{w}^T\mathbf{x} + b = 0$.

- the margin $\zeta_k(\mathbf{w}, b)$ of a training point x_k is defined as the distance between H and $\mathbf{x}_k$:
$$\zeta_k(\mathbf{w}, b) = y_k(\mathbf{w}^T\mathbf{x} + b).$$
- the margin $\zeta_D(\mathbf{w}, b)$ of a set of vectors $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_k\}$ is defined as the minimum distance from H to the vectors in $\mathcal{D}$:
$$\zeta_{\mathcal{D}}(\mathbf{w}, b) = \min_{\mathbf{x}_k \in \mathcal{D}} \zeta_k(\mathbf{w}, b)$$

□

Now consider the unit vector $\mathbf{w}^*$ and the constant b^* which maximize the margin of the training set $\zeta(\mathbf{w}, b)$ and also satisfy the condition (4.3) and (4.4). This pair of $(\mathbf{w}^*, b^*)$ defines the hyperplane which separates the positive from the negative examples with the largest margin. This hyperplane is also called *maximal margin hyperplane* or *optimal separating hyperplane*.

Definition 4.3 (**Optimal separating hyperplane**)

The Optimal hyperplane of a training set $\mathcal{D}$ is defined by

$$(\mathbf{w}^*, b^*) = \arg\max_{\mathbf{w}, b} \zeta_{\mathcal{D}}(\mathbf{w}, b)$$

□

Vapnik proves on the uniqueness of the optimal separating hyperplane in [124]. With a unique $\mathbf{w}^*$ we can describe b^* by

$$b^* = \frac{1}{2}\left(\min_{k \in I_+} \mathbf{w}^*\mathbf{x}_k - \max_{k \in I_-} \mathbf{w}^*\mathbf{x}_k\right),$$

where $I_+ = \{k \,|\, y_k = +1\}$ and $I_- = \{k \,|\, y_k = -1\}$.

4.3 Construction of the optimal hyperplane

Given a training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) : \mathbf{x}_k \in \mathbb{R}^n, y_k \in \{-1, +1\}\}$, the optimal separating hyperplane is the solution to the optimization problem

$$\begin{array}{rl} \max & \zeta_{\mathcal{D}}(\mathbf{w}, b) \\ \text{subject to} & \zeta_{\mathcal{D}}(\mathbf{w}, b) > 0 \\ & \|\mathbf{w}\| = 1 \end{array} \tag{4.5}$$

However, solving (4.5) is not straightforward, it involves nonlinear constraints, while the objective function itself is neither linear nor quadratic. We can rewrite the optimization problem equivalently as

$$\begin{array}{rll} \min & \frac{1}{2}\|\mathbf{w}\|^2 & \\ \text{subject to} & \mathbf{w}^T\mathbf{x} + b \geq +1 & \text{for } y_k = +1 \\ & \mathbf{w}^T\mathbf{x} + b \leq -1 & \text{for } y_k = -1 \end{array} \tag{4.6}$$

Theorem 4.4 *(Vapnik) [124] Vector* $\mathbf{w}_0$ *that solves problem (4.6) is related to the vector* $\mathbf{w}^*$ *solving problem (4.5) by the equality*

$$\mathbf{w}^* = \frac{\mathbf{w}_0}{\|\mathbf{w}_0\|}.$$

This implies that the construction of the optimal hyperplane first needs to solve problem (4.6) with linear constraints, where the constraints can be simplified as

$$y_k(\mathbf{w}^T\mathbf{x}_k + b) - 1 \geq 0, \ k = 1, ..., N. \tag{4.7}$$

The Lagrangian for this problem is

$$\mathcal{L}(\mathbf{w}, b, e; \alpha) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{k=1}^{N} \alpha_k\{y_k[\mathbf{w}^T\mathbf{x}_k + b] - 1\} \tag{4.8}$$

with Lagrange multipliers $\alpha_k \geq 0$ for $k = 1, ..., N$. The solution is characterized by the saddle point of the Lagrangian

$$\max_{\alpha} \min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b; \alpha). \tag{4.9}$$

This leads to

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 & \rightarrow \quad \mathbf{w} = \sum_{k=1}^{N} \alpha_k y_k \mathbf{x}_k \\ \frac{\partial \mathcal{L}}{\partial b} = 0 & \rightarrow \quad \sum_{k=1}^{N} \alpha_k y_k = 0 \end{cases} \tag{4.10}$$

with resulting classifier

$$y(\mathbf{x}) = \text{sign}[\sum_{k=1}^{N} \alpha_k y_k \mathbf{x}_k^T \mathbf{x} + b]. \tag{4.11}$$

Elimination of $\mathbf{w}$ from (4.8) with (4.10) gives the following Quadratic Programming (QP) problem as the dual problem in the Lagrange multipliers α_k

$$\begin{aligned} \max_{\alpha} \mathcal{J}_D(\alpha) = \quad & -\frac{1}{2} \sum_{k,l=1}^{N} y_k y_l \mathbf{x}_k^T \mathbf{x}_l \alpha_k \alpha_l + \sum_{k=1}^{N} \alpha_k \\ \text{such that} \quad & \sum_{k=1}^{N} \alpha_k y_k = 0 \end{aligned} \tag{4.12}$$

Note that this problem is solved in $\alpha = [\alpha_1 ... \alpha_N]^T$, and not in $\mathbf{w}$. This QP problem has a number of interesting properties [102]:

- global and unique solution:
  The matrix related to the quadratic term in α in this quadratic form is positive definite or positive semidefinite. In the case that the matrix is positive definite (all eigenvalues strictly positive) the solution α to this QP problem is global and unique. When the matrix is positive semidefinite (all eigenvalues are positive with some zero eigenvalues) then the solution is global but not necessarily unique [29]. It may happen that the solution of $(\mathbf{w}, b)$ is unique but α not. In terms of $\mathbf{w} = \sum_k \alpha_k y_k \mathbf{x}_k$ this means that there may exist equivalent expansions of $\mathbf{w}$ which require fewer support vectors.

- sparseness:
  An interesting property is that many of the resulting α_k values are equal to zero. Hence the obtained solution is sparse. This means that in the resulting classifier the sum should be taken only over non-zero α_k values instead of all training data points:

  $$y(\mathbf{x}) = \text{sign}[\sum_{k=1}^{\#SV} \alpha_k y_k \mathbf{x}_k^T \mathbf{x} + b].$$

  where $\#SV$ is the number of support vectors.

- geometric meaning of support vectors:
  The support vectors obtained from the QP problem are located close to the decision boundary.

- non-parametric/parametric issues:
  Note that the size of the solution vector α grows with the number of data points N. It means, the dual problem actually corresponds to a non-parametric approach, while in the primal problem it is parametric because the size of $\mathbf{w}$ is fixed independently of the number of data points. As a consequence, for large data sets it might be advantageous to solve the primal problem but in higher dimensional input space it is better to solve the dual problem as the size of the solution vector α does not depend on the dimension n of the input space.

4.4 Optimal hyperplane for linearly nonseparable case

In the previous section the SVM solution of a linearly separable classification problem is explained. However, in most real life problems we deal with noisy data which will render simple linear separation impossible. No feasible solution to the margin maximization problem can be found, due to the fact that the objective function (i.e. the dual Lagrangian) is growing arbitrarily large.

To generalize the previous result with this case, we need to relax the separability constraints, tolerating some misclassifications in the overlapping region. For each violation of the constraints will be punished with a misclassification penalty.

Cortes and Vapnik in 1995 [16] gave the extension of linear SVM to the nonseparable case. Basically it is done by introducing positive slack variable ξ_k in the constraints. The inequalities are transformed to

$$y_k[\mathbf{w}^T\mathbf{x}_k + b] \geq 1 - \xi_k, \;\; k = 1, ..., N. \tag{4.13}$$

We are interested in the smallest slack variable satisfying

$$\xi_k = \max\{0, 1 - y_k[\mathbf{w}^T\mathbf{x}_k + b]\}. \tag{4.14}$$

It measures how many points fail to have a margin of $1/\|\mathbf{w}\|$. The values of ξ_k indicate where $\mathbf{x}_k$ lies compared to the separating hyperplane.

- $\xi_k \geq 1 : y_k[\mathbf{w}^T\mathbf{x}_k + b] < 0$, misclassification.
- $0 < \xi_k < 1 : \mathbf{x}_k$ is classified correctly, but lies inside the margin.
- $\xi_k = 0 : \mathbf{x}_k$ is classified correctly, and lies outside the margin or exactly on the margin boundary.

In the primal weight space the optimization problem becomes

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \mathcal{J}_{\mathcal{P}} = \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + c\sum_{k=1}^{N}\xi_k \\ \text{such that} \quad & y_k[\mathbf{w}^T\mathbf{x}_k + b] \geq 1 - \xi_k, \ k = 1, ..., N \\ & \xi_k \geq 0, \ k = 1, ..., N \end{aligned} \tag{4.15}$$

where c is a positive real constant. We should then consider the following Lagrangian

$$\mathcal{L}(\mathbf{w}, b, \xi; \alpha, \nu) = \mathcal{J}_{\mathcal{P}}(\mathbf{w}, \xi) - \sum_{k=1}^{N}\alpha_k(y_k[\mathbf{w}^T\mathbf{x}_k + b] - 1 + \xi_k) - \sum_{k=1}^{N}\nu_k\xi_k \tag{4.16}$$

with Lagrange multipliers $\alpha_k \geq 0, \nu_k \geq 0$ for $k = 1, ..., N$. The second set of Lagrange multipliers is needed due to the additional slack variables ξ_k. The solution is given by the saddle point of the Lagrangian:

$$\max_{\alpha,\nu}\min_{\mathbf{w},b,\xi}\mathcal{L}(\mathbf{w}, b, \xi; \alpha, \nu). \tag{4.17}$$

One obtains

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial\mathbf{w}} = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{k=1}^{N}\alpha_k y_k \mathbf{x}_k \\ \frac{\partial\mathcal{L}}{\partial b} = 0 \quad \rightarrow \quad \sum_{k=1}^{N}\alpha_k y_k = 0 \\ \frac{\partial\mathcal{L}}{\partial\xi_k} = 0 \quad \rightarrow \quad 0 \leq \alpha_k \leq c, \ k = 1, ..., N \end{cases} \tag{4.18}$$

which gives the following dual QP problem after substituting (4.18) into (4.16):

$$\max_{\alpha} \mathcal{J}_{\mathcal{D}} = \quad -\frac{1}{2}\sum_{k,l=1}^{N} y_k y_l \mathbf{x}_k^T \mathbf{x}_l \alpha_k \alpha_l + \sum_{k=1}^{N} \alpha_k \tag{4.19}$$

$$\text{such that} \quad \sum_{k=1}^{N} \alpha_k y_k = 0,$$
$$0 \leq \alpha_k \leq c, \ k = 1, ..., N.$$

In comparison with the linearly separable case (4.13) this problem has additional box constraints.

4.5 Kernel-based classifiers

In the previous sections we have seen how linear classifiers in the input space can easily be solved by standard optimization techniques. Although linear classifiers are easy to handle, they sometimes pose severe restrictions on the learning task. The target concept may be too complex to be expressed as a linear combination of the given attributes. This problem can be overcome by an approach called *kernel technique*, which dates back to the early 1960s and was introduced as the method of potential functions by Aizerman et al. [1].

Let $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^n$ denote a real valued random input vector, and $y \in \{-1, +1\}$ discrete real valued random output variable and let $\Omega \subseteq \mathbb{R}^{n_{\mathcal{H}}}$ denote a high dimensional feature space. The SVM method basically maps the input vector $\mathbf{x}$ into the high dimensional feature space Ω through some nonlinear mapping $\varphi : \mathcal{D} \to \Omega$. In this feature space, one consider the linear function

$$f(x) = sign[\mathbf{w}^T \varphi(x) + b]. \tag{4.20}$$

This linear function may generalize well in solving classification problems, however, it remains a problem to solve the calculation in the high dimensional feature space. Interestingly, no explicit construction of the nonlinear mapping $\varphi(\mathbf{x})$ is needed. This is motivated by the following result. The inner product in the feature space $\varphi(x_k)^T \varphi(x_l)$ can be replaced with the corresponding kernel $K(x_k, x_l)$ satisfying Mercer's condition.

Theorem 1 *(Mercer) [69]. Let $K \in L^2(C), g \in L^2(C)$ where C is a compact subset of $\mathbb{R}^d$. To guarantee that a continuous symmetric function K has an expansion*

$$K(t,z) = \sum_{k=1}^{\infty} a_k \phi(t)\phi(z) \tag{4.21}$$

with positive coefficients $a_k > 0$*, it is necessary and sufficient that the condition*

$$\int_C \int_C K(t,z)g(t)g(z)dtdz \geq 0 \tag{4.22}$$

be valid for all $g \in L^2(C)$*.*

Suppose $w = \sum_{k=1}^{N} \alpha_k \varphi(x_k)$, and based on the Mercer theorem above, the linear function in the feature space (4.20) has the following equivalent representation in input space

$$f(\mathbf{x}) = sign[\sum_{k=1}^{N} \alpha_k K(\mathbf{x}, \mathbf{x}_k) + b],\ b \in \mathbb{R}, \alpha_k \in \mathbb{R} \tag{4.23}$$

where $\mathbf{x}_k$ are vectors and $K(\mathbf{x}, \mathbf{x}_k)$ is a given function satisfying Mercer's condition. While in the primal space the parameter w may have a range over an "infinite dimensional" parameter set, in the dual problem the dimension of the parameter vector α grows with the number of data points.

Using Mercer's theorem to replace the inner product $\varphi(x_k)^T \varphi(x_l)$ with its corresponding kernel $K(x_k, x_l)$ is often called the *kernel trick*. It enables us to work in a huge dimensional feature space without actually having to do explicit computations in this space. Computations are done in another space after applying this kernel trick. In the case of support vector machines, one starts from a formulation in the primal weight space with a high dimensional feature space by applying transformations $\varphi(\cdot)$. The solution is calculated not in this primal weight space, but in the dual space of Lagrange multipliers after applying the kernel trick. In this way classification is done implicitly in a high dimensional feature space rather than in the original input space.

4.6 Nonlinear SVM classifiers

By combining the idea of an optimal separating hyperplane with a kernel induced mapping to a high dimensional feature space, we extend the idea from linear to nonlinear classifiers. One can formally replace $\mathbf{x}$ by $\varphi(\mathbf{x})$ and apply the kernel trick where possible. However, notice that $\varphi(\mathbf{x})$ can be infinite dimensional, and hence also the $\mathbf{w}$ vector. While for linear SVM one can in fact

equally well solve the primal problem in $\mathbf{w}$ as the dual problem in the support values α, this is no longer the same for the nonlinear SVM case because in the primal problem the unknown $\mathbf{w}$ can be infinite dimensional.

With slight modification, for the nonlinear case we can write

$$\mathbf{w}^T\varphi(\mathbf{x}_k) + b \geq +1 \quad \text{for} \quad y_k = +1 \tag{4.24}$$
$$\mathbf{w}^T\varphi(\mathbf{x}_k) + b \leq -1 \quad \text{for} \quad y_k = -1 \tag{4.25}$$

which is equivalent to

$$y_k[\mathbf{w}^T\varphi(\mathbf{x}_k) + b] \geq 1, \ k = 1, ...N \tag{4.26}$$

in the case of separable data. The classification can be written as

$$y(\mathbf{x}) = \text{sign}[\mathbf{w}^T\varphi(\mathbf{x}) + b]. \tag{4.27}$$

The optimization problem becomes

$$\min_{\mathbf{w},b,\xi} \mathcal{J}_P = \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + c\sum_{k=1}^{N}\xi_k \tag{4.28}$$
$$\text{such that} \quad y_k[\mathbf{w}^T\varphi(\mathbf{x}_k) + b] \geq 1 - \xi_k, \ k = 1, ..., N$$
$$\xi_k \geq 0, \ k = 1, ..., N. \tag{4.29}$$

The Lagrangian is constructed:

$$\mathcal{L}(\mathbf{w}, b, \xi; \alpha, \nu) = \mathcal{J}_P(\mathbf{w}, \xi) - \sum_{k=1}^{N}(\alpha_k y_k[\mathbf{w}^T\varphi(\mathbf{x}_k) + b] - 1 + \xi_k) - \sum_{k=1}^{N}\nu_k\xi_k \tag{4.30}$$

with Lagrange multipliers $\alpha_k \geq 0, \nu_k \geq 0$ for $k = 1, ..., N$. The solution is given by the saddle point of the Lagrangian:

$$\max_{\alpha,\nu} \min_{\mathbf{w},b,\xi} \mathcal{L}(\mathbf{w}, b, \xi; \alpha, \nu). \tag{4.31}$$

This leads to

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 & \rightarrow \quad \mathbf{w} = \sum_{k=1}^{N} \alpha_k y_k \varphi(\mathbf{x}_k) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 & \rightarrow \quad \sum_{k=1}^{N} \alpha_k y_k = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_k} = 0 & \rightarrow \quad 0 \leq \alpha_k \leq c,\ k = 1, ..., N. \end{cases} \tag{4.32}$$

The dual problem becomes

$$\begin{aligned} \max_{\alpha} \mathcal{J}_D(\alpha) = \quad & -\frac{1}{2} \sum_{k,l=1}^{N} y_k y_l K(\mathbf{x}_k, \mathbf{x}_l) \alpha_k \alpha_l + \sum_{k=1}^{N} \alpha_k \\ \text{such that} \quad & \sum_{k=1}^{N} \alpha_k y_k = 0 \\ & 0 \leq \alpha_k \leq c,\ k = 1, ..., N. \end{aligned} \tag{4.33}$$

In this quadratic form, the kernel trick is applied

$$K(\mathbf{x}_k, \mathbf{x}_l) = \varphi(\mathbf{x}_k)^T \varphi(\mathbf{x}_l) \tag{4.34}$$

for $k = 1, ..., N$. Finally the nonlinear SVM classifier takes the form

$$y(\mathbf{x}) = \text{sign}[\sum_{k=1}^{N} \alpha_k y_k K(\mathbf{x}, \mathbf{x}_k) + b] \tag{4.35}$$

with α_k positive real constants which are the solution to a QP problem. The next problem is the determination of the constant b. Karush-Kuhn-Tucker (KKT) [29] complementarity conditions state that the product of the dual variables and the constraints should be zero at the optimal solution. Therefore, using the KKT conditions, it yields

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \rightarrow & \mathbf{w} = \sum_{k=1}^{N} \alpha_k y_k \varphi(\mathbf{x}_k) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow & \sum_{k=1}^{N} \alpha_k y_k = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_k} = 0 \rightarrow & c - \alpha_k - \nu_k = 0 \\ & \alpha_k \{ y_k [\mathbf{w}^T \varphi(\mathbf{x}_k) - 1 + \xi_k] \} = 0, \quad k = 1, ..., N \\ & \nu_k \xi_k = 0, \qquad\qquad\qquad\quad k = 1, ..., N \\ & \alpha_k \geq 0, \qquad\qquad\qquad\quad k = 1, ..., N \\ & \nu_k \geq 0, \qquad\qquad\qquad\quad k = 1, ..., N. \end{cases} \tag{4.36}$$

From $\nu_k \xi_k = 0$ we have for the solutions $\mathbf{w}^\star, b^\star, \xi^\star, \alpha^\star, \nu^\star$ to this problem that $\xi_k^\star = 0$ for $\alpha_k^\star \in (0, c)$. Hence

$$y_k [\mathbf{w}^T \varphi(\mathbf{x}_k) + b] - 1 = 0 \quad \text{for } \alpha_k \in (0, c), \tag{4.37}$$

which means that one can take any training data point for which $0 \leq \alpha_k \leq c$ and use that equation to compute the bias term b.

Some properties of the nonlinear SVM classifier and its solution are the following [102]:

- choice of kernel function:
  Four common choices of kernels are

  1. $K(x, z) = x^T z$ (linear SVM)
  2. $K(x, z) = (\tau + x^T z)^d$ (polynomial SVM of degree d)
  3. $K(x, z) = \exp(-\|x - z\|_2^2 / \sigma^2)$ (RBF kernel)
  4. $K(x, z) = \tanh(\kappa_1 x^T z + \kappa_2)$ (Multilayer Perceptron/MLP)

  The Mercer condition holds for all σ values in the RBF kernel case and τ values in the polynomial case but not for all possible choices of κ_1, κ_2 in the MLP case.

- global and unique solution:
  As in the linear SVM case the solution to the convex QP problem is again global and unique provided that one chooses a positive definite kernel for $K(\cdot, \cdot)$. This choice guarantees that the matrix involved in the

QP problem is positive definite as well, and the kernel trick is applicable. For a positive semidefinite kernel the solution to the QP problem is global but not necessarily unique.

- sparseness:
  As for the linear SVM classifier case many α_k values are equal to zero in the solution vector. In the dual space the nonlinear SVM classifier takes the form
  $$y(\mathbf{x}) = \text{sign}[\sum_{k=1}^{\#SV} \alpha_k y_k K(\mathbf{x}, \mathbf{x}_k) + b]$$
  where the sum is taken over the non-zero α_k values which correspond to support vectors $\mathbf{x}_k$ of the training data set.

- geometric meaning of support vectors:
  The support vectors obtained from the QP problem are located close to the nonlinear decision boundary.

- non-parametric/parametric issues:
  Both the primal and the dual problem have neural network representation interpretations. The problem in the primal weight space is parametric, while the dual is non-parametric. Note that in the dual problem the size of the QP problem is not influenced by the dimension n of the input space.

4.7 Least Squares SVM classifiers

The advantages of nonlinear SVM classifiers are their ability to solve classification problems by means of convex quadratic programming, as well as the sparseness as a result of this QP problem. Suykens [101] in 1999 proposed the idea of modifying the Vapnik's SVM formulation by adding a least squares term in the cost function, which transformed the problem from solving a QP problem to solving a set of linear equations. This approach significantly reduces the costs in complexity and computation time for solving the problem.

Suykens formulated the SVM modification as follows:

$$\begin{aligned} \min_{\mathbf{w},b,e} \mathcal{J}_P = \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma \sum_{k=1}^{N} e_k^2 \\ \text{such that} \quad & y_k[\mathbf{w}^T\varphi(\mathbf{x}_k) + b] = 1 - e_k, \ k = 1, ..., N \end{aligned} \tag{4.38}$$

which corresponds to a classifier in the primal space:

$$y(\mathbf{x}) = \text{sign}[\mathbf{w}^T \varphi(\mathbf{x}) + b] \tag{4.39}$$

where $\varphi(\cdot)$ is the mapping to the high dimensional feature space as in the standard SVM. The Vapnik formulation is modified here at two points. First, instead of inequality constraints, equality constraints are used, where the value 1 at the right hand side is considered as a target value instead of a threshold value. Error variable e_k allows some tolerance of misclassification in the case of overlapping distributions. These error variables function similarly as the slack variable ξ_k in SVM formulation. Second, a squared loss function is taken for this error variable. As explained below, these modifications will greatly simplify the problem.

In the case of a linear classifier the primal problem can be solved easily, however $\mathbf{w}$ in general might become infinite dimensional. Therefore let us derive the dual problem for the LS-SVM nonlinear classifier formulation. The Lagrangian for the problem is

$$\mathcal{L}(\mathbf{w}, b, e; \alpha) = \mathcal{J}_P(\mathbf{w}, e) - \sum_{k=1}^{N} \alpha_k \{y_k[\mathbf{w}^T \varphi(\mathbf{x}_k) + b] - 1 + e_k\} \tag{4.40}$$

where the α_k values are the Lagrange multipliers, which can be positive or negative now due to the equality constraints.

The conditions for optimality yield

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 & \rightarrow \quad \mathbf{w} = \sum_{k=1}^{N} \alpha_k y_k \varphi(\mathbf{x}_k) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 & \rightarrow \quad \sum_{k=1}^{N} \alpha_k y_k = 0 \\ \frac{\partial \mathcal{L}}{\partial e_k} = 0 & \rightarrow \quad \alpha_k = \gamma e_k \\ \frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 & \rightarrow \quad y_k[\mathbf{w}^T \varphi(\mathbf{x}_k) + b] - 1 + e_k = 0 \end{cases} \tag{4.41}$$

for $k = 1, ..., N$. These can be written as the linear system [29]

$$\left[\begin{array}{ccc|c} \mathbf{I} & 0 & 0 & -\mathbf{Z}^T \\ 0 & 0 & 0 & -\mathbf{Y}^T \\ 0 & 0 & \gamma\mathbf{I} & -\mathbf{I} \\ \hline \mathbf{Z} & \mathbf{Y} & \mathbf{I} & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \mathbf{w} \\ b \\ \mathbf{e} \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ \hline \mathbf{1} \end{array}\right] \tag{4.42}$$

where $\mathbf{Z} = [\varphi(\mathbf{x}_1)^T y_1 ... \varphi(\mathbf{x}_N)^T y_N]^T$, $\mathbf{Y} = [y_1 ... y_N]^T$, $\mathbf{1} = [1...1]^T$, $\mathbf{e} = [e_1 ... e_N]^T$,

$\alpha = [\alpha_1 ... \alpha_N]^T$. Elimination of $\mathbf{w}$ and e gives

$$\left[\begin{array}{c|c} 0 & \mathbf{Y}^T \\ \hline \mathbf{Y} & \mathbf{Z}^T\mathbf{Z} + \mathbf{I}/\gamma \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{1} \end{array}\right]. \tag{4.43}$$

The kernel trick is applied to the matrix $\Omega = Z^T Z$ with

$$\begin{array}{rcl} \Omega_{kl} & = & y_k y_l \, \varphi(\mathbf{x}_k)^T \varphi(\mathbf{x}_l) \\ & = & y_k y_l \, K(\mathbf{x}_k, \mathbf{x}_l), \;\; k, l = 1, ...N. \end{array} \tag{4.44}$$

The classifier in the dual space takes the form

$$y(\mathbf{x}) = \text{sign}\left[\sum_{k=1}^{N} \alpha_k y_k K(\mathbf{x}, \mathbf{x}_k) + b\right] \tag{4.45}$$

similar to the standard SVM case.

The LS-SVM classifier has the following properties:

- choice of kernel function:
  A chosen kernel function should be positive definite and satisfy the Mercer condition. All comments on the kernel functions equally well apply to the use of kernels in the LS-SVM context. However, in this thesis we will focus on the use of linear and RBF kernels.
- global and unique solution:
  The dual problem for linear and nonlinear LS-SVMs corresponds to solving a linear KKT system which is a square system with a unique solution when the matrix has full rank.
- lack of sparseness and interpretation of support vectors:
  As clearly shown from the condition for optimality

  $$\alpha_k = \gamma e_k,$$

  meaning that in this simplified formulation sparseness is lost. Every data point in the LS-SVM classifier is a support vector, because in general no α_k values will be exactly equal to zero.

4.8 Conclusions

Support Vector Machines (SVM) is a method of calculating the optimal separating hyperplane in the feature space. Optimal separating hyperplane is

defined as the maximum-margin hyperplane in the higher dimensional feature space.

The use of the maximum-margin hyperplane is motivated by statistical learning theory, which provides a probabilistic test error bound which is minimized when the margin is maximized.

The parameters of the maximum-margin hyperplane are derived by solving a quadratic programming (QP) optimization problem. There exists several specialized algorithms for quickly solving the QP problem that arises from SVMs.

The original SVM was a linear classifier. However, Vapnik suggested using the kernel trick (originally proposed by Aizerman [1]). In the kernel trick, each dot product used in a linear algorithm is replaced with a non-linear kernel function. This causes the linear algorithm to operate in a different space. For SVMs, using the kernel trick makes the maximum margin hyperplane be fit in a feature space. The feature space is a non-linear map from the original input space, usually of much higher dimensionality than the original input space. In this way, non-linear SVMs can be created. If the kernel used is a radial basis function, the corresponding feature space is a Hilbert space of infinite dimension.

Least Squares SVM (LS-SVM) simplifies the formulation by replacing the inequality constraint in SVM with an equality constraint. This approach significantly reduces the complexity and the computation times, solving a set of linear equations instead of solving the QP problem.

Chapter 5

Magnetic Resonance Spectroscopy

5.1 Introduction

Brain tumours are the second leading cause of cancer death in children under 15 years old and young adults up to the age of 34. These tumours are also the second fastest growing cause of cancer death among humans older than 65 years [110]. Early detection and correct treatment based on accurate diagnosis are important steps to improve disease outcome.

In this thesis, we investigate the following real life case study: automated brain tumour classification using Magnetic Resonance Spectroscopy. The first section will start with a brief description of brain tumours, followed by a description of brain tumour diagnosis and treatments.

Next we motivate the use of Magnetic Resonance (MR) techniques, especially MRS as a non-invasive method for brain tumour diagnosis to complement the use of MRI in clinical practice.

5.2 Brain tumours

The human body consists of building blocks called cells. During life, the body has the capability of regenerating its cells. The adult body only forms new cells to replace old or wear-and-tear damaged ones when they are needed. Infants and children form new cells to complete their development besides those needed

for repair. A tumour is formed when certain cells are generated when they are not needed.

The brain is made from two sorts of cells: **neurons** or **nerve cells** and supporting or **glial** cells. Through neurons, the brain sends messages to other parts of the body. Supporting (glial) cells on the other hand provide scaffolding for the neurons. The brain is covered by a series of membranes or linings (the meninges).

A brain tumour is a mass of unnecessary and abnormal cells growing in the brain. The *cell of origin* determines the type of the brain tumour. A tumour that starts in the brain is a primary brain tumour. Glioblastoma multiforme, astrocytoma, medulloblastoma and ependymoma are examples of primary brain tumours. If cells have spread to the brain from a cancerous tumour in another part of the body, they are known as secondary brain tumours or metastatic brain tumours. Brain tumours are usually located in the posterior third of the brain in childhood and in the anterior two-thirds of the brain in adulthood.

Some benign tumours grow for many years and reach a large size before being detected. Malignant tumours tend to grow more rapidly and will have been present for a shorter period when they are discovered. The onset of symptoms cannot be used to detect how long the tumour has been present or indicate whether it is cancerous or non-cancerous.

Primary brain tumours are named according to the cell types, from which they originated. Frequently encountered histologic brain tumour types are glioma, glioblastoma, astrocytoma, oligodendroglioma, medulloblastoma, meningioma and neuroglioma. Tumours can be benign and are usually but not necessarily localized in a small area, or they can be malignant and invasive (spreading to neighbouring areas). Brain cells can be damaged by tumour cells (i) directly when they are pushed or displaced due to growth of the tumour, (ii) indirectly when affected through ongoing inflammation in and around the tumour mass, (iii) via brain edema (swelling) or (iv) increased pressure in the skull (due to brain edema or to the blockage of the circulation of the cerebrospinal fluid).

The most common types of brain tumour are those which originate from malignant primary tumours arising somewhere else in the body and spreading out to the brain. Those secondaries or metastatic brain tumours are always malignant. Approximately, one quarter of metastatic cancers spread to brain. Lungs and breasts are most common locations from which secondary brain tumours originate. Tumour cells travel to the brain by blood vessels. Since the brain has no lymphatic drainage system like other organs (the cerebrospinal fluid system acts like a lymphatic system in the brain), spreading of tumour cells by the lymphatic route (which is very typical for cancers of other organs) is impossible. In contrast to primary brain tumours, metastatic tumour masses may occur in various remote locations in the brain. Highly aggressive brain

tumours like glioblastomas may also be observed in more than one location but this is usually the case in the advanced stages of the disease. Symptoms, diagnosis and treatment are quite similar to those of primary tumours. However in case of secondary tumours the initial location of the tumour cells must be identified and treated, as well.

Primary or secondary brain tumours may cause herniation of the brain (displacement of one part of the brain tissue due to mass effect of a lesion, usually causing the compression of the neurons controlling the respiratory system in the brainstem and eventually death) and permanent neurologic changes including intellectual decline.

5.3 Brain tumour diagnosis

Currently, patients will be sent to a neuroscience unit for investigation. The first scan is usually either a Magnetic Resonance Imaging (MRI) scan or a Computerised tomography (CT) scan.

MRI scan
Pictures of the head and brain are made by magnetic fields and radio waves. The person undergoing the scan lies in a long tube, and may have MRI contrast injected to enhance the tumour detection.

CT scan
Using a special type of x-ray, this scan takes pictures of the brain from different angles. Sometimes a dye or CT contrast is also injected into a vein to enhance the detection of the tumour.

EEG (Electroencephalography)
Sometimes an EEG may be carried out if the first symptom was a seizure. EEG is performed by attaching wires to the scalp for 20-30 minutes, during which time the electrical activity of the brain is recorded.

Abnormalities of the brain may be categorized according to their association with structural (i.e., anatomy-related), physiological (e.g., reduced or interrupted blood supply), or biochemical mechanisms. CT, MRI, and MR angiography are capable of evaluating structural and physiological alterations that are potentially responsible for brain diseases. Magnetic resonance spectroscopy provides means of assessing the biochemical aspects of tissue abnormalities.

5.4 Brain tumour treatment

There are three main types of brain tumour treatment:

1. Chemotherapy
2. Radiotherapy
3. Surgery

Chemotherapy uses drugs to destroy tumour cells, similar to the use of antibiotics to kill bacteria. This technique is mainly used for the treatment of malignant brain tumours in children. Adults may also receive chemotherapy at some point. There are many different types of chemotherapy. Some are taken orally, others are given via a drip into the veins. Although this therapy is not applicable in all cases, it is often powerful.

Chemotherapy is usually given by the radiation oncologist, who has expertise in both radiation and chemotherapy. In the case of children, a pediatric oncologist together with a radiotherapist may be in charge.

Side effects caused by chemotherapy vary considerably depending on the drug used. Some possible side effects are tiredness, nausea, hair loss and increased susceptibility to infection.

Radiotherapy is often used following a surgery, occasionally for treating benign tumours, but mostly used for treating malignant ones. This treatment is given by a radiotherapist who will also be an oncologist.

Radiotherapy uses focused high energy gamma rays to destroy the tumour. This therapy decelerates tumour growth and extends the time before it recurs. Radiotherapy may cure the tumour in some circumstances.

Common side effects are inflammation of the skin which may lead to temporary hair loss, or a temporary worsening of the neurological complaints. Most people feel tired towards the end of their treatment, which normally takes four to six weeks of daily radiotherapy.

Surgery may be performed for two different reasons (i) to make a diagnosis, (ii) as a treatment.

Brain scanning may be repeated for technical reasons to better plan a surgery. Occasionally an angiogram is performed to show the blood supplies of the suspected tumour. This involves injecting a dye through a cathether.

Biopsy is a procedure which takes away a small sample of tumour for pathological analysis. This operation aims at providing a diagnosis rather than treat-

ment. A preliminary diagnosis can also be made during the surgery. Biopsy is carried out through a *burr hole*, i.e. an opening in the skull with a diameter of about 1 cm. A needle is passed through this hole reaching the tumour, and takes a sample of it. The needle is usually guided to the target with the use of a scan and a special frame which is placed on the head. This is known as stereotactic biopsy. Recent advances allow this to be done without the frame in some cases, in which case it is called frameless stereotaxis. Inevitably, as with any surgical procedure, there are risks. The symptoms may be worse after biopsy or it may cause seizures.

Brain surgical treatment tries to remove the tumour entirely, without damage to the surrounding brain tissue. Although this is often possible for benign tumours, surgery of malignant tumours will likely invade the surrounding normal brain tissue, which must be left behind to avoid serious disability. Therefore, this surgery brings greater risks as those of a biopsy, since it involves a bigger operation.

5.5 Nuclear Magnetic Resonance (NMR)

Magnetic induction by nuclear spins was observed for the first time in 1946 independently by Edward Purcell [86] and Felix Bloch [7] (both of them received the Nobel prize for physics in 1952). They found that the nuclei of certain atoms, when placed under a magnetic field, absorbed energy in the radiofrequency range and re-emitted this energy during transition to their original state. This phenomenon was called Nuclear Magnetic Resonance (NMR).

This achievement opened a new field of research, leading to many important applications, such as molecular structure determination, dynamics studies both in the liquid and solid state [24], and magnetic resonance imaging [58, 65].

In 1946 Bloch introduced the concepts of relaxation times and introduced the so-called Bloch equations to describe the NMR phenomenon [6]. NMR was used for physical and chemical molecular analysis and rapidly progressed to become the most powerful non-destructive analytical method in chemistry.

Paul Lauterbur discovered that the introduction of gradients in the magnetic field made it possible to create two-dimensional images of structures that could not be visualized by other techniques. In 1973, he described how the addition of gradient magnets to the main magnet made it possible to visualize a cross section of tubes with ordinary water surrounded by heavy water. No other imaging method can differentiate between ordinary and heavy water.

Peter Mansfield utilized gradients in the magnetic field in order to more precisely show differences in the resonance. He showed how the detected signals

could be rapidly and effectively analysed and transformed to an image. This was an essential step in order to obtain a practical method. Mansfield also showed how extremely rapid imaging could be achieved by very fast gradient variations (so called echo-planar scanning). This technique became useful in clinical practice some years later.

In 2003 both Lauterbur and Mansfield received the Nobel prize in Medicine for these crucial achievements in the development of MRI in medical applications. It also needs to be mentioned that the first potential medical application of NMR was described by Raymond Damadian in 1971, who reported that the NMR water signals of malignant tumours were different from normal tissues.

5.6 Magnetic Resonance Imaging (MRI)

The MR image represents the MR characteristics of tissue at a predefined position in the human body. To create an image of a specific slice within the body and also to localise individual voxels (volume elements) within the slice, all voxels have to experience a unique magnetic field. In conventional MR imaging three different functions are used for this purpose: slice selective excitation, frequency encoding, and phase encoding. These can be accomplished by applying temporary gradients in the magnetic field.

In slice selection, protons in the imaging slice are selectively excited by means of a combination of a frequency selective RF pulse and a magnetic field gradient perpendicular to the imaging plane. The bandwidth of the RF pulse and the amplitude of the slice selection gradient determine the slice thickness. Spatial encoding with this slice can be done by frequency encoding and phase encoding. When a gradient is switched on, spins at different locations will experience different magnetic fields, and will present with different Larmor frequencies. This procedure is called *frequency encoding* and causes the resonance frequency to be proportional to the position of the spin. *Phase encoding* uses a gradient to induce a specific phase angle to the transverse magnetisation vector. This phase angle depends on the location of the individual spin and the duration and amplitude of the gradient. By applying frequency and phase encoding both the frequency and phase of the spins contain spatial information, which is recorded in the MR signal. In the process of MR imaging gradients with different strengths and/or durations are applied consecutively, and the MR signals recorded are stored in a raw data matrix called k-space. This data contains all spatial information and can be processed using Fourier transformation to produce an image.

5.7 Magnetic Resonance Spectroscopy (MRS)

Although MRI has proven to be an indispensable tool for clinicians to detect the presence of various diseases, it is often nonspecific in defining the underlying pathology. In many cases, diagnosis must await microscopic evaluation of tissue obtained by biopsy. On the other hand, Magnetic Resonance Spectroscopy (MRS) can achieve a high degree of diagnostic specificity, since it is able to detect the biochemical changes that accompany specific diseases. MRS is the medical name for NMR spectroscopy. MRS attempts to apply the rigorous and quantitative approaches of NMR to the characterization, diagnosis and monitoring of disease and conditions in humans [96].

5.7.1 Basic principles

Certain nuclei, such as ^{1}H or ^{31}P possess an inherent angular momentum or spin. This can be thought as small spinning spheres. Since nuclei bear electric charges, their spinning produces a magnetic field similar to the field produced by electric current flowing through a coil of wire. This results in a magnetic dipole moment oriented along the spinning axis. The nucleus of the hydrogen atom (^{1}H) consists of a proton and possesses a magnetic moment. Because of its high natural abundance, sensitivity, and presence in water and fat in the human body, it is the most important nucleus in medical MR.

^{1}H or ^{31}P both have a spin quantum number of $\frac{1}{2}$, corresponding to two different spin or energy states and energy difference between both states is proportional to the applied magnetic field

$$\Delta E = \gamma \hbar B_0, \ \hbar = h/(2\pi), \tag{5.1}$$

with γ is the gyromagnetic ratio, and h is the Planck constant. To induce transitions between these energy states, an oscillating magnetic field is applied in a plane perpendicular to the static field B_0. The magnetic moment does not align exactly with the axis of the external magnetic field but precesses around it at a rate given by the Larmor relationship

$$f_0 = \gamma B_0/(2\pi). \tag{5.2}$$

The amount of energy absorbed by the nuclei (and emitted in a later stage) depends on the population difference between the two energy states. The population density of the nuclei in the different energy states is determined by their Boltzmann distribution. In thermal equilibrium at absolute temperature T, the relative amounts n^- of the nuclei in the highest and n^+ of the nuclei in the lowest energy states are given by

$$n^-/n^+ = e^{-\Delta E/kT}, \tag{5.3}$$

where k is the Boltzmann constant. The ratio n^-/n^+ is very close to unity in normal circumstances, resulting in a low absorption of energy. This explains the inherently low SNR of the signal emitted when the spin system returns to equilibrium after excitation. The emitted signal is called a *free induction decay* (FID) signal.

The intensity of the signal is proportional to the number of nuclei that contributes to it. In order to obtain a reasonable SNR, a number of data acquisition is performed consecutively and the final signal is the sum of all measured signals. As shown in Eqs. (5.1) and (5.3) increasing the field strength B_0 or decreasing the temperature T results in a higher energy difference between both states. This will also increase the amount of emitted energy and therefore gives signals with higher SNR. At the same field strength and temperature, the sensitivity of the ^{1}H nucleus is higher than that of the ^{31}P because ^{1}H has a larger γ.

Nuclei of different atoms can easily be distinguished from one another since they have a different γ and consequently resonate at other frequencies. Nuclei of the same isotope, but belonging to different molecules are differentiable due to the effect of shielding. The effective magnetic field experienced by a nucleus is

$$B_{\text{eff}} = B_0(1 - \sigma'), \tag{5.4}$$

where σ' is the shielding constant, a dimensionless unit that depends on the electrical environment of a nucleus. Therefore, nuclei in a different chemical environment emit signals with different frequencies. So, depending on its chemical environment, every nucleus has its frequency shift or *chemical shift*. All chemical shifts are expressed in relative terms. The shift, is expressed in parts per million (ppm) versus some reference compound. For ^{1}H and ^{13}C, this reference compound is normally tetramethylsilane (TMS), $(CH_3)_4Si$. A compound having a chemical shift of 1.00 ppm would show a peak 60 Hz downfield of TMS if run on a 60 MHz spectrometer, or 400 Hz downfield of TMS on a 400 MHz spectrometer. Chemical shift is thus independent of the instrument's field strength.

For ^{1}H spectra of most organic compounds, chemical shifts can range from 0-10 ppm. The spectrum is divided into several general regions, although there is a large amount of overlap.

The FID signal approaches exponential decay if the molecules are sufficiently mobile such that interactions are averaged out. However, the local field inho-

mogeneities introduced by variations in tissue homogeneity and operator shim introduce departure from this exponential decay.

In its simplest form a MRS experiment thus consists of applying a radio frequency (RF) pulse with a bandwidth chosen to excite all the nuclei within the required frequency range and acquiring the FID signal as the response. This process is repeated until the desired SNR is obtained.

5.7.2 Relaxation

When an excitation pulse is switched off, the spins start returning to their equilibrium. However, from their excited state, the nuclei can only return to the ground state by dissipating their excess energy to their surroundings. This process is called longitudinal (T_1) relaxation. T_1 relaxation is promoted by a surrounding nuclear environment with magnetic fields fluctuating at a frequency close to the Larmor frequency. The molecular environment of the nuclei determines the T_1 relaxation time.

To modify this relaxation time, MR contrast agents may be applied, Gadolinium chelates are often used for this purpose. Gadolinium has a large magnetic moment because of its unpaired electrons, providing a strong fluctuating magnetic field at the Larmor frequency for protons in a magnetic field of ~ 1.5 T. As a result, the T_1 will drop when protons get near the contrast agent.

While the longitudinal magnetisation recovers to its equilibrium value, the transverse net magnetisation decays to zero. Nuclear magnetic moments contributing to the transverse net magnetisation get out of phase because of their mutual interaction, a process called T_2 relaxation. All protons are affected by the magnetic fields of their neighbours, making them precess at slightly different Larmor frequencies. As a result, the magnetisation in the transverse plane begins to dephase, usually at a rate much faster than the longitudinal magnetisation recovers.

5.7.3 Sequences

Spectroscopy can be performed with either a single- or multivoxel technique. The single voxel method acquires data from an 8-cm^3 volume element, typically placed in the occipital gray matter and or the parietal white matter [35]. Brain scanning with TR=2000 ms / TE=35 ms with 128 acquisition averages takes about three minutes of scan time.

Two common pulse sequences are available [18]. Stimulated echo acquisition mode (STEAM) is the first widely used MRS pulse sequence. Another pulse sequence is point resolved spectroscopy (PRESS). STEAM can achieve a shorter echo times but at the expense of less signal-to-noise.

The echo time (TE) of MRS affects the obtained information. An acquisition with a short TE (e.g. 30 ms), metabolites with both short and long T_2 relaxation times are observed, meaning that the spectrum acquired contains the signals from most of the metabolites. While using a long TE (e.g. 135 ms), the spectrum only contains metabolites with a long T_2, producing a spectrum primarily *NAA* (N-acetyl-aspartate), *Cr* (creatine) and *Cho* (choline). Two examples of long and short echo time MRS spectra are given in Figure 5.1.

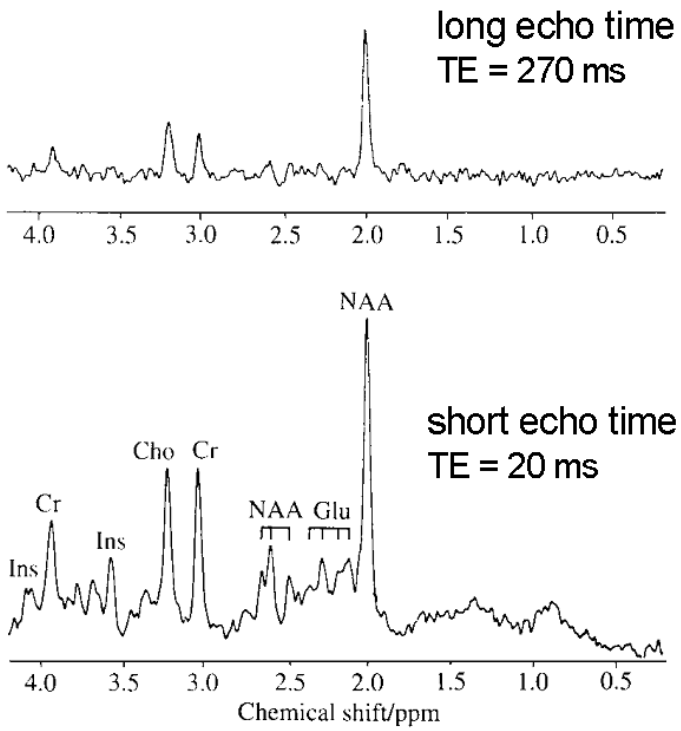


Figure 5.1: *Comparison of MRS spectra acquired by long echo time and short echo time. Much more metabolite peaks can be detected in short echo, however they strongly overlap and are often influenced by noise and artefacts.*

5.7.4 Mathematical model of MRS signal

A model function commonly used to represent the digitized FID signal is the Lorentzian model :

$$y_n = \bar{y}_n + e_n = \sum_{k=1}^{K} (a_k e^{j\phi_k})\, e^{(-d_k + j2\pi f_k)t_n} + e_n \quad n = 0, 1, ..., N-1, \quad (5.5)$$

K represents the number of different resonances, $j = \sqrt{-1}$. The amplitude a_k is related to the concentration of the metabolite and the frequency f_k characterizes it. The damping d_k provides, among others, information about its mobility and molecular information. The quantity $2\pi f_k t_0$ is also called the first-order phase. The noise term e_n is assumed to be circular complex white Gaussian noise (which means that the real and imaginary parts of the noise sequence are not correlated and have equal variance). It mainly consists of thermal noise from the sample and noise from the electronic components.

An alternative model function often used is the Gaussian model

$$y_n = \bar{y}_n + e_n = \sum_{k=1}^{K} (a_k e^{j\phi_k})\, e^{(-g_k t_n + j2\pi f_k)t_n} + e_n \quad n = 0, 1, ..., N-1. \quad (5.6)$$

And recently, a combination of both Lorentzian and Gaussian model, known as the Voigt model [67, 95] has received some attention

$$y_n = \bar{y}_n + e_n = \sum_{k=1}^{K} (a_k e^{j\phi_k})\, e^{(-d_k - g_k t_n + j2\pi f_k)t_n} + e_n \quad n = 0, 1, ..., N-1. \quad (5.7)$$

5.8 The role of MRS in brain tumour diagnosis

MRS studies have been shown to be of significant clinical value, particularly for evaluation of diseases that affect the brain. MRS has shown special advantages over MRI for the evaluation of brain tumours. For example, MRS can be used to :

- provide a better differential diagnosis,
- help to define the tumour grade, aggressivity, and relevant biochemistry of the lesion,
- monitor the tumour response to nonsurgical treatments,
- determine an earlier presence of tumour recurrence.

Currently, Magnetic Resonance Spectroscopy (MRS) in combination with Magnetic Resonance Imaging (MRI) are important tools to identify the location, size and type of brain tumours. So far, MRS has been proven to be an accurate non-invasive technique which can give detailed chemical information of metabolites present in the suspected brain tumours [74, 96]. Under physiological conditions, several important metabolites are observed: *NAA* as a neuronal marker; *Cho* as membrane precursors and degradation products; *Cr* as a measure of the energy status; *glucose*; and *mI* (myo-inositol). Under pathological conditions, the presence of some resonances can be indicative: a doublet of *Lac* (lactate); *lipids* and/or *some low molecular weight proteins* which might occur even under normal conditions; *Ace* (acetate) and certain amino acids, such as *Ala* (alanine), *Gln* (glutamine), *Glu* (glutamate) and *Gly* (glycine).

MRS can be used to determine the degree of malignancy. As a general rule, as malignancy increases, *NAA* and *Cr* decrease, while *Cho*, *Lac* and *lipids* increase. *NAA* decreases as tumour growth displaces or destroys neurons. Very malignant tumours have high metabolic activity and deplete the energy stores, resulting in reduced *Cr*. Very hypercellular tumours with rapid growth elevate the *Cho* levels. Lipids are found in necrotic portions of tumours, and lactate appears when tumours outgrow their blood supply and start utilizing anaerobic glycolysis.

Figure 5.2 illustrates an automated classification scheme of brain tumour classification to improve diagnosis. Some known observable proton metabolites of the brain are shown in Table 5.1 acquired by MRS. Table 5.2 shows some metabolite ratios which are commonly used to detect abnormality of the brain [40]. Figure 5.3 shows the advantage of MRS in acquiring metabolite information of a brain tumour.

5.9 Conclusions

NMR Spectroscopy has opened the way to investigation of brain metabolism. The most exciting aspect of MRS is that biochemical information can be obtained *non-invasively in vivo*. MRS can be performed in conjunction with MRI. Significant alterations in the metabolite pattern peaks can be observed in tumours. Advances in the MRS acquisition, together with the accurate classification of the data will significantly contribute to a better treatment of the brain tumour patient.

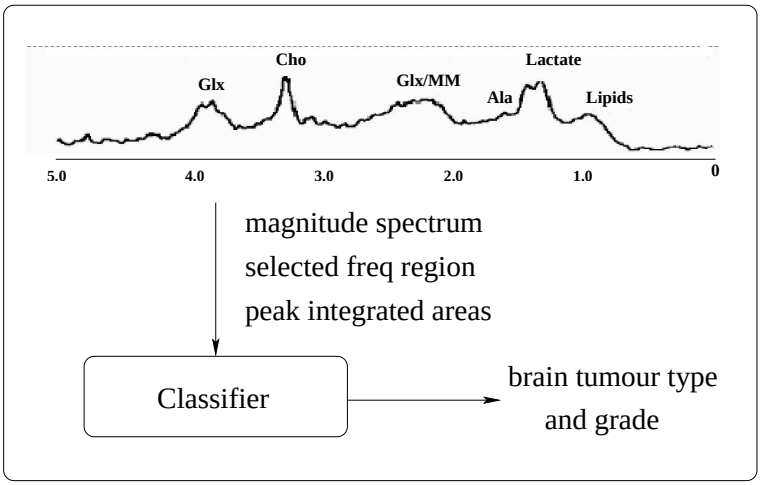


Figure 5.2: *An automated analysis of MRS requires reliable, robust and efficient automated classification for improved diagnosis.*

metabolite	properties
lipids	products of brain destruction
lactate	product of anaerobic glycolysis
NAA	neuronal marker
Glutamine/GABA	neurotransmitter
creatine	energy metabolism
choline	cell membrane marker
myo-inositol	glial cell marker
alanine	present in meningiomas

Table 5.1: *Observable proton metabolites using MRS. For further explanations, see [33].*

	Normal	Abnormal
NAA/Cr	2.0	< 1.6
NAA/Cho	1.6	< 1.2
Cho/Cr	1.2	> 1.5

Table 5.2: *Metabolite ratios using MRS commonly used to detect abnormality.*

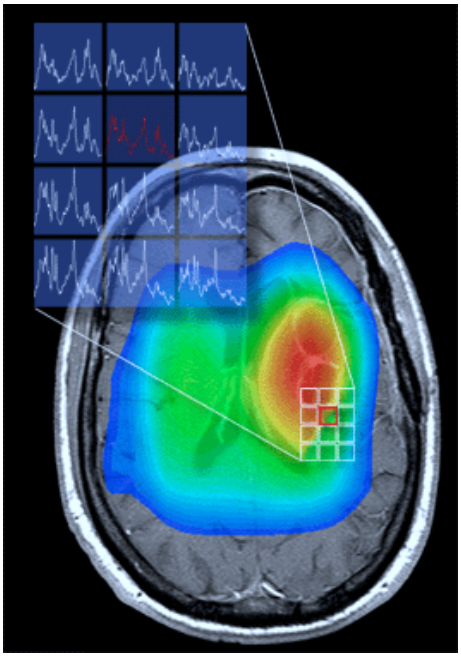


Figure 5.3: *Brain scans using MRS show the chemical characteristics of the brain tumour, and provide information on the speed of growth and the aggressivity of the tumour. (Image copyright by the Wallace-Kettering Neuroscience Institute)*

Chapter 6

Classification of brain tumours using long echo time MRS data

6.1 Introduction

This chapter reports the classification of brain tumours using long echo time Magnetic Resonance Spectroscopy. As explained in Chapter 5, MRS allows non-invasive diagnosis to the patient.

In comparison with *in vitro* spectroscopy, *in vivo* spectroscopy signals are more difficult to analyze because of their broader resonances, strongly overlapping peaks, lower signal-to-noise ratio and higher number of artefacts. Cousins [18] discusses the influence of the echo time TE on the spectral pattern of a MRS signal. The above mentioned metabolites can be detected in short echo ^{1}H MRS signals. However, short echo ^{1}H MRS signals are more difficult to analyze than long echo time ^{1}H MRS signals due to a higher number of overlapping peaks, a stronger baseline and a higher sensitivity to artefacts. In comparison, long echo time ^{1}H MRS signals are poorer in information but they allow a more reliable analysis and testing of classification methods.

Many studies have been performed to classify MRS signals. Lindon *et al.* [59] overviewed pattern recognition methods and their applications in biomedical magnetic resonance. Several studies [39, 94, 105, 107, 109, 126] also show some progress in automated pattern recognition for brain tumour classification

based on MR data. These studies are either based on MRI (e.g. [126]), MRI combined with MR Spectroscopy Imaging (MRSI) (e.g. [94]), long echo time (e.g. [39, 105, 107]) or short echo ^{1}H MRS (e.g. [109]), but most of the papers investigate only one classification method and restrict data collection to one center only. As performance measure either the training performance is considered or test performance on a specifically selected set. In our study we measure the binary classification performance based on the ROC curve analysis over 200 stratified random samplings of training and test set. This is in contrast to performance measures like the accuracy, e.g. used in [126], which is only based on a single cut-off point (e.g. for one specific value of the false-positive rate). Various clinical studies focus on the prediction of the malignancy of tumours, more specifically for brain gliomas (e.g. [39, 126]). Thereby, they consider only two classes: low-grade and high-grade gliomas. In our study, astrocytomas of grade II and glioblastomas (also called astrocytomas of grade IV) are included, which are large subtypes of respectively low-grade and high-grade gliomas. Additionally, we consider two other common brain tumour types, namely metastases and meningiomas.

Moreover, this thesis reports the results of a comparative study on a *multicenter* dataset of MRS signals. This dataset was developed in the framework of the EU funded INTERPRET project [45]. Several INTERPRET partners [52, 57, 62, 94, 106, 107, 108, 109] have already published results for classification of brain tumours based on MR data available within the project. The papers [57, 94, 106] focus on the use of ^{1}H MRSI data, while others consider the use of short or long echo time ^{1}H MRS. Nevertheless, most of these studies are based on a previous version of the dataset or focus on a specific technique. E.g. in [109], 144 short echo ^{1}H MRS spectra from 3 contributing centers were used, originating from three groups of brain tumours; meningiomas, low-grade astrocytomas and aggressive tumours. The latter group includes glioblastomas and metastases. Note that these groups correspond to the same four tumour groups as considered in this thesis. But Tate *et al.* selects a specific training and test set; the data from two centers formed the training set (94 spectra) and the data from the third center were used for testing (50 spectra). Based on this specific test set an accuracy of 96% was obtained using LDA.

In this study several methods are applied on all histopathologically validated long echo time ^{1}H MRS data from four common brain tumour types as available in the final status of the database development. We mention three additional points differing with previous classification studies within the framework of the INTERPRET project. First, we investigate what can be obtained as typical performance on a representative test set. Therefore, we construct 200 different combinations of training and independent test set. Second, the discrimination ability was judged by the AUC, which is, as opposed to the accuracy, a global measure. Only in one other INTERPRET study [57] ROC analysis was also applied to compare two diagnostic methods for classification based on ^{1}H MRSI.

Third, four different techniques are applied for classification, linear as well as nonlinear techniques. We investigate binary as well as multiclass classification. Moreover, this analysis includes optimal input variable selection and (hyper-) parameter estimation.

The chapter is organized as follows. Sections 6.2 explains the material used for the classification, while section 6.3 explains the applied classification techniques and the experimental setting. Section 6.4 summarizes the results of binary classification using complete spectra, selected frequency regions and peak integration values, consecutively. Afterwards, results of the multiclass classification approach are also mentioned. In section 6.5, we discuss the classification performance of the classifiers, the limitations of the dataset and the influence of dimensionality reduction. Finally, section 6.6 presents the conclusions.

6.2 Material

The data were provided by CDP (Centre Diagnòstic Pedralbes, Barcelona, Spain), IDI (Institut de Diagnòstic per la Imatge, Barcelona, Spain), SGHMS (St. George's Hospital Medical School, London, UK) and UMCN (University Medical Center Nijmegen, Nijmegen, The Netherlands) in the framework of the INTERPRET project. It concerns long echo time ^{1}H MRS data, acquired both with and without water suppression using a PRESS sequence (the repetition time TR is between 1500 and 2020 ms, the echo time TE=135 or 136 ms, the spectral width SW=1000 Hz or 2500 Hz, the number of data points is 512 or 2048).

Four main classes are considered, corresponding to four brain tumour types, i.e. glioblastomas, meningiomas, metastases and astrocytomas (grade II). They are labeled as class 1 (glio), class 2 (meni), class 3 (meta) and class 4 (astroII), respectively. All data have passed a quality control and validation process, which was regulated by strict rules agreed on by all INTERPRET partners. After thorough examinations, the brain tumours were histopathologically classified by three pathologists. These class assignments were based on the histological classification of Tumours of the Central Nervous System (CNS) set up by the World Health Organization (WHO).

The raw data are acquired in the time domain at the aforementioned centers. A few preprocessing steps are carried out: frequency alignment and phase correction with Klose's method [49] and filtering of the dominating residual water peak using Hankel Singular Value Decomposition (HSVD) [2]. The initial point of the time domain signal was removed, because it was often affected by artefacts. The resulting signal is transformed to the frequency domain by a Fast Fourier Transform (FFT). For each signal the L2-normalized magnitude spectrum (of unit length) is considered only in the frequency region of interest

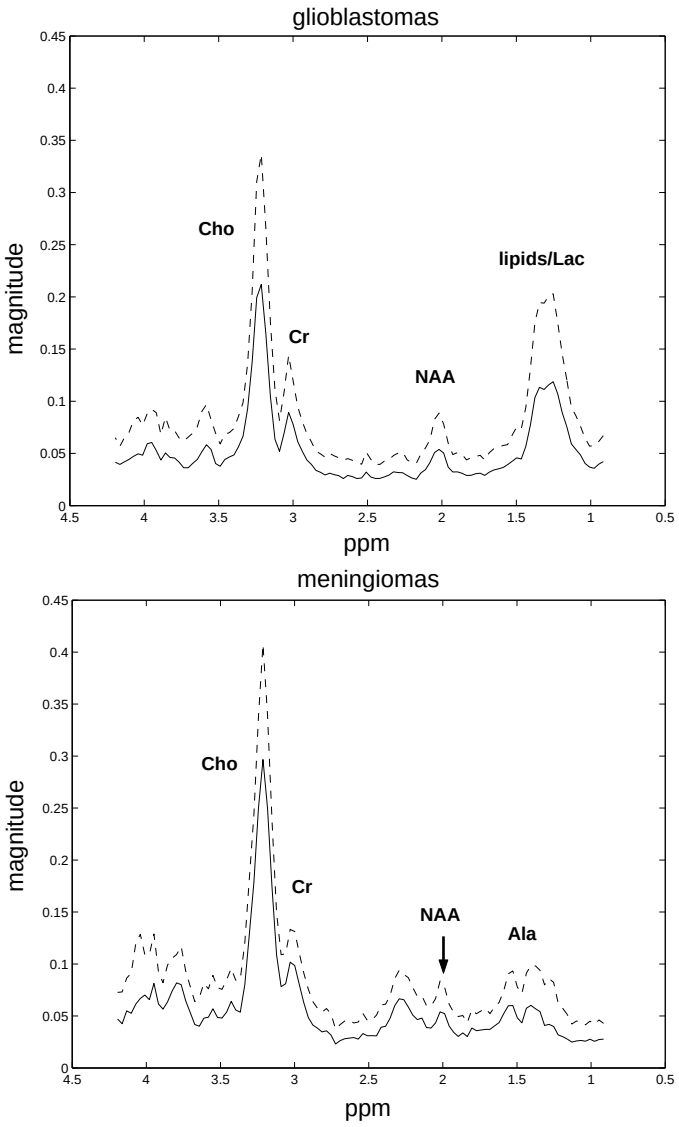


Figure 6.1: *Mean L2-normalized magnitude frequency spectra of the four considered classes: class 1 (top), class 2 (bottom) correspond to the glioblastomas, meningiomas, respectively. The solid lines are the means, while the dotted lines are the means plus the standard deviations of each class.*

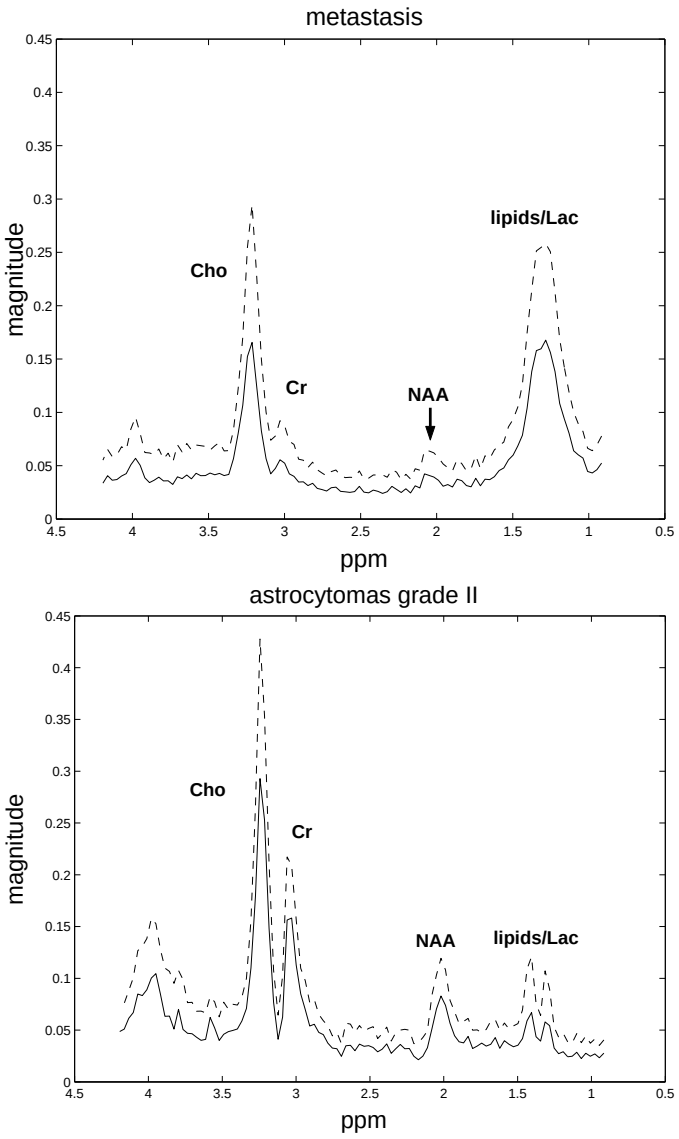


Figure 6.2: *Mean L2-normalized magnitude frequency spectra of the four considered classes: class 3 (top) and class 4 (bottom) correspond to metastases and astrocytomas (grade II) respectively. The solid lines are the means, while the dotted lines are the means plus the standard deviations of each class.*

Center (acquisition scheme)	1	2	3	4	total
CDP (PRESS, TE=135ms)	38	16	5	6	65
IDI (PRESS, TE=136ms)	28	27	16	6	77
SGHMS (PRESS, TE=136ms)	10	9	11	7	37
UMCN (PRESS, TE=136ms)	1	1	0	2	4
total	77	53	32	21	183

Table 6.1: *Number of long echo time ^1H MRS data of glioblastomas (class 1), meningiomas (2), metastases (3) and astrocytomas grade II (4). The rows correspond to the acquisition center, while the columns mention the type of brain tumour. The acquisition scheme is a PRESS sequence and TE denotes the echo time.*

(4.17 to 0.94 ppm), corresponding to 108 input variables. Figures 6.1 and 6.2 depict the mean magnitude frequency spectra of the four considered classes.

6.3 Methods

Several classification techniques can be applied to separate the given MR spectra. The techniques that we apply in this chapter are chosen so that we consider linear as well as nonlinear methods: LDA, SVM and LS-SVM. A detailed description of these methods is given in Chapters 3 and 4.

Linear Principal Component Analysis (PCA) is applied here to select the input variables. It reduces the 108 given spectral variables to a minimal set of variables which cover 75% variance of the data.

Quite often, different classes do not have equally distributed data points and their distributions are also overlapping among classes, which causes the problem to be linearly nonseparable. Here, two kernel based classifiers SVM and LS-SVM as explained in Chapter 4 are assessed.

The MRS spectra were classified using Steve Gunn's MATLAB Support Vector Machines toolbox http://www.isis.ecs.soton.ac.uk/isystems/kernel [34, 100] and KULeuven's MATLAB/C LS-SVMlab toolbox http://www.esat.kuleuven.ac.be/sista/lssvmlab [60, 82, 102] for LS-SVM classification with both linear and RBF kernels.

6.3.1 Selected frequency regions

It is well known that characteristic peaks at certain frequencies correspond to important metabolites in the brain [44, 55, 64, 74, 75, 96]. These peaks might

be used as discriminatory features to distinguish tumour types. In particular when their appearances clearly differ in size and shape between spectra of different tumour types. Instead of using complete spectra as input variables to the classifier, selection of the most explanatory input features can be used. One approach is based on selected frequency regions: therefore, the input variables within certain regions of the magnitude spectrum which are assumed to contain most of the information as input features are selected. Hence, the redundancy produced by spectral noise and artefacts in the spectrum is reduced. Characteristic metabolites can be observed in the following regions of the magnitude MRS spectrum: *Cho* and *Cr* (2.95 - 3.3 ppm); *NAA* (1.95 - 2.1 ppm); *Lac*, *Ala* and *lipid1* (1.15 - 1.55 ppm); *lipid2* (0.9 - 1.0 ppm). Note that these selected regions are based on the metabolites that are assumed to be most characteristic according to prior knowledge available from field experts participating in this study. Nevertheless this selection is still subjective as the size of the regions could be altered or some other resonances (e.g. from metabolites with a typically lower intensity at a long echo time (e.g. *mI, Gln, Gly,...*) could also have been included.

6.3.2 Peak integration

Another approach to select the most explanatory input is based on peak integration [41]. The amplitude of a resonance is proportional to the integral of the corresponding peak in the spectrum. However, precise estimation of the peak integrals is difficult due to several factors, including nonzero baseline, peak overlap, noise and also the discrete nature of the spectrum. Peak integration is performed here by using the trapezoidal rule. For each selected metabolite the area under the frequency peak in the magnitude spectrum is calculated. These regions cover: *Cho* (3.1 - 3.3 ppm); *Cr* (2.95 - 3.05 ppm); *NAA* (1.95 - 2.1 ppm); *Lac* and *Ala* (1.25 - 1.55 ppm); and *lipid1* (1.1 - 1.25 ppm).

6.3.3 Training and test data

Binary classification

Binary classification can be used to distinguish two different tumour types. Instead of using a one-against-all scheme, the classes are pairwise compared by means of a binary classifier. Consider four types of brain tumours, then six binary classifiers can be constructed to separate the following pairs:

- 1 vs. 2 : glioblastomas vs. meningiomas,
- 1 vs. 3 : glioblastomas vs. metastases,
- 1 vs. 4 : glioblastomas vs. astrocytomas grade II,
- 2 vs. 3 : meningiomas vs. metastases,

- 2 vs. 4 : meningiomas vs. astrocytomas grade II, and
- 3 vs. 4 : metastases vs. astrocytomas grade II.

By classifying in pairs, we obtain more information about:

1. the distribution of two classes and their overlap,
2. the balance of the data distribution of the classes, and
3. the performance of the classifier which can be measured using ROC analysis.

The dimension of the input features to LDA is reduced by PCA. Note that PCA is not used when peak integration values are taken as input features, as peak integration already significantly reduces the dimension.

To achieve a high level of performance in SVMs, some hyperparameters must be tuned. These adjustable hyperparameters include: a regularization parameter, which determines the tradeoff between minimizing the training errors and minimizing the model complexity. In case of a RBF kernel, also a kernel parameter (the width σ) must be selected. We choose the value of hyperparameters c for SVM, γ for LS-SVM with a linear kernel and (σ, γ) for LS-SVM with a RBF kernel through leave-one-out (loo) cross validation, while bounding the search to avoid overfitting.

The experiment consists of the following steps:

1. the data are divided in a training set (2/3 of the data) and a test set (remainder) using stratified random sampling,
2. train the classifiers and use the test set to evaluate the performance,
3. the index of the misclassified spectra is noted.

This randomization is repeated 200 times to avoid bias possibly introduced by selection of a specific training and test set. In this way we try to obtain a representative performance on the test set. ROC [103] analysis is used to evaluate the binary classifiers. The performance is then measured by the mean AUC and its pooled standard error calculated from 200 randomizations.

Multiclass classification

Using a binary classifier, a new MRS spectrum is assigned to one of the two considered classes. Nevertheless, in medical practice, the number of possible

tumour types is mostly not restricted to two types. Therefore a multiclass scheme is developed to handle all classes in one construction, thereby extending the restricted use of the binary classifiers mentioned in the previous section. In this setup, we expect the classifier to assign a certain spectrum to one of the four tumour types.

Numerous studies have been performed to distinguish MRS spectra of class 1 (glioblastomas) and class 3 (metastases), but to our knowledge none gives satisfactory results on an independent test set [83, 107]. Preul *et al.* [84] reported on a pattern recognition study, using LDA on data from six different brain tissue types: astrocytomas of grade II, of grade III and grade IV (glioblastomas), meningiomas, metastases and non-tumourous brain tissues (note these include our classes 1 and 3). Six metabolite intensity ratios (relative to the *Cr* resonance intensity in normal brain), obtained from long echo time ^{1}H MRSI (TE=272ms), were used as input features: *Cho*, *Cr*, *NAA*, *Ala*, *Lac* and lipids. They obtained a leave-one-out performance of 104 of 105 correctly classified spectra. Nevertheless, this does not necessarily imply that a similar result would be obtained on an independent test set.

Tate *et al.* [109] suggested to merge the two classes 1 and 3, resulting into a new group of aggressive (aggr) tumours, labelled as class 5. As shown in Figure 6.3 on the left part, a scheme to discriminate the classes 2, 4 and 5 is depicted as step one. A voting scheme is applied to decide which class is chosen based on the three outputs of the contributing binary classifiers. A certain class is taken if two of the binary classifiers give the same output, otherwise the classifier considers the output as *undecided*. Step two is carried out, as illustrated in the right part: if the output of step one is class 5, then the spectrum is further classified either into class 1 (glioblastomas) or 3 (metastases) using the binary classifier 1 vs. 3. The output of step two does not change if the voting scheme from step one gives class 2 or 4 as output. Four binary classifiers are the building blocks of this multiclass classifier: binary classifiers 1 vs. 3 and 2 vs. 4 are available from the previous section, and two additional binary classifiers are required:

- meningiomas vs. aggressive tumours (class 2 vs. class 5),
- astrocytomas of grade II vs. aggressive tumours (class 4 vs. class 5).

6.3.4 Statistical analysis

From 200 runs, the mean AUCs ($\overline{\text{AUC}}$) are listed in the tables, as well as the standard error SE on the AUC. For each binary classifier C, the mean and standard error of the AUC are calculated as explained in Section 2.2.3.

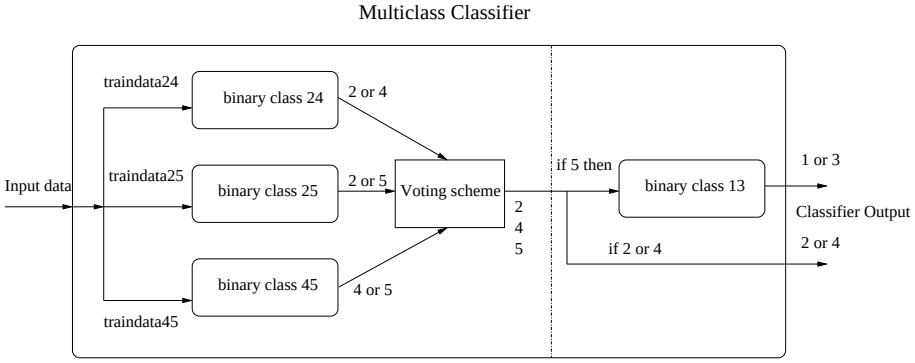


Figure 6.3: *Two-steps classification. The left part shows step one, classification of three tumour classes: meningiomas (class 2), astrocytomas of grade II (4), and aggressive tumours (5). The right part, or step two, further refines the classification if the output is class 5 and assigns the spectra of this class either to glioblastomas (class 1) or metastases (3).*

6.4 Results

6.4.1 Classification using complete spectra

In the following the classification performance of LDA, SVM, and LS-SVM (using linear and RBF kernels) are reported. The result using the complete spectra are summarized in Table 6.2, while Figures 6.4-6.6 show the boxplots corresponding to the same cases. Note that the boxplots display the median of the AUC values and the Interquartile Range (IQR, the distance between the 75 th percentile and the 25 th percentile), while the tables display the mean and standard error of the AUC values. The latter can be used to calculate the z-value (section 2.2.3).

In the tables, PCA/LDA means that PCA is applied to reduce the input dimension before classification using LDA. SVM lin denotes the linear classification using SVM with linear kernel, and similarly for LS-SVM lin, classification using LS-SVM with the linear kernel. LS-SVM RBF denotes the LS-SVM classifier using the nonliner kernel (radial basis function).

6.4.2 Classification using selected frequency regions

By selecting the values within specific frequency ranges in the spectra, the number of data points is reduced from 108 to 30. For the LDA classifier, PCA is applied to further reduce this input dimension, covering at least 75% of the variance. These input variables are different from those obtained for

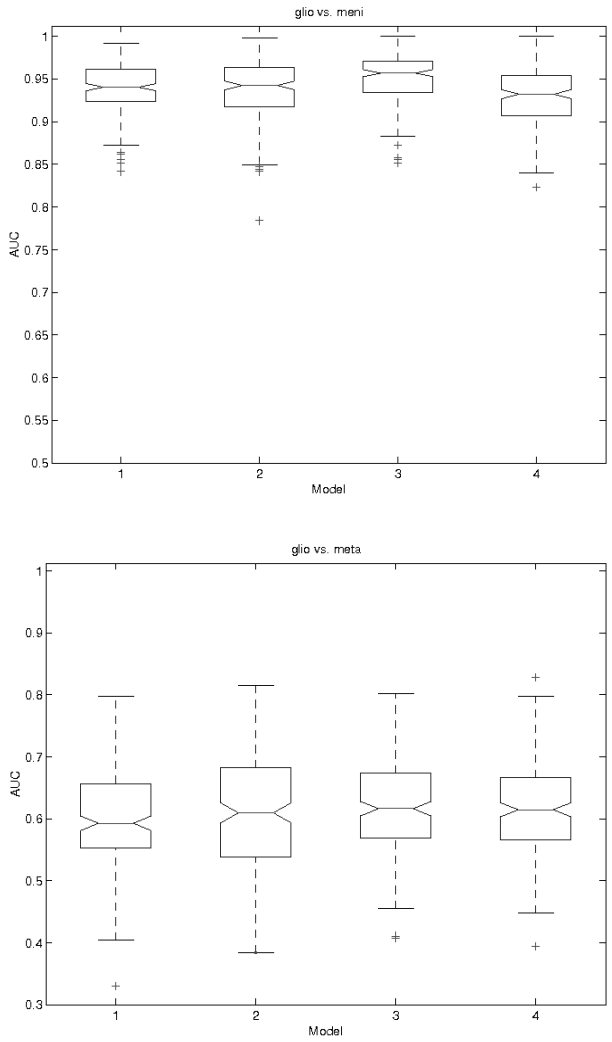


Figure 6.4: *Boxplots of the Area under ROC curves (AUC) on 200 stratified randomly sampled test sets of the four models (1) PCA-LDA, (2) SVM with Linear kernel, (3) LS-SVM with Linear kernel, (4) LS-SVM with RBF kernel. The figures correspond to the binary classifiers using complete spectra: glio vs. meni (top), glio vs. meta (bottom). Each box has lines at the lower quartile (lowest line), median (line in the middle), and upper quartile (highest line) values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers.*

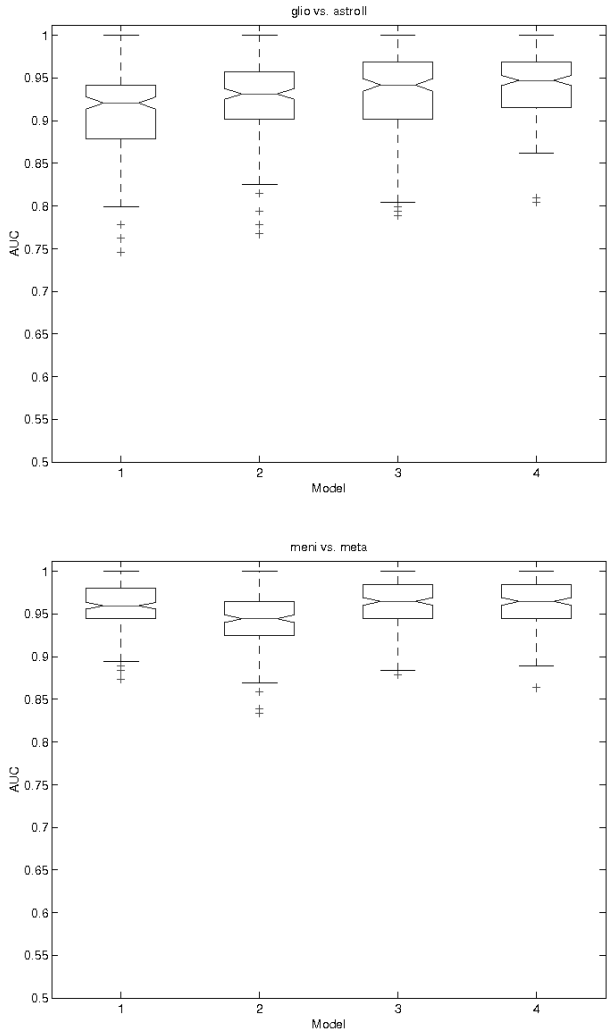


Figure 6.5: *Boxplots of the Area under ROC curves (AUC) on 200 stratified randomly sampled test sets of the four models (1) PCA-LDA, (2) SVM with Linear kernel, (3) LS-SVM with Linear kernel, (4) LS-SVM with RBF kernel. The figures correspond to the binary classifiers using complete spectra: glio vs. astroII (top). meni vs. meta (bottom). For further explanation about boxplot, see Figure 6.4.*

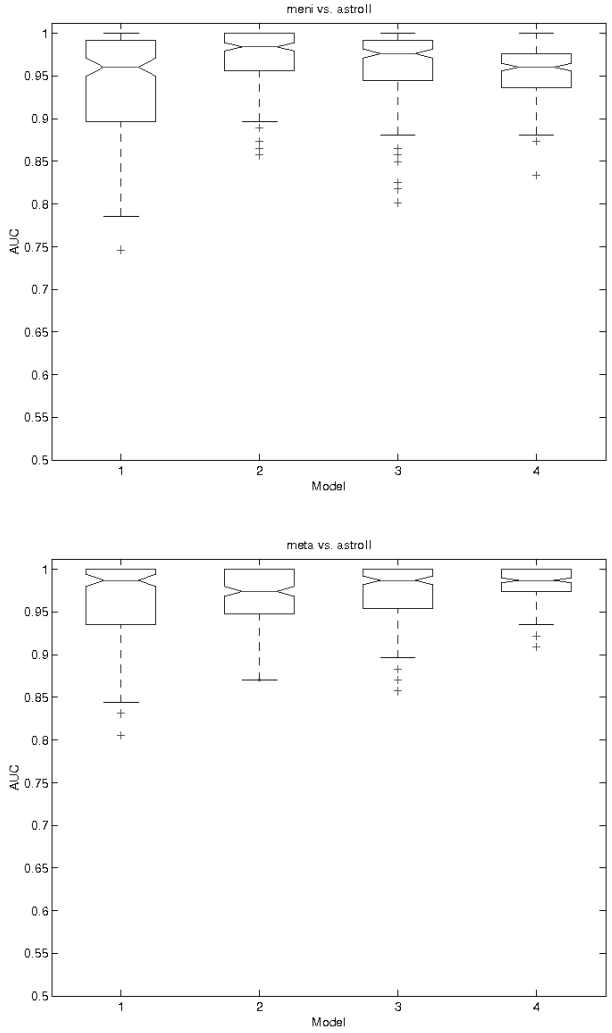


Figure 6.6: *Boxplots of the Area under ROC curves (AUC) on 200 stratified randomly sampled test sets of the four models (1) PCA-LDA, (2) SVM with Linear kernel, (3) LS-SVM with Linear kernel, (4) LS-SVM with RBF kernel. The figures correspond to the binary classifiers using complete spectra: meni vs. astroII (top) and meta vs. astroII (bottom). For further explanation, see Figure 6.4.*

the complete spectrum, due to the higher degree of freedom in the latter case. The results of 200 runs of stratified random samplings of the L2-normalized magnitude MRS spectra are shown in Table 6.3 and Figures 6.7-6.9.

6.4.3 Classification using peak integration

Table 6.4 and Figures 6.10-6.12 show the results of the ROC analysis for classification using peak integration. Five peak integration values are used as input of the classifiers. The linear classifier LDA is used without applying PCA.

6.4.4 Multiclass approach

As mentioned above, two additional binary classifiers are constructed by merging glioblastomas and metastases into one class of aggressive tumours. Table 6.5 shows the performance of these classifiers using the complete spectra as input.

Training performance

One way to train the multiclass classifier is by feeding all the spectra to the classifier and train each binary classifier with the corresponding classes. For example, use the spectra of class 2 and class 4 to train the binary classifier 24, and similarly for the others.

Table 6.6 shows a comparison of the multiclass classifier performance. The first row shows the percentage of correctly classified spectra in the class of *meningiomas*, *astrocytomas grade II* and *aggressive tumours*. One undecided case arose when using PCA/LDA with complete spectra classification, fifteen when using PCA/LDA and one when using LS-SVM classification both with the selected frequency regions as input variables.

In the second step, we use classifier 13 to further subclassify the *aggressive class*. Using this subclassification, the multiclass classifier's performance is shown in Table 6.7.

Test performance

Besides using all the spectra to choose the hyperparameters and to train the classifiers, one can also select 2/3 of the dataset as training set and use the remainder as test set. This stratified random sampling is repeated for 200 runs. The results are shown in Table 6.8 for one-step classification, which assigns the spectra to one of the 3 following classes: 2, 4 or 5. Table 6.9 shows the classifier performance after two-steps classification, which assigns the spectra to one of the 4 following classes: 1, 2, 3 or 4. Each spectrum of class 5 in step one, is either assigned to class 1 or class 3 in step two. In the Tables 6.8 and 6.9 we

Classes	PCA/LDA	SVM lin	LS-SVM lin	LS-SVM RBF
1 vs. 2	$0.9528 \pm 0.0306_{(8)}$	0.9519 ± 0.0335	0.9506 ± 0.0338	0.9560 ± 0.0304
1 vs. 3	$0.5926 \pm 0.1036_{(6)}$	0.6323 ± 0.0942	0.6431 ± 0.0983	0.5851 ± 0.1037
1 vs. 4	$0.9180 \pm 0.0627_{(7)}$	0.9159 ± 0.0565	0.9351 ± 0.0524	0.9385 ± 0.0486
2 vs. 3	$0.9605 \pm 0.0375_{(5)}$	0.9642 ± 0.0337	0.9711 ± 0.0307	0.9701 ± 0.0306
2 vs. 4	$0.9313 \pm 0.0725_{(10)}$	0.9661 ± 0.0390	0.9581 ± 0.0482	0.9595 ± 0.0456
3 vs. 4	$0.9612 \pm 0.0533_{(4)}$	0.9695 ± 0.0418	0.9740 ± 0.0393	0.9721 ± 0.0377

Table 6.2: *Classification using complete spectra. Average performance on the test set from 200 runs of stratified random samplings of the L2-normalized magnitude MRS spectra. As performance measure we use the mean AUC and its pooled standard error SE. The number between the brackets mentions the number of principal components used.*

Classes	PCA/LDA	SVM lin	LS-SVM lin	LS-SVM RBF
1 vs. 2	$0.7643 \pm 0.0722_{(2)}$	0.8532 ± 0.0575	0.8922 ± 0.0494	0.9187 ± 0.0413
1 vs. 3	$0.6381 \pm 0.1004_{(2)}$	0.5081 ± 0.1044	0.6368 ± 0.0998	0.5576 ± 0.1030
1 vs. 4	$0.8319 \pm 0.0776_{(2)}$	0.8692 ± 0.0713	0.8849 ± 0.0660	0.9012 ± 0.0594
2 vs. 3	$0.9212 \pm 0.0525_{(2)}$	0.9098 ± 0.0594	0.9339 ± 0.0475	0.9534 ± 0.0374
2 vs. 4	$0.9079 \pm 0.0645_{(3)}$	0.9592 ± 0.0410	0.9619 ± 0.0422	0.9617 ± 0.0411
3 vs. 4	$0.9173 \pm 0.0689_{(2)}$	0.9459 ± 0.0549	0.9698 ± 0.0389	0.9642 ± 0.0429

Table 6.3: *Classification using selected frequency regions. Average performance on the test set from 200 runs of stratified random samplings of the L2-normalized magnitude MRS spectra. As performance measure we use the mean AUC and its pooled standard error SE. The number between the brackets mentions the number of principal components used.*

Classes	PCA/LDA	SVM lin	LS-SVM lin	LS-SVM RBF
1 vs. 2	0.8504 ± 0.0586	0.8561 ± 0.0577	0.8448 ± 0.0593	0.8677 ± 0.0550
1 vs. 3	0.6252 ± 0.1007	0.6236 ± 0.1005	0.6434 ± 0.1006	0.6264 ± 0.0988
1 vs. 4	0.8773 ± 0.0635	0.8916 ± 0.0571	0.8787 ± 0.0628	0.8818 ± 0.0631
2 vs. 3	0.9103 ± 0.0628	0.9113 ± 0.0618	0.9191 ± 0.0585	0.9357 ± 0.0473
2 vs. 4	0.8441 ± 0.0858	0.8297 ± 0.0926	0.8485 ± 0.0851	0.8281 ± 0.0921
3 vs. 4	0.9592 ± 0.0461	0.9727 ± 0.0376	0.9597 ± 0.0453	0.9521 ± 0.0528

Table 6.4: *Classification using peak integration values. Average performance on the test set from 200 runs of stratified random samplings of the L2-normalized magnitude MRS spectra. As performance measure we use the mean AUC and its pooled standard error SE.*

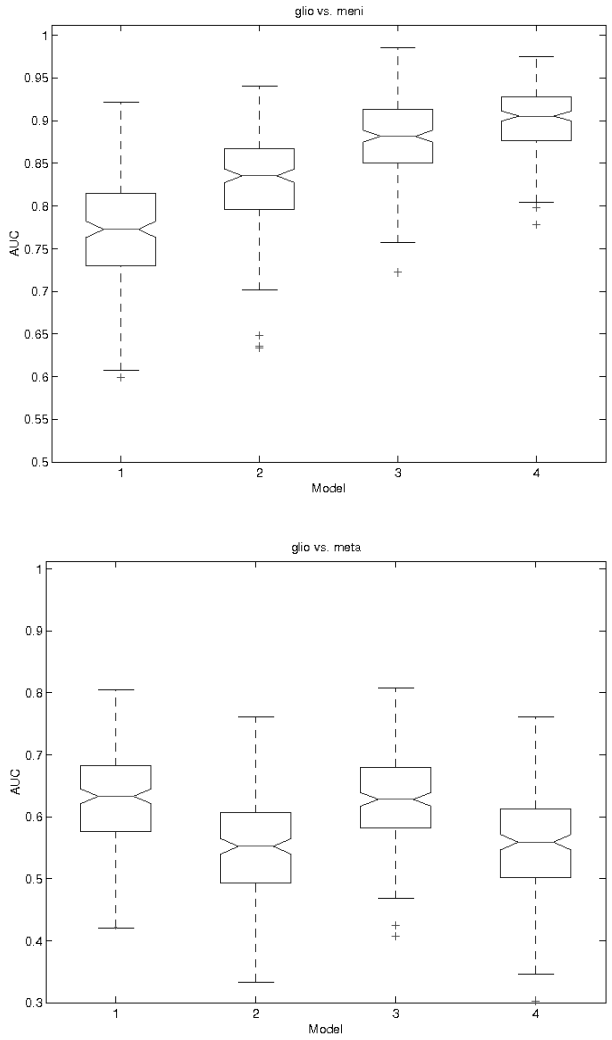


Figure 6.7: *Boxplots of the Area under ROC curves (AUC) on 200 stratified randomly sampled test sets of the four models (1) PCA-LDA, (2) SVM with Linear kernel, (3) LS-SVM with Linear kernel, (4) LS-SVM with RBF kernel. The figures correspond to the binary classifiers using selected frequency regions: glio vs. meni (top), glio vs. meta (bottom). For further explanation, see Figure 6.4.*

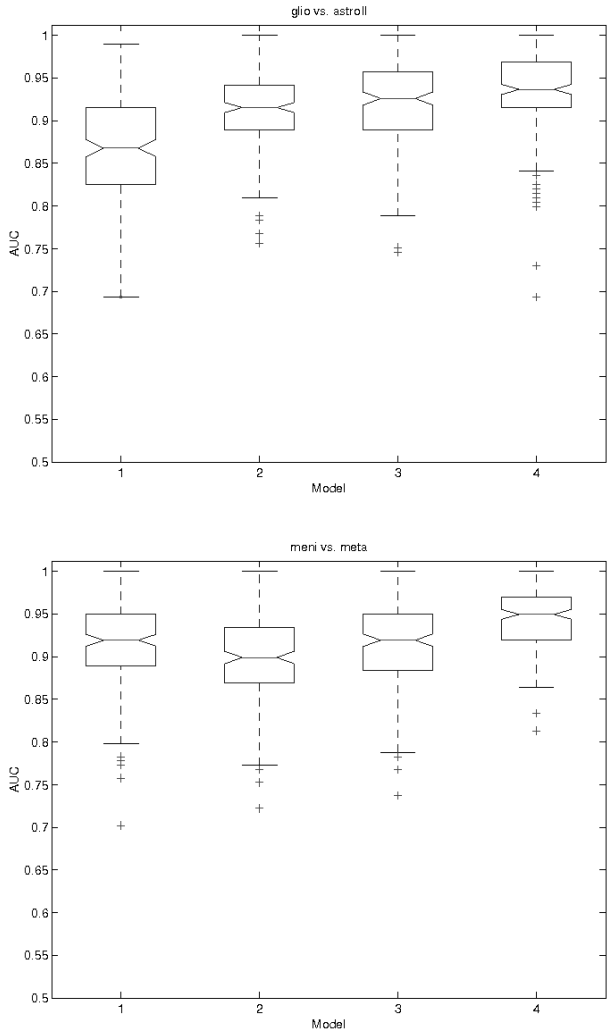


Figure 6.8: *Boxplots of the Area under ROC curves (AUC) on 200 stratified randomly sampled test sets of the four models (1) PCA-LDA, (2) SVM with Linear kernel, (3) LS-SVM with Linear kernel, (4) LS-SVM with RBF kernel. The figures correspond to the binary classifiers using selected frequency regions: glio vs. astroII (top), meni vs. meta (bottom). For further explanation, see Figure 6.4.*

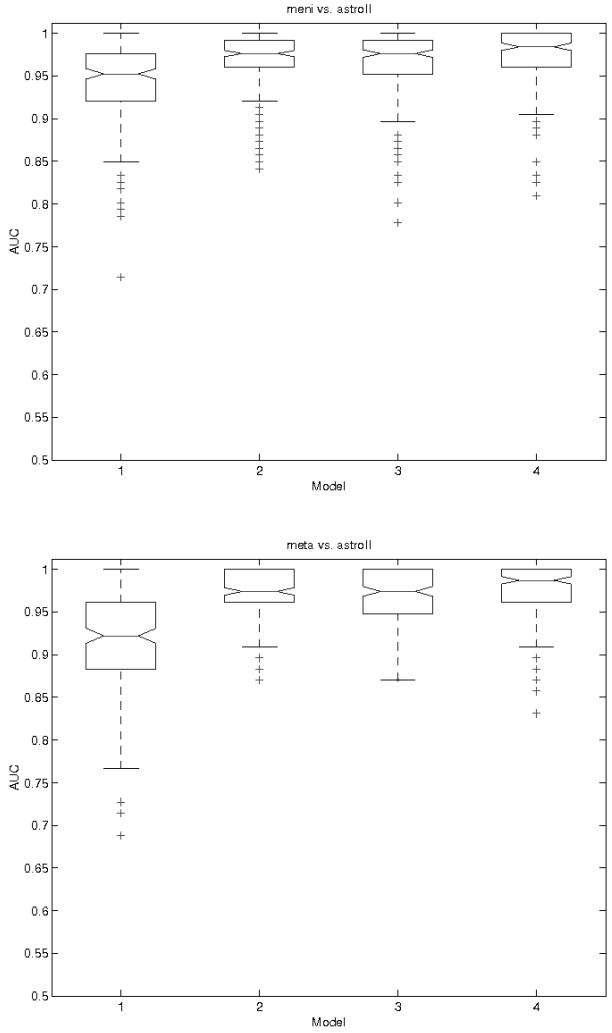


Figure 6.9: *Boxplots of the Area under ROC curves (AUC) on 200 stratified randomly sampled test sets of the four models (1) PCA-LDA, (2) SVM with Linear kernel, (3) LS-SVM with Linear kernel, (4) LS-SVM with RBF kernel. The figures correspond to the binary classifiers using selected frequency regions: meni vs. astroII (top) and meta vs. astroII (bottom). For further explanation, see Figure 6.4.*

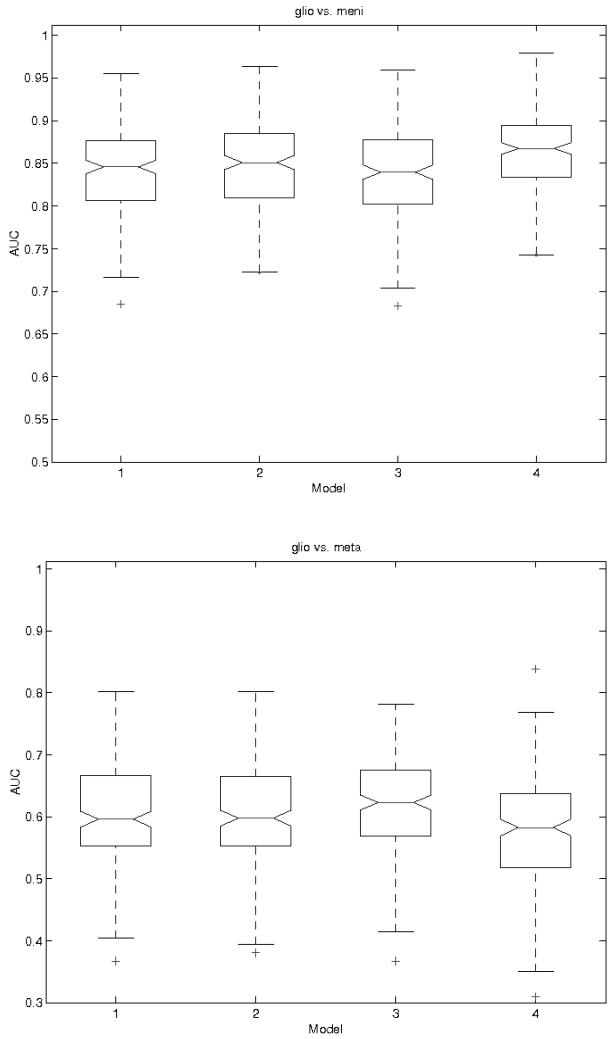


Figure 6.10: *Boxplots of the Area under ROC curves (AUC) on 200 stratified randomly sampled test sets of the four models (1) LDA, (2) SVM with Linear kernel, (3) LS-SVM with Linear kernel, (4) LS-SVM with RBF kernel. The figures correspond to the binary classifiers using peak integration: glio vs. meni (top), glio vs. meta (bottom). For further explanation, see Figure 6.4.*

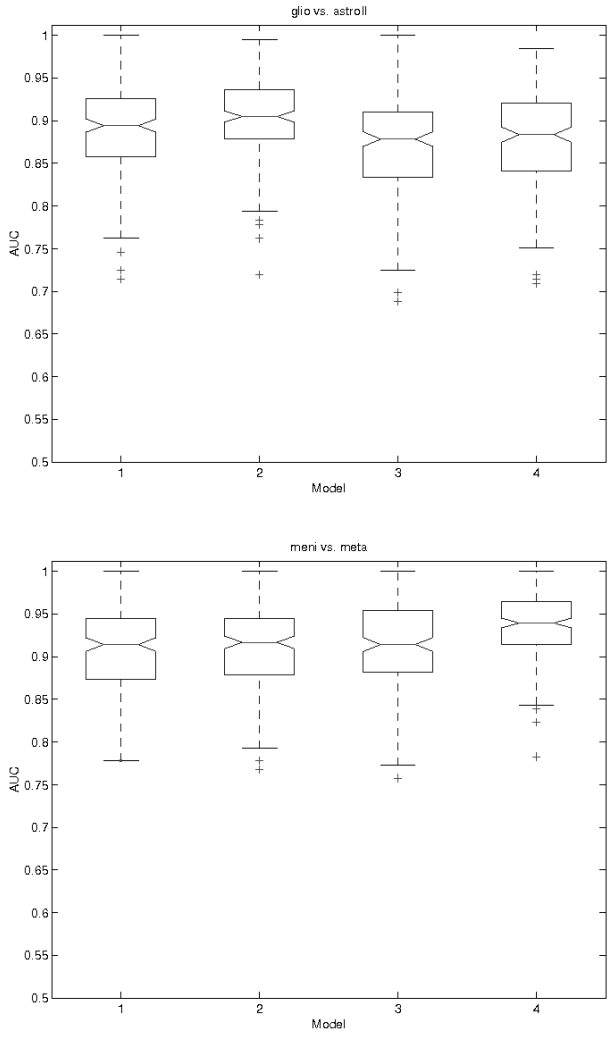


Figure 6.11: *Boxplots of the Area under ROC curves (AUC) on 200 stratified randomly sampled test sets of the four models (1) LDA, (2) SVM with Linear kernel, (3) LS-SVM with Linear kernel, (4) LS-SVM with RBF kernel. The figures correspond to the binary classifiers using peak integration: glio vs. astroII (top), meni vs. meta (bottom). For further explanation, see Figure 6.4.*

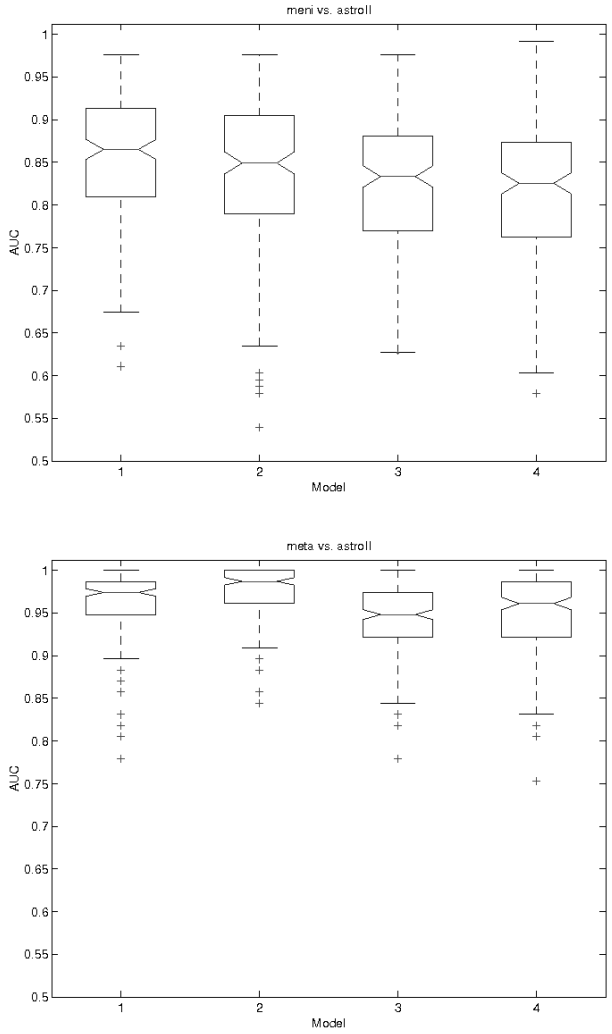


Figure 6.12: *Boxplots of the Area under ROC curves (AUC) on 200 stratified randomly sampled test sets of the four models (1) LDA, (2) SVM with Linear kernel, (3) LS-SVM with Linear kernel, (4) LS-SVM with RBF kernel. The figures correspond to the binary classifiers using peak integration: meni vs. astroII (top) and meta vs. astroII (bottom). For further explanation, see Figure 6.4.*

Classes	**PCA/LDA**	**SVM lin**	**LS-SVM lin**	**LS-SVM RBF**
meni-aggr	$0.9433 \pm 0.0306_{(6)}$	0.9409 ± 0.0343	0.9620 ± 0.0279	0.9110 ± 0.1121
astroII-aggr	$0.9230 \pm 0.0674_{(6)}$	0.9343 ± 0.0458	0.9416 ± 0.0502	0.9129 ± 0.1137

Table 6.5: *Classification using complete spectra. Average performance on the test set from 200 runs of stratified random samplings of the L2-normalized magnitude MRS spectra. As performance measure we use the mean AUC and its pooled standard error SE. The number between the brackets mentions the number of principal components used.*

	PCA/LDA (%)	LS-SVM lin (%)	LS-SVM RBF (%)
complete spectra	84.6995	93.9891	97.8142
selected freq. region	65.0273	84.6995	90.1639
peak integration	75.9563	77.0492	80.8743

Table 6.6: *Percentage of correctly classified spectra using all L2-normalized magnitude MRS spectra to assess the training performance after one-step classification.*

	PCA/LDA (%)	LS-SVM lin (%)	LS-SVM RBF (%)
complete spectra	71.0383	78.1421	83.6066
selected freq. region	50.2732	68.8525	74.8634
peak integration	61.7486	62.2951	67.7596

Table 6.7: *Percentage of correctly classified spectra using all L2-normalized magnitude MRS spectra to assess the training performance after two-step classification.*

mention the mean correct classification rate, the mean misclassification rate and the mean percentage of undecided cases and their standard deviation. The correct classification rate is defined as the percentage of correctly classified spectra, while the misclassification rate is the percentage of misclassified cases.

6.5 Discussion

In this section we discuss various issues concerning the results we obtained using the available long echo time ^{1}H MRS data. We do not necessarily claim that these remarks generally hold for similar analyses on other data.

6.5.1 Limitations

MRS signals of brain tumours contain chemical information about metabolites characteristic for the type of tumour. Nevertheless, there are still some factors making it hard to construct a classifier which is able to discriminate between different brain tumours using MRS signals:

1. the limited number of available spectra per type of tumour (see Table 6.1). Especially the amount of available metastases and astrocytomas grade II is low. This makes it difficult to construct a classifier with a high generalization capacity.

2. the presence of noise and artefacts in the spectra. Even after elimination of the dominating water peak, remaining artefacts might affect important peaks in the spectra.

3. the large variances within each class and the overlap between spectra of different brain tumour types (see Figure 6.1 and 6.2). E.g. the mean spectra of glio and meta show a very similar characteristic pattern, which makes the discrimination between glio and meta a very difficult problem. This problem is also observed in [83, 107]. Further discussion about this is addressed below (section 6.5.2).

6.5.2 Glioblastomas versus metastases

Although we obtained a low performance for distinguishing glioblastomas (glio) and metastases (meta), there are indications that these tumour types might be separable based on MR. In [105] Szabo De Edelenyi *et al.* introduced the so called nosologic images, which is an approach to analyze ^{1}H MRSI data of brain tumours. It is a tool that assigns the spectroscopic data of each voxel in the spectroscopic image to a histopathological class. Classification was carried

	PCA/LDA (%)	LS-SVM lin (%)	LS-SVM RBF (%)
correct	80.1855 ± 4.2853	82.7823 ± 3.3449	83.5726 ± 3.5058
misclass	14.0887 ± 4.0665	13.6532 ± 3.1140	12.5565 ± 3.3290
undecided	5.7258 ± 2.6110	3.5645 ± 2.0870	3.8710 ± 2.1144

Table 6.8: *One-step classification using complete spectra. Average performance on the test set from 200 runs of stratified random samplings (2/3 of the data used for training, 1/3 for testing). The first, second and third rows give, respectively, the mean correct classification rate, the mean misclassification rate and the mean percentage of undecided cases, each with their standard deviation.*

	PCA/LDA (%)	LS-SVM lin (%)	LS-SVM RBF (%)
correct	63.1532 ± 4.7255	65.7984 ± 3.3449	66.8145 ± 3.5058
misclass	31.1210 ± 4.6858	30.6371 ± 3.1706	29.3145 ± 3.5954
undecided	5.7258 ± 2.6110	3.5645 ± 2.0870	3.8710 ± 2.1144

Table 6.9: *Two-steps classification using complete spectra. Average performance on the test set from 200 runs of stratified random samplings (2/3 of the data used for training, 1/3 for testing). The first, second and third rows give, respectively, the mean correct classification rate, the mean misclassification rate and the mean percentage of undecided cases, each with their standard deviation. Note that the number of undecided cases is equal to that for the one-step classifier.*

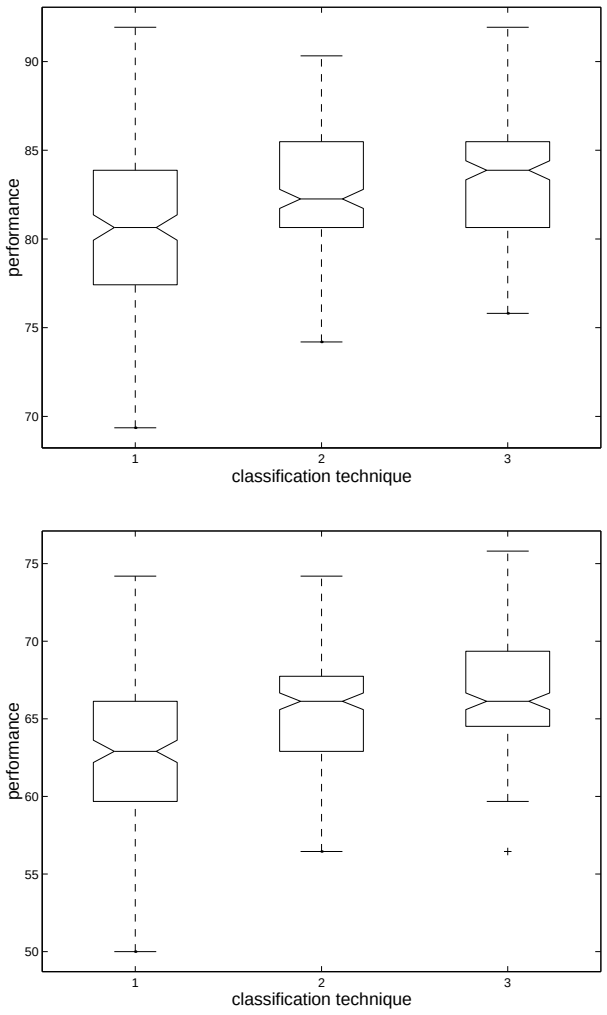


Figure 6.13: *Boxplots of the correct classification rate on the test set from 200 runs of stratified random samplings (2/3 of the data for training, 1/3 for testing) of the three models (1) LDA, (2) LS-SVM with Linear kernel, (3) LS-SVM with RBF kernel. The two figures correspond to the multiclass classifiers using complete spectra: output of step-one classifier (top), output of step-two classifier (bottom). For further explanation, see Figure 6.4.*

out by LDA applied on 6 metabolite values obtained from long echo time ^{1}H MRSI spectra (TE=272ms), together with the unsuppressed water area. Their study included 77 images, of which 24 high-grade gliomas and 10 metastases, for which they obtained a training performance of 87% following a leave-one-out (LOO) procedure. For the high-grade gliomas and metastases, respectively 19 and 6 spectra were correctly assigned.

Researchers [46, 53, 80] have found a few metabolite peaks or ratios which might contribute to the discrimination of high-grade gliomas and metastases. Law *et al.* concluded out of a study based on MRSI that, despite the small size of their dataset (11 high-grade gliomas, 6 metastases), the *Cho/Cr* ratio was significantly higher in high-grade gliomas than in metastases; this was the case for the tumoural region as well as the peritumoural region. Also based on perfusion-weighted MRI they have found different characteristics. Opstad *et al.* [80] have considered short echo ^{1}H MRS spectra (TE=30ms) from 25 glioblastomas and 34 metastases. Based on these data, they were able to find a significant difference in the ratio of the 1.3 ppm and the 0.9 ppm lipid/macromolecule peaks between these two tumour groups. This lipid peak area (LPA) ratio was 2.6 ± 0.6 for glioblastomas and 3.8 ± 1.4 for metastases ($P \leq 0.0001$). Based on ^{1}H MRS, Iishimaru *et al.* [46] have shown that the absence of *Cr* might indicate a diagnosis of metastasis, while in short echo the absence of lipids may exclude metastasis. In Figure 6.1 and 6.2 we do indeed notice a large mean lipid peak in metastases, but also in glioblastomas. This latter might be due to the occurrence of necrotic tissue in part of the glioblastomas. This partially explains the large variation we observe within this class and the similarity with the class of metastases.

6.5.3 Classification techniques

In general, LDA as linear classifier, preceeded by PCA (except for peak integration) performs quite well in solving the brain tumour classification problem. This is in correspondence with [109]. However, due to its linear boundary, overlapping classes are very difficult to handle.

As stated above, the small dataset available also forms a limitation for training. Therefore, the discrimination boundary will strongly correlate with the training set. Especially LDA requires a significant amount of data points to be able to draw a linear separating line between overlapping classes. In addition, it is possible that the separating line is very dependent on the selected training set.

Kernel based classifiers are less sensitive to the amount of data points; although the dimension is larger than the number of data points, these classifiers could draw an optimal separating boundary, without applying any dimensionality reduction (such as PCA). Kernel based classifiers, SVM and LS-SVM, feature the advantage of detecting automatically important characteristics independently of the input pattern.

Based on the statistical analysis described in section 2.2.3, no statistically significant difference was found between the AUC values for any of the classification techniques applied to the available long echo time ^{1}H MRS data. The highest z-value (1.72) was obtained, when comparing PCA/LDA with LS-SVM with a linear kernel based on the frequency selected regions; this is still lower than the cutoff value (1.96).

Also from visual inspection, we can not conclude that there is a clear difference between the considered classification techniques. In particular, the best performing technique depends on the considered classes and the type of input. To be more specific, when comparing the classification techniques we can group them in two ways (e.g. consider only the results with the complete spectra as input):

- linear (LDA, SVM lin, LS-SVM lin) vs. nonlinear techniques (LS-SVM RBF): in the cases 1 vs. 2 (glioblastomas vs. meningiomas), 1 vs. 4 (glioblastomas vs. astrocytomas grade II) the mean AUC values are slightly higher for the nonlinear technique, while in the other cases the AUC values are in the same range or slightly lower.
- LDA vs. kernel-based techniques (SVM lin, LS-SVM lin, LS-SVM RBF): in the cases 2 vs. 3 (meningiomas vs. metastases), 2 vs. 4 (meningiomas vs. astrocytomas grade II) and 3 vs. 4 (metastases vs. astrocytomas grade II) the kernel-based techniques perform slightly better.

Additionally, we can still consider the comparison of SVM vs. LS-SVM. Out of this we can only conclude that the best performing technique is also quite dependent on the case.

6.5.4 Influence of dimensionality reduction

For classification using selected frequency regions (section 6.4.2) and peak integration values (section 6.4.3) the input dimension is reduced by selecting only spectral regions which contain resonances of important metabolites. The underlying idea for this dimensionality reduction is to remove any redundant input features and reduce the influence of noise and artefacts. Hence, we try to enhance the discriminatory chemical information present in the spectra.

In contradiction with our expectations, we observe that the results in sections 6.4.2 and 6.4.3 on average are worse than in section 6.4.1. This seems to imply that this approach to dimensionality reduction also reduces part of the valuable information, that is present in the excluded frequency regions, which was important to explain the variance between the brain tumour classes. More specifically, here we are not considering resonances which typically have a low intensity at long echo times (because of a small T_2-value or cancellation due to

J-modulation) [25, 33]: *mI* (e.g. with triplets and multiplet at 3.26 and 3.57 ppm), *Glu* (e.g. multiplets at 2.33 ppm and 3.74 ppm), *Gln* (e.g. multiplets at 2.43 ppm and 3.75 ppm), *Gly* (singlet at 3.55 ppm). In Figure 6.1 and 6.2 we indeed notice a few small peaks around the specified resonances.

6.5.5 Multiclass classification

Multiclass classifiers handle all classes in one construction. We reduce this problem to a set of four binary classification problems, as explained in 6.3.3. Hence, we obtain four separating functions instead of one (one for each binary problem). Multiclass classifiers with the proposed scheme show a high learning capability. This is illustrated by the correct classification rates for the first step using the complete spectra as input: 84.7% (PCA/LDA), 93.9% (LS-SVM lin) and 97.8% (LS-SVM RBF). Given an independent test set as input, the classifiers on average give a quite good generalization performance: 80.2% (PCA/LDA), 82.8% (LS-SVM lin) and 83.6% (LS-SVM RBF).

In the second step of the multiclass classifier we combine the output of the first step with the binary classifier that separates glio and meta. The test performances reduce to 63.1% correct classification (LDA/PCA), 65.8% (LS-SVM lin), 66.8% (LS-SVM RBF). This can be explained by the most difficult binary case glio vs. meta. As observed in the binary classification, these two classes are very similar. Therefore separating them from one single class 5 into class 1 and class 3 deteriorates the total performance of the classifier.

Although no ROC analysis for multiclass classification was performed in order to test for significant differences between the classification techniques, the following indications can be noted, without drawing a general conclusion. The results, after step one as well as step two, yield a clearly higher training performance for the kernel-based methods than for LDA. Moreover, the kernel-based methods on average perform slightly better on an independent test set than LDA. This is clear from the mean percentage of correctly classified spectra which is a few percentages higher for LS-SVM (see Tables 6.8, 6.9 and Figure 6.13). Also the mean percentage of undecided cases differs slightly in favor of LS-SVM. This indicates that kernel-based methods can generalize at least as well as LDA based on a small dataset.

6.6 Conclusions

This chapter shows a comparative study of brain tumour classification based on long echo time ^{1}H MRS signals. Linear as well as nonlinear classifiers are compared. All techniques are applied automatically, including (hyper-) parameter selection, training and testing. Also for use in clinical practice, all techniques are easy to automate for analysis of independent data.

Binary classification gives more insight on the distributions of each class and their overlap. Except for the most difficult case glioblastomas vs. metastases, all classifiers based on the complete magnitude spectra are able to reach an AUC of more than 0.9. Based on the available data, we were not able to statistically prove any difference in performances between the classification techniques, for binary as well as multiclass classification. This indicates that kernel-based methods and LDA statistically perform as well for classification of brain tumours based on a small set of long echo time ^{1}H MRS data.

However, each of the applied techniques has its characteristics. LDA requires a prior dimensionality reduction of input variables (e.g. by applying PCA), while dimensionality reduction is done automatically in kernel-based methods.

We expected that dimensionality reduction, by selecting frequency regions or peak integration, would reduce the disturbing noise and artefacts in the spectra. However, the described approach for selecting resonance peaks of long echo time ^{1}H MRS spectra resulted in a lower performance. It might be necessary to include additional spectral information to increase classification performance. This also motivates further research in learning the peak pattern of short echo ^{1}H MRS, for which data are also provided within the INTERPRET project.

By using magnitude spectra, phasing problems are avoided. Nevertheless, with respect to real spectra, in magnitude spectra there occurs more peak overlap. Also, in real spectra at long echo time TE (TE=135, 136 ms) the peaks of Ala and Lac are inverted. This might reduce the ability to distinguish tumour types based on subtle differences in the spectral pattern. In order to test for this effect, in a future study also real spectra could be included as input features.

Discriminating aggressive tumour types, glioblastomas and metastases, using long echo time ^{1}H MRS spectra clearly is a very hard problem due to the highly similar pattern of the spectra from both classes, possibly due to the presence of some necrotic tissue. In order to address this problem, the discriminatory information present in the ^{1}H MRS spectrum should be enhanced, potentially by improvements in the acquisition of ^{1}H MRS signals. In particular, improvements are expected when processing short echo ^{1}H MRS signals, since more metabolites are visible in these spectra. Moreover, this spectral information is spread out over a larger number of peaks, thereby enlarging the number of possible discriminatory features. In the next chapter, we will report this experiment.

Chapter 7

Classification of brain tumours using short echo time MRS data

7.1 Introduction

This chapter presents the results of a comparative study on a multicentre dataset of short echo time ^{1}H MRS signals. In Chapter 6 we have compared the performance of automated binary classifiers and reached a mean area under the ROC curve (AUC) of more than 0.9 except for the most difficult binary classification to discriminate glioblastomas from metastases. In comparison with long echo time MRS spectra, short echo time MRS spectra are more difficult to analyze due to a higher number of overlapping peaks, a stronger baseline and a higher sensitivity to artefacts. However, short echo time MRS spectra are richer in information than long echo time MRS spectra, as several resonances vanish at longer echo times, because of a small T_2-value or cancellation due to J-modulation [26, 33]. For example, resonances of *mI* (myo-Inositol) (e.g. triplets and multiplet at 3.26 and 3.57 ppm), *Glu* (glutamate) (e.g. multiplets at 2.33 ppm and 3.74 ppm), *Gln* (glutamine) (e.g. multiplets at 2.43 ppm and 3.75 ppm) vanish into the baseline at longer echo times. In this chapter we report our results of an extended study based on short echo time MRS spectra, similar to [63]. We investigate whether the applied pattern recognition techniques are able to exploit the large amount of information available in short echo time MRS spectra, despite of the problems to analyse these data, in order to discriminate several types of brain tumours.

Linear as well as nonlinear classification techniques are applied to several input features, such as complete magnitude short echo time MRS spectra, or peak integration values obtained from these spectra. The algorithms are designed to extract the most important features which are then used to classify each spectrum into the corresponding tumour type.

The chapter is organized as follows. Section 7.2 gives a description of the data and its preprocessing, while section 7.3 explains the applied classification techniques and the experimental setting. Section 7.4 summarizes the results of binary classification using complete spectra, selected frequency regions and peak integrated values; comparing L2- with water normalization, magnitude and real spectra, with and without baseline correction. Results are then shown using a multiclass classification approach. In Section 7.5, we discuss the performance of binary as well as multiclass classifiers and the influence of dimensionality reduction. Finally, section 7.6 presents the conclusions.

7.2 Material

The short echo time INTERPRET database contains single voxel ^{1}H MR spectra from 6 centres: CDP (Centre Diagnòstic Pedralbes), Barcelona (Spain); FLENI (Fundación para la Lucha contra las Enfermedades Neurológicas de la Infancia), Buenos Aires (Argentina); IDI (Institut de Diagnòstic per la Imatge), Barcelona (Spain); MUL (Uniwersytet Medyczny w Łodzi), Łodzi (Poland); SGHMS (St George's Hospital Medical School), London (United Kingdom) and UMCN (Universitair Medisch Centrum Nijmegen), Nijmegen (The Netherlands). Table 7.1 mentions the most important acquisition parameters. Data are acquired either by a STEAM- or PRESS-sequence, at an echo time (TE) between 20 and 32 ms. Note that the echo time can differ somewhat from one centre to another, or even between data from the same centre. This range of TE typically will affect the appearance of coupled spins signals (e.g. *Glx*, *Lac*, *Ala*) and the intensity of short T_2-signals from macromolecules. The difference between the acquisition sequences PRESS and STEAM can also be observed in the coupled spins signals. Nevertheless, Tate *et al.* [109] concludes that pattern recognition is less sensitive to differences in acquisition parameter than had been expected.

Four main classes are considered, corresponding to four common brain tumour types, i.e. glioblastomas, meningiomas, metastases and astrocytomas of grade II. They are labeled as class 1 (glioblastomas), class 2 (meningiomas), class 3 (metastases) and class 4 (astrocytomas of grade II), respectively (see Table 7.2). The INTERPRET database also contains data from other brain tumours and other abnormal and normal brain masses. However, in order to restrict the unbalancedness for classification, in this chapter we focus the anal-

ysis on the brain tumour types with more than 20 data available. Moreover, the selected brain tumour types are the major groups in brain tumours and include both benign/low grade (meningiomas and astrocytomas grade II) and malignant (glioblastomas and metastases) tumour types, even within the glial type (glioblastomas vs. astrocytomas grade II). All data have passed a quality control and validation process, which was regulated by strict rules agreed on by all INTERPRET partners. This includes spectroscopic as well as clinical aspects (for details see [45]); the spectral quality control was carried out by expert spectroscopists, while for the clinical validation a Clinical Data Validation Committee (CDVC) was formed. Concerning the spectral quality control, spectra with either large water linewidths (> 8 Hz), a low signal-to-noise ratio (SNR < 10) or obvious artefacts are discarded. The spectral water linewidth is obtained by fitting a Lorentzian lineshape to the real part of the water unsuppressed spectrum after eddy current correction of the water unsuppressed FID (Free Induction Decay) using the phase information of the FID itself. The SNR is defined as the ratio of the highest signal intensity in the range 0-3.4 ppm and the standard deviation of the noise in the range 9-11 ppm of the water suppressed spectrum.

Spectra approved by one expert spectroscopist, are checked by another expert spectroscopist. If both expert spectroscopists agree, the spectrum is approved. In case of disagreement, a third expert spectroscopist will give the final decision. The clinical validation was done by the clinical committee CDVC. Each case was examined and based upon the quality of the available clinical information, biopsy and MRI scans, the case was validated or not. Cases with inadequate data (for example the spectral localization was not within the site of the brain tumour) or cases where there was no histological agreement about the diagnosis were discarded. The histopathological classification of the brain tumours was performed by two pathologists (three in case of disagreement between the first two, similar as for the spectral quality control).

The raw data (FID) are acquired in the time domain at the aforementioned centres. A few preprocessing steps are carried out: frequency alignment and phase correction with Klose's method [49] and filtering of the dominating residual water peak using HSVD [2], with 15 singular values and (4.09-5.31 ppm) as residual water region. The initial point of the time-domain signal was removed, because it was often affected by artefacts. The FIDs acquired with a GE-scanner had a different spectral width and number of points than the Siemens- and Philips-data (see Table 7.1). In order to match these parameters, the GE-data were interpolated using piecewise cubic splines in the time domain. Hence, after preprocessing all FIDs had a spectral width of 1000 Hz and 511 points (512 minus the first removed point).

The resulting signal is transformed to the frequency domain by a FFT (Fast Fourier Transform). Each MRS spectrum is considered only in the frequency region of interest (4.17 to 0 ppm), corresponding to 138 input variables. Figures

7.1 and 7.2 depict the mean magnitude L2-normalized spectra without baseline correction of the four considered classes. We note that no additional referencing is made, apart from the frequency alignment with Klose's method. This seemed to produce slight misalignment (up to 0.08 ppm).

Center	manufacturer	sequence	TE(ms)	TR(ms)	SW(Hz)	N
CDP	GE	STEAM	20	1600,2000	2500	2048
SGHMS	GE	STEAM	30	2000,2018,2020	2500	2048
		PRESS	30	2000,2018,2020	2500	2048
FLENI	GE	PRESS	30	2000	2500	2048
IDI	Philips	PRESS	30,31,32	2000	1000	512
UMCN	Siemens	STEAM	20	2000	1000	512
MUL	Siemens	STEAM	20	2000	1000	512

Table 7.1: *Overview of the acquisition parameters of the short echo time ^1H MRS data from all centres. The columns respectively correspond to the name of the centre, the manufacturer of the scanner, the sequence used, the echo time, the repetition time, the spectral width and the number of points in the original FID.*

Center	acq. scheme	glio	meni	meta	astroII	total
CDP	STEAM	30	12	5	4	51
SGHMS	STEAM	12	4	5	6	27
	PRESS	6	8	11	4	29
FLENI	PRESS	2	0	0	1	3
IDI	PRESS	31	30	17	5	83
UMCN	STEAM	2	1	0	1	4
MUL	STEAM	4	2	1	1	8
	total	87	57	39	22	205

Table 7.2: *Number of short echo time ^1H MRS data of glioblastomas (class 1), meningiomas (2), metastases (3) and astrocytomas of grade II (4). The rows correspond to the acquisition centre, while the columns mention the type of brain tumour.*

7.3 Methods

7.3.1 Classification techniques

Similar classification techniques (PCA, LDA and LS-SVM) as in Chapter 6 are applied here. The (hyper-) parameters of the classifiers are trained using the given data.

7.3.2 Input features

Normalization

According to Cousins [18], there are two approaches to interpret MRS signals: quantitative and qualitative. Pattern recognition can either be based on metabolite concentrations (quantitative approach) or on spectra (qualitative). The first approach requires absolute concentrations using an external or internal reference signal to which the metabolite concentrations can be calibrated [25, 38, 50]. Absolute quantitation involves corrections for T_1- and T_2-relaxation behaviour. As T_1- and T_2-values differ in between different brain regions [72] and water T_2-values are elevated in several brain tumour tissues compared to normal white matter [44], assessment of good estimates of T_1- and T_2-values for all relevant types of pathological tissues is very difficult. This motivates the use of qualitative approach in which the spectral values are fed to a classification algorithm as carried out in this study.

Nevertheless, also in the qualitative approach a normalization procedure should be applied that tries to compensate for intensity differences between spectra due to effects independent of the type of brain tissue. In this chapter, two kinds of normalization are compared. A first approach is L2-normalization which divides each spectrum by the L2-norm of its component values. Another approach is water normalization which uses the water amplitude as reference of the metabolite peaks. Both approaches were elaborated in the following way:

- L2-normalization: if you consider a spectrum in the frequency domain (full spectral range; here 1000 Hz) as a vector, then the L2-norm of that vector is defined as the square root of the sum of the squares of the vector modulus. In this study we divided each spectral value by the L2-norm of the spectrum. Hence, a L2-normalized spectrum has unity norm.
- water normalization: for each water suppressed signal, a corresponding water unsuppressed signal (internal standard) was available, acquired with the same acquisition parameters and originating from the same voxel. The amplitude of the water unsuppressed signal was estimated with software developed by Leentje Vanhamme [120]. The water unsuppressed signal is modeled in the time domain by a Voigt model [67, 95], with an additional first order term that corrects for eddy currents. HSVD is used to obtain initial estimates of the model parameters, which are subsequently used as starting values in a nonlinear least squares algorithm to obtain the final parameter estimates. Each spectral value in the water suppressed spectrum is then divided by the resulting estimate of the intensity of the water peak.

Within the INTERPRET project, quite some effort has been made to use two standard protocols, using a PRESS- or a STEAM-sequence, to obtain MRS

data measured under acquisition parameters that are as uniform as possible (see Table 7.1). Moreover, water normalization also compensates for differences in voxel size, assuming that different brain regions and tissues have the same water content. Note that this assumption is not completely valid as was concluded by experiments discussed in [44]. Also the use of different types of machines and pulse sequences (PRESS and STEAM) might still have a significant influence on the signal intensities and by consequence might also affect the classification performance. Carefully performed absolute quantitation might solve this problem.

Magnitude versus real spectra

Fourier transformed MRS signals consist of a real and an imaginary part, containing a mixture of absorption and dispersion spectra [11]. The combination of both spectra is typically used to reveal phase information. Another way of representing each signal is by calculating the magnitude spectrum, which makes the signal phase independent. In this chapter, we compare the use of real and magnitude spectra for classification. This extends our previous study on long echo time MRS signals, where we only considered magnitude MRS spectra [63].

Baseline correction

Short echo time ^{1}H MRS signals are characterized by the presence of an unknown broad baseline underlying the resonances of the metabolites of interest, that hinders the assessment of the intensity (i.e. quantitation) of low weight molecules. Several parametric and non-parametric approaches to model the macromolecular baseline are described in literature [3, 42, 43, 85, 92, 98], but most of these techniques are applied in combination with quantitation of MRS spectra. Independently of quantitation, the baseline can be modelled e.g. by polynomes or splines, but this requires the selection of several spectral points through which the baseline should be fitted, which is heavily user-dependent.

In this chapter, another approach is used, which requires the selection of only one parameter. The preprocessed FID is multiplied by an exponentially decreasing apodization function [10, 94], given by $g_n = e^{-\beta n \Delta t}, n = 0, \ldots, N-1$, with N the number of points in the FID and a parameter β. The resulting FID is a model for the baseline. The value $\beta = 0.15$ was chosen to avoid on the one hand a too flat baseline and overestimation of several metabolite resonances (β too high) and on the other hand the underestimation of several metabolite resonances (β too low). This baseline model was subtracted from the original signal to obtain the baseline corrected time-domain signal. The performance of the classification using the baseline corrected spectra is compared to the classification using the non-baseline corrected spectra.

7.3.3 Dimensionality reduction by prior knowledge

Complete spectra

As input features we use the complete spectra, obtained after preprocessing (see Section 7.2). Each spectrum consists of 138 variables (spectral points).

Selected frequency regions

Similar to the approach used in Section 6.3.1, the dimension of the MRS spectra is reduced by taking only the input variables within certain regions of the spectrum. By taking a range of ± 0.075 ppm around the resonance frequencies [33] of several characteristic metabolites, the number of variables is reduced to 71. The following regions of the MRS spectrum are selected: *L2* (lipids at $\approx$ 0.9 ppm; 0.825 - 0.975 ppm); *L1* (lipids at $\approx$ 1.2 ppm; 1.125 - 1.275 ppm); *Lac* ($^3CH_3$-group; 1.235 - 1.385 ppm); *Ala* ($^1CH_3$-group; 1.395 - 1.545 ppm); *NAA* ($^2CH_3$-group; 1.935 - 2.085 ppm); *Glx* ($^4CH_2$-group; 2.270 - 2.510 ppm); *Cr* ($N(CH_3)$-group; 2.955 - 3.105 ppm); *Cho* ($N(CH_3)_3$-group; 3.115 - 3.265 ppm); *mI* (^{5}CH-group) + *Tau* ($^2CH_2$-group) (3.195 - 3.345 ppm); *mI/Gly* (^{1}CH-, ^{3}CH-, ^{4}CH- and ^{6}CH-groups; 3.450 - 3.680 ppm); *Tau* ($^1CH_2$-group; 3.345 - 3.495 ppm); *Glx* + *Ala* (^{2}CH-groups; 3.665 - 3.815 ppm); *Cr* ($^2CH_2$-group; 3.845 - 3.995 ppm). This set of metabolite peaks is an extension of the one used in the previous study on long echo time MRS [63], because several additional metabolite peaks (e.g. *Glx*, *mI*, *Tau*) are more visible in short echo time MR spectra.

Peak integration

Another approach to select the most discriminatory input is based on peak integration. The amplitude of a resonance is proportional to the integral of the corresponding peak in the spectrum. However, precise estimation of the peak integrals is difficult due to several factors, including nonzero baseline, peak overlap, noise and also the discrete nature of the spectrum. Peak integration is performed here by using the trapezoidal rule [41]. For each selected metabolite the area under the frequency peak in the magnitude spectrum is calculated. We use the same frequency ranges as for the selected frequency regions above, which reduces the number of variables to 13.

7.3.4 Experimental setting

Binary classification

Binary classification provides more insight into each class distribution and its overlapping regions. We can also obtain information about the balance of the data distribution between the classes, and the classifier performance can be

measured using ROC analysis.

Given four types of brain tumours, then six binary classifiers can be constructed to separate the following pairs : glioblastomas vs. meningiomas (class 1 vs. 2), glioblastomas vs. metastases (1 vs. 3), glioblastomas vs. astrocytomas grade II (1 vs. 4), meningiomas vs. metastases (2 vs. 3), meningiomas vs. astrocytomas grade II (2 vs. 4) and metastases vs. astrocytomas grade II (3 vs. 4).

We applied the aforementioned techniques in the following way:

1. LDA and LS-SVM with linear kernel are used as linear techniques, while LS-SVM with RBF kernel is used as nonlinear technique.

2. LS-SVM techniques are applied using the complete spectra, using the selected frequency regions as well as using peak integration.

3. Feature selection using PCA prior to LDA is applied except for the case of peak integration where a specific dimensionality reduction is already applied.

LS-SVM classifiers require the tuning of a set of hyperparameters to achieve a high level of performance. These adjustable hyperparameters include a regularization parameter γ, which determines the tradeoff between the model complexity and training error rate. In case of an RBF kernel, the kernel width σ must be selected. We choose the values of hyperparameters γ for LS-SVM with a linear kernel, and (σ, γ) for LS-SVM with a RBF kernel using leave-one-out (loo) cross validation, while limiting the search space to avoid overfitting.

The experiment consists of the following steps:

1. divide the dataset in a training set (2/3 of the data) and a test set (remainder),

2. train the classifiers using the training set,

3. evaluate the performance using the test set.

Stratified random sampling is used while dividing the dataset to preserve the proportion of the classes. The aforementioned procedure is repeated 100 times to avoid bias possibly introduced by selection of a specific training and test set. In this way we try to obtain a representative performance on the test set. ROC analysis as mentioned in Chapter 2.2.2 is used to evaluate the binary classifiers. The classification performance is then measured by the mean AUC and its pooled standard error calculated from 100 randomizations.

7.3.5 Multiclass classification

We use here a similar scheme as explained in 6.3.3 to classify multiple classes.

7.3.6 Statistical analysis

From 100 runs, the mean AUCs ($\overline{\text{AUC}}$) are listed in the tables, as well as the standard error SE on the AUC. For each binary classifier C, the mean and standard error of the AUC are calculated as explained in Section 2.2.3.

7.4 Results

7.4.1 Binary classification

In the following the classification performance of PCA/LDA and LS-SVM (using linear and RBF kernels) are reported. The performance of the classifiers using the complete spectra of L2-normalized magnitude MRS spectra without baseline correction is shown in Table 7.3. This table differs from each of the following tables in one aspect; Table 7.4 shows the classification performance using water normalized magnitude MRS spectra without baseline correction. In Table 7.5, the classification results are shown using the L2-normalized magnitude spectra with baseline correction. Finally, Table 7.6 shows the classifiers performance using real spectra instead of magnitude spectra.

Selecting frequency regions of the spectra using knowledge of characteristic metabolites as mentioned in Section 7.3.3, reduces the number of datapoints from 138 to 71. Applying PCA (prior to LDA) further reduces this input dimension, retaining at least 75% of the variance. The results of the ROC analysis for classification using selected frequency regions and using peak integration obtained from the L2-normalized magnitude spectra without baseline correction are shown in Table 7.7 and Table 7.8, respectively.

As mentioned before, PCA is applied prior to LDA in the case of complete spectra and selected frequency regions. However, PCA requires a criterion in order to decide how many PCs to retain. Rencher [87] mentions four criteria, which do not necessarily take the same decision. Because our study is based on 100 runs for each classification problem, such a criteria should be easily automatable and computationally efficient. Therefore, only two of the mentioned criteria are feasible to use and with which we experimented in our study. A first criterion retains those components that account for more variance than the average for the individual variables. A second criterion takes the largest components that explain a certain percentage of total variance.

Although the first criterion is the default in many software packages, our tests revealed that, based on the available data, it retains many components (up to

Classes	PCA/LDA	LS-SVM lin	LS-SVM RBF
1 vs. 2	$0.9556 \pm 0.0278_{(4)}$	0.9725 ± 0.0210	0.9730 ± 0.0204
1 vs. 3	$0.5907 \pm 0.0967_{(4)}$	0.5917 ± 0.0939	0.5873 ± 0.0942
1 vs. 4	$0.9660 \pm 0.0292_{(3)}$	0.9635 ± 0.0305	0.9615 ± 0.0314
2 vs. 3	$0.9540 \pm 0.0441_{(4)}$	0.9770 ± 0.0263	0.9781 ± 0.0254
2 vs. 4	$0.9965 \pm 0.0093_{(11)}$	0.9959 ± 0.0100	0.9902 ± 0.0168
3 vs. 4	$0.9856 \pm 0.0248_{(2)}$	0.9916 ± 0.0178	0.9903 ± 0.0194

Table 7.3: *Binary classification using complete spectra. Average performance on the test set from 100 runs of stratified random samplings of the L2-normalized magnitude MRS spectra without baseline correction. As performance measure we use the mean AUC and its pooled standard error SE. The number between the brackets mentions the number of principal components used (column PCA/LDA).*

Classes	PCA/LDA	LS-SVM lin	LS-SVM RBF
1 vs. 2	$0.8399 \pm 0.0607_{(1)}$	0.9294 ± 0.0397	0.9581 ± 0.0281
1 vs. 3	$0.5492 \pm 0.0935_{(1)}$	0.5917 ± 0.0928	0.5950 ± 0.0935
1 vs. 4	$0.9485 \pm 0.0403_{(1)}$	0.9560 ± 0.0339	0.9598 ± 0.0326
2 vs. 3	$0.8924 \pm 0.0618_{(1)}$	0.9616 ± 0.0470	0.9704 ± 0.0309
2 vs. 4	$0.7427 \pm 0.1199_{(1)}$	0.9930 ± 0.0139	0.9793 ± 0.0303
3 vs. 4	$0.9784 \pm 0.0297_{(1)}$	0.9757 ± 0.0386	0.9837 ± 0.0284

Table 7.4: *Binary classification using complete spectra. Average performance on the test set from 100 runs of stratified random samplings of the water normalized magnitude MRS spectra without baseline correction.* For further explanation, see Table 7.3.

Classes	PCA/LDA	LS-SVM lin	LS-SVM RBF
1 vs. 2	$0.9591 \pm 0.0265_{(5)}$	0.9625 ± 0.0271	0.9616 ± 0.0269
1 vs. 3	$0.5762 \pm 0.0939_{(5)}$	0.5914 ± 0.0931	0.6002 ± 0.0930
1 vs. 4	$0.9650 \pm 0.0295_{(4)}$	0.9584 ± 0.0335	0.9587 ± 0.0334
2 vs. 3	$0.9607 \pm 0.0388_{(5)}$	0.9685 ± 0.0308	0.9736 ± 0.0281
2 vs. 4	$0.9965 \pm 0.0099_{(13)}$	0.9973 ± 0.0082	0.9960 ± 0.0100
3 vs. 4	$0.9725 \pm 0.0411_{(3)}$	0.9813 ± 0.0309	0.9836 ± 0.0282

Table 7.5: *Binary classification using complete spectra. Average performance on the test set from 100 runs of stratified random samplings of the L2-normalized magnitude MRS spectra with baseline correction. For further explanation, see Table 7.3.*

Classes	PCA/LDA	LS-SVM lin	LS-SVM RBF
1 vs. 2	$0.9543 \pm 0.0279_{(6)}$	0.9743 ± 0.0196	0.9742 ± 0.0199
1 vs. 3	$0.5628 \pm 0.0984_{(4)}$	0.5655 ± 0.0970	0.5753 ± 0.0973
1 vs. 4	$0.9755 \pm 0.0240_{(5)}$	0.9717 ± 0.0264	0.9691 ± 0.0278
2 vs. 3	$0.9562 \pm 0.0425_{(6)}$	0.9702 ± 0.0319	0.9734 ± 0.0288
2 vs. 4	$0.9959 \pm 0.0102_{(13)}$	0.9947 ± 0.0119	0.9917 ± 0.0154
3 vs. 4	$0.9924 \pm 0.0167_{(4)}$	0.9893 ± 0.0210	0.9875 ± 0.0233

Table 7.6: *Binary classification using complete spectra. Average performance on the test set from 100 runs of stratified random samplings of the L2-normalized real MRS spectra without baseline correction. For further explanation, see Table 7.3.*

Classes	PCA/LDA	LS-SVM lin	LS-SVM RBF
1 vs. 2	$0.9060 \pm 0.0430_{(2)}$	0.9748 ± 0.0202	0.9733 ± 0.0200
1 vs. 3	$0.5959 \pm 0.0932_{(2)}$	0.5898 ± 0.0945	0.6085 ± 0.0930
1 vs. 4	$0.9500 \pm 0.0358_{(1)}$	0.9636 ± 0.0318	0.9620 ± 0.0315
2 vs. 3	$0.9495 \pm 0.0468_{(2)}$	0.9712 ± 0.0316	0.9683 ± 0.0336
2 vs. 4	$0.9958 \pm 0.0103_{(7)}$	0.9956 ± 0.0105	0.9971 ± 0.0086
3 vs. 4	$0.9947 \pm 0.0133_{(1)}$	0.9924 ± 0.0172	0.9929 ± 0.0165

Table 7.7: *Binary classification using selected frequency regions. Average performance on the test set from 100 runs of stratified random samplings of the L2-normalized magnitude MRS spectra without baseline correction. For further explanation, see Table 7.3.*

Classes	LDA	LS-SVM lin	LS-SVM RBF
1 vs. 2	0.9675 ± 0.0243	0.9704 ± 0.0221	0.9713 ± 0.0210
1 vs. 3	0.5054 ± 0.0940	0.5762 ± 0.0956	0.6073 ± 0.0935
1 vs. 4	0.9580 ± 0.0335	0.9616 ± 0.0328	0.9539 ± 0.0355
2 vs. 3	0.9641 ± 0.0364	0.9640 ± 0.0378	0.9632 ± 0.0379
2 vs. 4	0.9968 ± 0.0090	0.9965 ± 0.0092	0.9875 ± 0.0201
3 vs. 4	0.9892 ± 0.0219	0.9938 ± 0.0147	0.9938 ± 0.0146

Table 7.8: *Binary classification using peak integration values. Average performance on the test set from 100 runs of stratified random samplings of the L2-normalized magnitude MRS spectra without baseline correction. For further explanation, see Table 7.3.*

26). Certainly in one case this gave raise to rank deficiency. Hence, we decided to use the second criterion, which requires to fix a threshold for the total variance explained by retained components. On the one hand, a too low threshold could give us too few components and a large loss of information On the other hand, if we fix the threshold too high, we risk to include components that are specific to the data [87] and might not be generalizable to independent data. Moreover, this implies that a larger amount of components included does not always guarantee a better test performance. Based on these rules, we decided to take a moderate percentage of 75% of total variance. When analyzing the results, we should take into account that some valuable information, present in the smaller components, might be lost.

7.4.2 Multiclass approach

As mentioned above, two additional binary classifiers are constructed by merging glioblastomas and metastases into one class of aggressive tumours. Table 7.9 shows the performance of these classifiers using the complete spectra as input.

Training performance

Table 7.10 shows a comparison of the multiclass classifier's training performance, based on all data coming from *meningiomas*, *astrocytomas grade II* and *aggressive tumours*. In the second step, we use classifier 1 vs. 3 to further subclassify the class of *aggressive tumours*. Using this subclassification, the multiclass classifier's performance is shown in Table 7.11.

Test performance

The training performance gives an indication about the classifier's ability to learn from the given data. However, to investigate its generalization capacity, the classifier should be applied to an independent test set. Therefore, 2/3 of the dataset is selected as training set and the remainder is used as test set. This stratified random sampling is repeated for 100 runs. The results are shown in Table 7.12 for one-step classification, which assigns the spectra to one of the 3 following classes: 2, 4 or 5. Table 7.13 shows the classifier performance after two-steps classification, which assigns the spectra to one of the 4 following classes: 1, 2, 3 or 4. Each spectrum of class 5 in step one, is either assigned to class 1 or class 3 in step two. In Tables 7.12 and 7.13 we mention the mean correct classification rate, the mean misclassification rate and the mean percentage of undecided cases and their standard deviation. The correct classification rate is defined as the percentage of correctly classified spectra, while the misclassification rate is the percentage of misclassified cases.

Classes	PCA/LDA	LS-SVM lin	LS-SVM RBF
2 vs. 5	$0.9570 \pm 0.0245_{(4)}$	0.9803 ± 0.0148	0.9801 ± 0.0148
4 vs. 5	$0.9701 \pm 0.0232_{(3)}$	0.9728 ± 0.0225	0.9725 ± 0.0226

Table 7.9: *Binary classification using complete spectra. Average performance on the test set from 100 runs of stratified random samplings of the L2-normalized magnitude MRS spectra without baseline correction. As performance measure we use the mean AUC and its pooled standard error SE.*

	PCA/LDA (%)	LS-SVM lin (%)	LS-SVM RBF (%)
complete spectra	86.7647	89.7059	88.2353
selected freq. range	80.8824	92.6471	92.6471
peak integration	89.7059	85.2941	83.8235

Table 7.10: *One-step multiclass classification. Percentage of correctly classified spectra using all L2-normalized magnitude MRS spectra without baseline correction to asses the training performance.*

	PCA/LDA (%)	LS-SVM lin (%)	LS-SVM RBF (%)
complete spectra	70.5882	72.0588	73.5294
selected freq. range	66.1765	73.5294	73.5294
peak integration	70.5882	67.6471	66.1765

Table 7.11: *Two-step multiclass classification. Percentage of correctly classified spectra using all L2-normalized magnitude MRS spectra without baseline correction to asses the training performance.*

	PCA/LDA (%)	LS-SVM lin (%)	LS-SVM RBF (%)
correct	85.0882 ± 4.2939	87.2794 ± 3.6466	86.6324 ± 3.4857
misclass	9.3676 ± 3.6722	8.7941 ± 3.5409	9.8971 ± 3.4336
undecided	5.5441 ± 2.4183	3.9265 ± 2.0592	3.4706 ± 1.9102

Table 7.12: *One-step multiclass classification. Average performance on the test set from 100 runs of stratified random samplings of the L2-normalized magnitude MRS spectra without baseline correction. The first, second and third rows give, respectively, the mean correct classification rate, the mean misclassification rate and the mean percentage of undecided cases, each with their standard deviation.*

7.5 Discussion

In this section we discuss various issues concerning the results we obtained using the available short echo time ^{1}H MRS data. We do not necessarily claim that these remarks generally hold for similar analyses on other data.

7.5.1 Classification techniques

Linear Discriminant Analysis, as a linear classifier, preceeded by PCA (except for peak integration), performs quite well in discriminating brain tumours given complete MRS spectra. Except for the most difficult binary classification to discriminate glioblastomas and metastases, classification of L2-normalized non-baseline corrected magnitude spectra reaches a mean AUC of at least 0.95 (Table 7.3). PCA helps in selecting the necessary principal components (PCs) to cover 75% of the data variance. PCA reduces the input up to 1 to 3 PCs for water normalized and 2 to 17 PCs for L2-normalized MRS complete spectra. This means that some classes can easily be separated by a straight line.

Kernel-based methods, such as LS-SVMs are less sensitive to the amount of datapoints. Although the dimension is larger than the number of datapoints, these classifiers can draw an optimal separating boundary, even without applying any dimensionality reduction (such as PCA). Linear as well as nonlinear classifiers can be selected by choosing an appropriate kernel, e.g. LS-SVM with linear kernel as linear classifier, while LS-SVM with RBF kernel as nonlinear classifier.

Although visual inspection points out some differences in the classification performance, for the considered cases no statistically significant difference was found between the AUC values from different classification techniques. The statistical analysis as described in Section 2.2.3 shows that linear techniques such as LDA perform well in linearly separable cases. Due to the limited data available, nonlinear techniques, in this case LS-SVMs with RBF kernel, can not fully exploit the advantage of a nonlinear method.

7.5.2 Influence of input features

Normalization tries to make a uniform scale of the measurement data. Water normalization is applied to have a volume-independent measure of the metabolite concentrations dividing each spectrum by the amplitude of the water peak in the unsuppressed spectrum as reference. In the L2-normalized short echo time MRS spectra, each spectrum is divided by its L2-norm. Using the magnitude MRS complete spectra without baseline correction, a significantly lower performance was found for water normalization for the classification of classes

1 vs. 2 (using PCA/LDA and LS-SVM with a linear kernel) and of classes 2 vs. 4 (using PCA/LDA).

Although the spectra are already (partially) phase corrected by Klose's method, there is still quite some variation in the phase of the real spectra, which makes it more difficult to recognize the pattern of the different classes. By using magnitude spectra, phasing problems are avoided. However, in contrast to real spectra, more peak overlap occurs in magnitude spectra. No significant differences were found between the use of real and magnitude MRS complete spectra without baseline correction.

In contrast to long echo time MRS spectra, short echo time MRS spectra are heavily influenced by overlapping peaks and the presence of a macromolecular baseline. Therefore, baseline correction is applied to reduce the influence of the macromolecules. We expected that spectra with baseline correction will have clearer pattern peaks and result in a more robust classification. However, the results show no significant difference between the two approaches. Although baseline correction simplifies the identification of metabolite peaks, also macromolecules included in the baseline contribute to biochemical fingerprint of the tumours. Hence, by applying baseline correction also some valuable information might be lost, which could explain why classification does not improve after baseline correction.

7.5.3 Influence of dimensionality reduction

In classifying short echo time MRS spectra using selected frequency regions and peak integrated values (Section 7.4.1) the input dimension is reduced by selecting only spectral regions which contain resonances of important metabolites. The underlying idea for this dimensionality reduction is to remove any redundant input features and reduce the influence of noise and artefacts. Hence, we try to enhance the discriminatory chemical information present in the spectra.

As shown in Table 7.7 and 7.8, all of the classifiers perform well. Except for the binary classification problem to discriminate glioblastomas and metastases, at least a mean AUC of 0.90 can be achieved. In comparison to classification using complete spectra (see Table 7.3), classifying using selected frequency regions or peak integration can reach about the same performance. This indicates that the selected metabolite regions or peaks cover most of the available information and can be used to correctly distinguish brain tumour types. Only for the discrimination of glioblastomas and meningiomas PCA/LDA performed significantly lower based on selected frequency regions (L2-normalized magnitude MRS spectra without baseline correction) than based on the complete spectra.

7.5.4 Multiclass classification

The multiclass classifier described in Section 6.3.3 shows a high learning and generalization capability. This is illustrated by the correct classification rates on the training set for the first step using the complete spectra as input: 86.8% (PCA/LDA), 89.7% (LS-SVM lin) and 88.2% (LS-SVM RBF). Given an independent test set as input, the classifiers on average give 85.1% (PCA/LDA), 87.3% (LS-SVM lin), and 86.6% (LS-SVM RBF), which does not differ much from the training performance.

The classified tumour type can be further refined by extending the multiclass classifier's output from three to four classes. In the second step of the multiclass classifier we forward the output of the first step to a binary classifier that separates glioblastomas and metastases. The test performances reduce to 67.6% (PCA/LDA), 69.4% (LS-SVM lin), and 68.5% (LS-SVM RBF) because of the difficulties in separating glioblastomas from metastases. As observed in the binary classification, these two classes share very similar pattern peaks. Therefore separating class 5 (aggressive tumours) into class 1 (glioblastomas) and 3 (metastases) deteriorates the total performance of the classifier.

7.6 Conclusions

This chapter presents a comparison of classification techniques for brain tumour classification using short echo time MRS spectra. Linear as well as nonlinear classifiers are applied automatically, including feature selection, (hyper-) parameter estimation, training and testing. These aim at ease of use and at the automated analysis of independent data in clinical practice.

Classifying tumour types in pairs (binary classification) gives us more insight on the distribution of each class and their overlap. All classifiers based on the L2-normalized magnitude MRS complete spectra without baseline correction are able to reach a mean AUC of more than 0.95, except for the most difficult binary classification to discriminate glioblastomas from metastases. This finding is in line with our previous study on long echo time MRS spectra reported in [63] and results described in [109].

Although in a few cases visual inspection reveals some differences in performance of the classification techniques, in general all classifiers performed very well. Linear techniques (PCA/LDA and LS-SVM with linear kernel) found the best separating linear boundary for most of the cases. Dimensionality reduction by selecting characteristic metabolite regions helps the PCA/LDA classifier to reach a mean AUC of at least 0.90 (except for the case of glioblastomas vs. metastases) with a lower number of PCs. Moreover, when using peak integration prior to LDA, the performances are of the same order as when using

complete spectra. Kernel-based methods, e.g. LS-SVMs, have the advantage of robustly learn the peak pattern without the use of any feature reduction. Even in case of a large input dimension (138 in the case of the complete spectra as input features), LS-SVMs are able to reach a high mean AUC (except for glioblastomas vs. metastases).

Dimensionality reduction, by selecting frequency regions or peak integration could possibly reduce the influence of disturbing noise and artefacts in the spectra. The described approach for selecting resonance peaks of short echo time ^{1}H MRS spectra makes it easier to reach a similar performance as using complete spectra, which also speeds up the computation by reducing the complexity. Although peak integration is a rough and inaccurate quantitation method, it yields a high classification performance. This indicates the potential use of absolute quantitation (e.g. by a parametric time-domain model) prior to classification.

Based on the available data we have not found any significant effect on the classification performance of the use of real or magnitude spectra, or the use of baseline correction like described in this chapter.

Normalizing MRS spectra by the water content results in a significant reduction of the number of selected PCs. For some cases, even 1 to 3 PCs are enough to proceed with linear classification. Meanwhile, L2-normalized MRS spectra require more PCs covering 75% to achieve a high classification accuracy. However, in a few cases water normalization yields a significantly lower performance than L2-normalization. This could be partially caused by the lower amount of PCs selected by the aforementioned criterion in the case of water normalization. However, as mentioned in Section 7.3.2 the use of different types of machines and pulse sequences might affect the classification performance. This effect could also be cancelled out by accurate absolute quantitation.

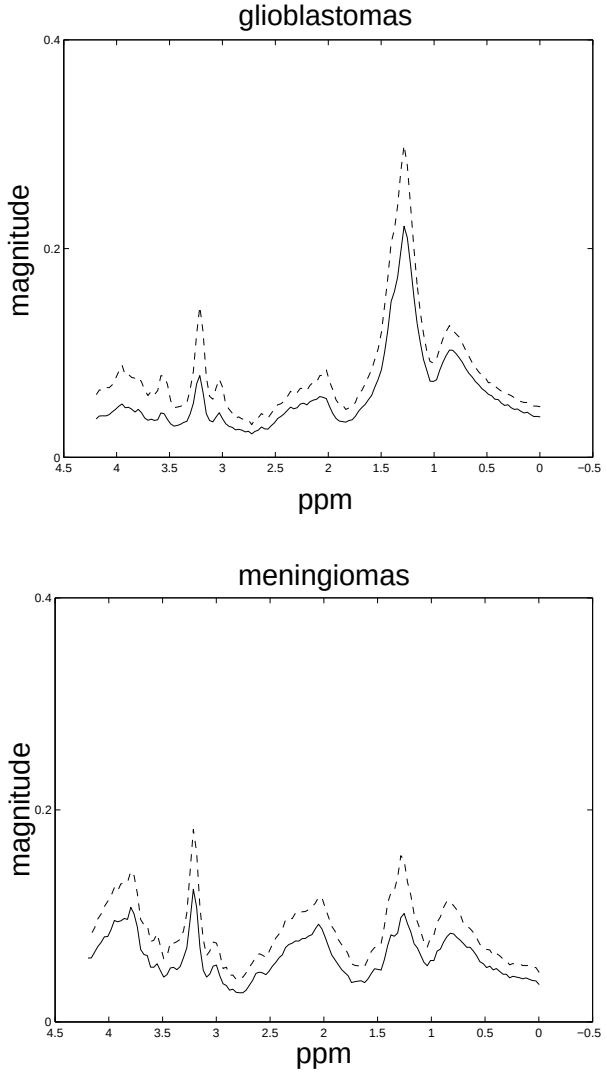


Figure 7.1: *Mean L2-normalized magnitude complete MRS spectra without baseline correction of the four considered classes: class 1 (top), class 2 (bottom) correspond to the glioblastomas, meningiomas respectively. The solid lines are the means, while the dotted lines are the means plus the standard deviations of each class.*

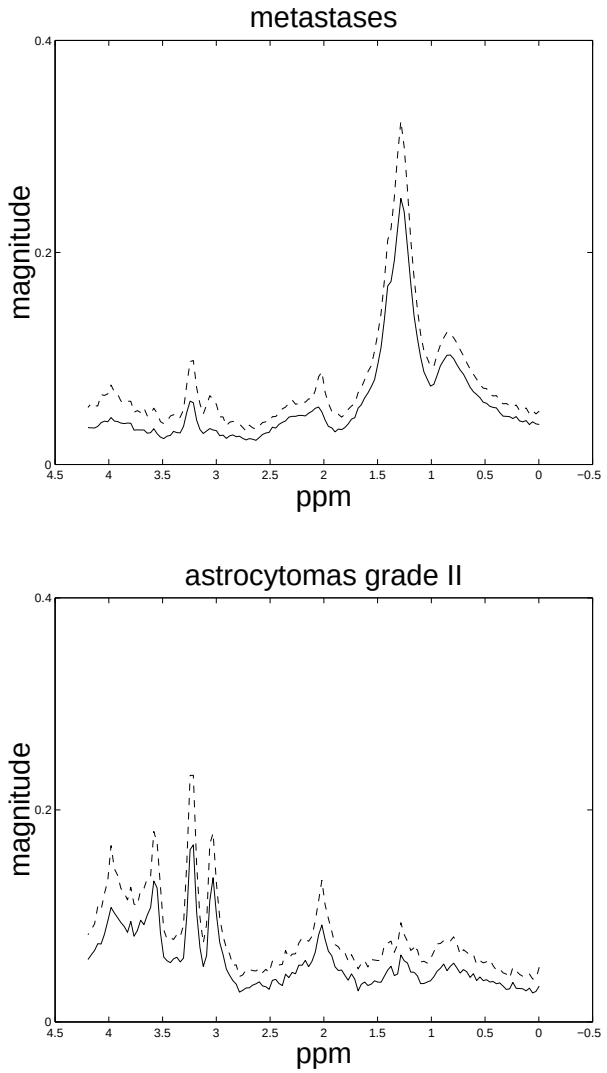


Figure 7.2: *Mean L2-normalized magnitude complete MRS spectra without baseline correction of the four considered classes: class 3 (top) and class 4 (bottom) correspond to the metastases and astrocytomas (grade II) respectively. The solid lines are the means, while the dotted lines are the means plus the standard deviations of each class.*

	PCA/LDA (%)	**LS-SVM lin (%)**	**LS-SVM RBF (%)**
correct	67.6176 ± 4.9418	69.4118 ± 3.6082	68.5441 ± 4.0417
misclass	26.8382 ± 4.5043	26.6618 ± 3.6122	27.9853 ± 4.0984
undecided	5.5441 ± 2.4183	3.9265 ± 2.0592	3.4706 ± 1.9102

Table 7.13: *Two-step multiclass classification. Average performance on the test set from 100 runs of stratified random samplings of the L2-normalized magnitude MRS spectra without baseline correction. The first, second and third rows give, respectively, the mean correct classification rate, the mean misclassification rate and the mean percentage of undecided cases, each with their standard deviation. Note that the number of undecided cases is equal to that for the one-step classifier.*

Chapter 8

Brain tumour classification using MRSI and MRI data

8.1 Introduction

Magnetic Resonance Imaging (MRI) is an important noninvasive tool for identifying the location and size of brain tumours. However, MRI is often nonspecific in determining the underlying type of brain tumour and grade [78]. In comparison, a second MR technique of potentially large diagnostic value is Magnetic Resonance Spectroscopy (MRS). As mention in Chapter 5, two types of MRS signals can be distinguished, depending on their type of localization technique. Single voxel ^{1}H MRS acquires a spectral profile (spectrum) from a single volume element or voxel, while multivoxel MRS or Magnetic Resonance Spectroscopy Imaging (MRSI) acquires spectra from a grid of selected voxels. During MRSI, for each voxel, a spectrum is measured that contains chemical information from metabolites present in the suspected brain tumours. In constrast to single voxel MRS, MRSI also provides spatial information to visualize the heterogeneity and spatial extent of the lesion.

In general, diagnosis of brain tumours is based on the microscopic examination of tissue obtained by a biopsy. Typically, a MRI study is carried out prior to the biopsy, in order to localize the tumour area. Accessing the tumour area by biopsy is invasive. Therefore, the procedure carries with it some risk of brain damage that could leave the patient with some loss of brain function. In comparison, this risk can be avoided by the use of noninvasive techniques like MRS and MRSI.

Lindon *et al.* [59] give an overview of pattern recognition methods and their applications in biomedical magnetic resonance. Several studies [39, 84, 94, 105,

107, 109, 126] also show some progress in automated pattern recognition for brain tumour classification based on MR data. These studies are either based on MRI (e.g. [126]), long echo time ^{1}H MRSI (e.g. [84]), short echo time ^{1}H MRS (e.g. [109]) or long echo time ^{1}H MRS (e.g. [39, 105, 107]). In order to enhance the diagnostic capabilities in clinical practice, MR images and MRSI spectra can also be used collectively. However, up to our best knowledge, currently only few studies (e.g. [94]) have used a combination of MRI and MRSI for classification of brain tumours. The purpose of this chapter is to investigate the relative contribution of MRSI and MRI to the discrimination of several brain tumour types and normal brain tissue types. Therefore, the image information as well as the MRSI spectra are used separately and in combination with each other. By comparing the obtained results, we get an idea about the discriminatory value of MRI and MRSI of brain tumours.

MRS (and MRSI) provides metabolic information, which can be quantified by computing an amplitude for each of the metabolites present in the examined tissue. Each amplitude is related to the concentration of the corresponding metabolite. There is a clear indication that MRS is able to uncover differences in metabolic concentrations and their ratios between normal and cancerous tissue ([96] and references therein): *NAA* (N-acetyl aspartate) and *Cr* (total creatine) levels are reduced with respect to normal tissue, *Cho* (choline containing compounds) is increased, and the ratios *Cr/Cho* and *NAA/Cho* are significantly reduced in all tumour types relative to control tissue [15, 66]. Differences in levels of *Asp* (aspartate), *Ala* (alanine), *Ins* (inositol) and *Glx* (glutamine/glutamate) were found between normal cells and various tumour types. Metabolites specific to meningiomas (e.g. *Ala*) and astrocytomas (e.g. *Lac*) were reported [66]. Differences in metabolite concentrations and ratios between different types of tumours (gliomas vs. meningiomas vs. glioblastomas vs. noncerebral tumours) have also been reported [51]. Lipids are found in both areas of necrosis and in tissues of high-grade astrocytomas and metastases [74]. In this chapter we consider data from normal tissue from volunteers and patients, cerebro spinal fluid (CSF) as healthy brain tissue types, as well as grade II, grade III gliomas, glioblastomas and meningiomas as brain tumour tissue types. Binary classification is carried out for the discrimination of e.g. benign and malignant tumours, grade II and grade III astrocytomas.

We consider linear as well as nonlinear classification techniques applied to several input features, such as magnitude short echo time MRSI spectra, MRI information, peak integration values obtained from the MRSI spectra and a combination of the latter two. As the input features consist of several variables and as classification is required to be objective and user-friendly, all techniques are automated. The algorithms are designed to extract the most important features which are then used to classify each spectrum into the corresponding tumour type.

8.2 Material

Data from 25 brain tumour patients and 4 volunteers selected from the database developed in the framework of the EU funded INTERPRET project (IST-1999-10310) [45] and were provided by the acquisition centre UMCN (Universitair Medisch Centrum Nijmegen), Nijmegen (The Netherlands). Each case was examined and based on the brain biopsy, the case was validated or not. For one of the 25 patients no consensus was reached. Therefore, the data from the tumour region of this patient were not used.

The dataset contains MR images as well as MRSI spectra, acquired and pre-processed as described in [94]. For each subject, MR images of cross-sections of the whole brain at four contrasts were acquired: T_1- and T_2-weighted images, a proton density image and a gadolinium-enhanced (Gd-DTPA) T_1-weighted image (256x256, FOV = 200 mm, slice thickness = 5 mm). The image values will further be labeled as T_1, T_2, PD and GD. Besides MR images, also ^{1}H MRSI data were acquired for each subject, both with and without water suppression using a 16x16 2D STEAM ^{1}H MRSI sequence with acquisition parameters TR = 2000 or 2500 ms, TE = 20 ms, slice thickness 12.5 or 15 mm, FOV 200 mm, spectral width = 1000 Hz, 1024 data points.

The set of validated data consisted of 10 grade II, 4 grade III tumours, 7 glioblastomas and 3 meningiomas, which gives us 4 classes of brain tumours. Furthermore for each patient, data were selected from several voxels, identified as lying in the tumour. Besides brain tumour data, also data from cerebrospinal fluid (CSF) from patients and normal brain tissue from volunteers and patients were selected. Data from all subjects with the same pathology were combined into one class, which gives rise to the following 7 classes of pathologies (see Table 8.1 for the total amount of data) :

1. normal tissue from volunteers: normal tissue from healthy persons
2. normal tissue from patients: normal tissue from the contralateral half of the brain
3. CSF: CSF from patients, where the ventricels were clearly visible and the voxels were located as far from the tumour as possible
4. grade II tumours: diffuse astrocytomas, oligodendrogliomas and mixtures
5. grade III tumours: astrocytomas, oligodendrogliomas and mixtures
6. glioblastomas
7. meningiomas.

To ensure that image pixels from subsequent images originate from the same spatial location, the images were co-aligned [94]. The location of the ^{1}H MRSI slice was aligned with the GD image which showed the largest enhanced tumour area. All MRSI data are semi-automatically preprocessed (cfr [94]), which involves:

- the k-space data were filtered by a Hanning filter of 50 % using the LUISE software package (Siemens, Erlangen, Germany),
- zero filling to 32 x 32,
- 2D Fourier transformation to obtain time domain signals for each voxel,
- correction for eddy current effects in the spectra using a method which prevents the occasional occurrence of eddy current correction induced artefacts [93],
- filtering of the dominating residual water using HLSVD, with 12 singular values and 4.0-6.0 ppm as residual water region,
- frequency alignment by setting the NAA peak to 2.02 ppm,
- first order phase correction,
- Fourier transformation to obtain frequency spectra.

8.3 Methods

8.3.1 Input features

This chapter addresses classification problems where the input is a vector that we call a pattern of n-components or features. The classifiers learn the collection of patterns of each particular class (tumour type) and try to recognize the class pattern which reveals the best contrast with respect to the other classes.

A pattern is either a MRSI spectrum, a set of quantified values (by peak integration) from the MRSI spectrum, a set of MRI values or a combination of the latter two input types. This enables us to investigate what the relative contribution is of the image information and the spectroscopic information for pattern recognition of brain tumours. The following input features are considered:

- water normalized magnitude spectra (see Figures 8.1-8.3): for each water suppressed signal, a corresponding water unsuppressed signal (internal standard) was available, acquired with the same acquisition parameters

and originating from the same voxel. The amplitude of the water unsuppressed signal was estimated with a software developed by Leentje Vanhamme [120]. The water unsuppressed signal is modeled in the time domain by a Voigt model [67, 95], with an additional first order term that corrects for eddy currents. HSVD is used to obtain initial estimates of the model parameters, which are subsequently used as starting values in a nonlinear least squares algorithm to obtain the final parameter estimates. Each spectral value in the preprocessed water suppressed spectrum is then divided by the resulting estimate of the intensity of the water peak. Only the spectral values in the region of interest (0.5-4.0 ppm) are used as input features.

- metabolite amplitudes obtained by peak integration: short echo time ^{1}H MRSI signals are characterized by the presence of an unknown broad baseline underlying the resonances of the metabolites of interest, that hinders the assessment of the intensity (e.g. by peak integration) of low weight molecules. In order to remove this broad baseline a simple baseline correction [10, 22, 94] was applied as additional preprocessing step prior to the first order phase correction. This was performed by exponential filtering of the time domain signal of each MRSI spectrum with a filter width of 5 ms, followed by the subtraction of the residual from the original time domain signal [94]. Amplitude estimates were then obtained from the baseline corrected frequency spectra using peak integration within a spectral range of 0.13 ppm around ten different resonance peaks. This involves resonances from metabolites and lipids that are assumed to be characteristic to distinguish between tumour types [44, 55, 64, 74, 75, 96]: lipids (0.90 ppm), *Lac* (1.33 ppm), *NAA* (2.02 ppm), *Glx* (2.20 and 3.75 ppm), *Cr* (3.02 and 3.95 ppm), *Cho* (3.20 ppm), *Glx* (3.44 ppm), *mI* (3.56 ppm). The resulting quantified values are then water normalized as described as above.

- image information: for each of the images (T_1, T_2, PD and GD), an image value was extracted, producing four additional variables. In order to obtain the same spatial resolution of both MRI and MRSI data, the resolution of the MRI data was lowered to that of the MRSI grid by averaging the image pixels within each spectroscopy voxel [94]. The values within each low resolution image were scaled to the same range as the spectral data.

8.3.2 Classification techniques

Binary classification is performed by Linear Discriminant Analysis (LDA), Least Squares Support Vector Machines with linear or Radial Basis Function (RBF) kernel, which are also applied in classification based on long echo (Chapter 6) and short echo time MRS (Chapter 7).

We apply here LS-SVM with linear as well as RBF kernel. The PCA/LDA technique has also been used as discussed in Chapter 6.

Label	Pathology	No. of data	No. of subjects
1	normal from volunteers	142	4
2	normal from patients	76	4
3	CSF	100	8
4	grade II	176	10
5	grade III	57	4
6	glioblastoma	70	7
7	meningioma	48	3
	Total	669	29

Table 8.1: *Number of MRI/MRSI data for each type of brain tissue (brain tumour or healthy tissue) provided by UMCN. The first and second columns give the label of the classes (from 1 till 7) and the pathology. The third and fourth columns display the number of data and subjects for each class. Note that the total number of subjects is not simply the summation of the number of subjects per class, because for several patients data are available from brain tumour as well as from healthy tissue.*

8.3.3 Experimental setting

As data from many types of brain tissue are available (see Table 8.1), many classification problems can be focussed on. Classifiers can be developed that distinguish healthy (classes 1 till 3) from tumour tissue (classes 4 till 7). Data from normal brain tissue from volunteers as well as from brain tumour patients are available, which can be used to test whether volunteers and patients can be distinguished based on their normal brain tissue. Furthermore, the dataset includes data from low as well as high grade (astrocytic) tumours, astrocytic tumours and meningiomas, which makes it possible to consider several other classification problems.

We applied the aforementioned techniques in the following way:

1. LDA and LS-SVM with linear kernel are used as linear techniques, while LS-SVM with RBF kernel is used as a nonlinear technique.
2. All techniques are applied using the image information, peak integration values, magnitude spectra and the combination of image information and peak integration values.
3. Feature selection using PCA prior to LDA is applied when using magnitude spectra.

8.4 Results

8.4.1 Healthy vs. tumour tissue

Although there exists a high degree of variability in human brain tumour types [74, 96], some general characteristic differences between normal and brain tumour tissues have been reported: *NAA* and *Cr* levels are reduced, *Cho* levels are increased. Also the ratios of creatine to choline-containing compounds (*Cr*/*Cho*) and *NAA* to choline-containing compounds (*NAA*/*Cho*) are significantly reduced in all tumour types relative to control tissue [15, 66]. Several brain tumour types have elevated T_2-components and increased water content compared to normal brain tissue [44, 66].

This section shows the classification results for the discrimination of healthy and tumour tissue. The data from healthy tissue correspond to classes 1,2 and 3 (see Table 8.1), while the class of tumour tissue is obtained by combining classes 4,5,6 and 7. This results in 318 healthy tissue spectra and 351 tumour spectra. Classification is performed either on the MRI data (4 variables), the peak integration values (10), the combination of MRI and peak integration values (14), or the complete magnitude spectra (231).

Table 8.2 shows the classification performance for distinguishing healthy brain tissue from brain tumour MRI/MRSI data. The resulting mean AUC is for all techniques and input patterns very high (higher than 0.95). Although the performance based on MRI only is the lowest, the combination of MRI and MRSI information improves classification with respect to the performance by using MRSI only.

Classifier	LDA	LS-SVM lin	LS-SVM RBF
MRI	0.9547 ± 0.0134	0.9549 ± 0.0132	0.9728 ± 0.0116
peak integration	0.9923 ± 0.0039	0.9927 ± 0.0037	0.9959 ± 0.0051
MRI/peak integration	0.9991 ± 0.0009	0.9992 ± 0.0008	0.9996 ± 0.0006
spectra	$0.9741 \pm 0.0087_{(5)}$	0.9882 ± 0.0063	0.9961 ± 0.0038

Table 8.2: *Classification of healthy vs. tumour tissue. Average performance on the test set from 100 runs of stratified random splittings.* MRI*: image information;* peak integration*: peak integration values;* MRI/peak integration*: MRI and peak integration values;* spectra*: complete magnitude spectra. Performance measures for each classifier are the mean AUC and its pooled standard error. The number between the brackets mentions the number of principal components used (column PCA/LDA).*

8.4.2 Normal tissue of brain tumour patients vs. normal tissue of volunteers

To the best of our knowledge there is no study that investigates whether a brain tumour also affects the condition of normal tissue. We consider the binary classification for the discrimination of normal tissue of volunteers (class 1; 142 data) from normal tissue of brain tumour patients (class 2; 76 data), which classification performances are shown in Table 8.3.

Classifier	LDA	LS-SVM lin	LS-SVM RBF
MRI	0.9996 ± 0.0011	0.9996 ± 0.0011	0.9983 ± 0.0028
peak integration	0.8697 ± 0.0471	0.8709 ± 0.0470	0.9320 ± 0.0320
MRI/peak integration	0.9995 ± 0.0014	0.9995 ± 0.0014	0.9986 ± 0.0024
spectra	$0.8263 \pm 0.0510_{(23)}$	0.8170 ± 0.0578	0.9254 ± 0.0330

Table 8.3: *Classification of normal tissue of patients vs. volunteers. Average performance on the test set from 100 runs of stratified random splittings. For further explanation, see Table 8.2.*

8.4.3 Benign vs. malignant tumours

Accurate determination of the grade of malignancy is one of the most important factors for some types of tumours. Generally, fast growing tumours are much more likely to emerge again after treatment than slowly growing tumours. On the other hand, high grade tumours do respond better to radiotherapy and chemotherapy. Lipids, *Lac* and macromolecules are significant markers of tumour malignancy [19, 44].

Astrocytomas of grade II (class 4) and the meningiomas in our dataset (7) are considered to be benign (224 data), while astrocytomas of grade III (5) and glioblastomas (6) are assigned to the class of malignant tumours (127 data). Table 8.4 shows that classification is poor based on MRI only. Also in case of PCA/LDA discrimination based on the spectral pattern is difficult. Based on peak integration values and the combination with MRI, a mean AUC of at least 0.91 is obtained.

8.4.4 Low grade astrocytomas vs. high grade astrocytomas

According to the World Health Organization (WHO) [81], well-differentiated and low-grade astrocytomas may be classified as grades I and II, anaplastic astrocytomas as grade III and glioblastomas as grade IV. Low-grade astrocytomas show e.g. mild anaplasia and no mitosis and necrosis, while astrocytomas

Classifier	LDA	LS-SVM lin	LS-SVM RBF
MRI	0.6239 ± 0.0543	0.6239 ± 0.0543	0.8469 ± 0.0385
peak integration	0.9210 ± 0.0243	0.9328 ± 0.0220	0.9827 ± 0.0124
MRI/peak integration	0.9195 ± 0.0255	0.9260 ± 0.0240	0.9918 ± 0.0072
spectra	$0.8480 \pm 0.0402_{(2)}$	0.9573 ± 0.0189	0.9797 ± 0.0161

Table 8.4: *Classification of low grade vs. high grade tumour. Average performance on the test set from 100 runs of stratified random splittings. For further explanation, see Table 8.2.*

of grade III are characterized by moderate anaplasia and mitotic figures [74]. Glioblastomas, represent the extreme malignant variant of astrocytic tumours and are marked by e.g. pronounced anaplasia, numerous mitoses and necrosis [74]. There is no doubt that glioblastomas arise through a process of anaplasia from astrocytomas, but it can not be said that this is an absolute rule [91]. Sometimes, no sign indicating the preexistence of an astrocytoma is found.

Table 8.5 shows the classification results for the discrimination of low-grade (class 4; 176 data) and high-grade astrocytomas (classes 5 and 6; 127 data). Similar classification results are obtained as in section 8.4.3, but here a slightly higher performance is reached. This might be partly due to the lower heterogeneity in the two compared classes, while in section 8.4.3 also non-astrocytomas (namely meningiomas) were included.

Classifier	LDA	LS-SVM lin	LS-SVM RBF
MRI	0.7429 ± 0.0517	0.7431 ± 0.0517	0.8774 ± 0.0354
peak integration	0.9452 ± 0.0223	0.9453 ± 0.0222	0.9774 ± 0.0158
MRI/peak integration	0.9563 ± 0.0203	0.9589 ± 0.0193	0.9920 ± 0.0084
spectra	$0.8621 \pm 0.0389_{(2)}$	0.9809 ± 0.0114	0.9896 ± 0.0081

Table 8.5: *Classification of low grade vs. high grade astrocytomas. Average performance on the test set from 100 runs of stratified random splittings. For further explanation, see Table 8.2.*

8.4.5 Astrocytic tumours vs. meningiomas

Astrocytomas are the most common primary intracerebral tumors in adults, while the most common primary extracerebral tumour is meningioma [74]. In astrocytomas, the level of *NAA* is reduced, which indicates degradation of the neuronal activity [30]. Another characteristic of astrocytomas is the marked elevation of choline with respect to normal brain tissue [74]. However, not all studies [114] were able to confirm this finding for all grades of astrocytomas. Also in some meningiomas, *Cho* may be elevated. Creatine levels in astrocy-

tomas remain unchanged or are decreased compared to normal brain tissue [74, 114]. Many researchers regard the presence of alanine to be characteristic for meningiomas [9]. However, alanine is not always detectable [66, 74].

Table 8.6 shows the results for the discrimination of all astrocytic tumours (classes 4,5 and 6; 303 data) from meningiomas (7; 48 data). Note that the same data are used as in section 8.4.3, only now astrocytomas are differentiated from meningiomas. At least a mean AUC of 0.85 is reached when based on MRI, peak integration or the combination of both. Only when using LDA based on the first PCs of the MRSI spectra the performance is poor.

Classifier	LDA	LS-SVM lin	LS-SVM RBF
MRI	0.8593 ± 0.0406	0.8590 ± 0.0407	0.9520 ± 0.0303
peak integration	0.9945 ± 0.0056	0.9947 ± 0.0055	0.9985 ± 0.0025
MRI/peak integration	0.9961 ± 0.0040	0.9964 ± 0.0039	0.9989 ± 0.0019
spectra	$0.6323 \pm 0.0750_{(2)}$	0.9889 ± 0.0092	0.9912 ± 0.0092

Table 8.6: *Classification of astrocytomas tumours vs. meningiomas. Average performance on the test set from 100 runs of stratified random splittings. For further explanation, see Table 8.2.*

8.4.6 Grade II vs. grade III astrocytic tumours

As already mentioned above, accurate determination of the grade is very important (see Section 8.4.3). In this section we focus on the specific discrimination of grade II and grade III astrocytomas. We consider 176 grade II astrocytomas (class 4) and 57 grade III astrocytomas (5). The results for the discrimination of the tumours are shown in Table 8.7.

Classifier	LDA	LS-SVM lin	LS-SVM RBF
MRI	0.7799 ± 0.0689	0.7799 ± 0.0688	0.9018 ± 0.0410
peak integration	0.9278 ± 0.0310	0.9290 ± 0.0306	0.9706 ± 0.0217
MRI/peak integration	0.9480 ± 0.0313	0.9488 ± 0.0308	0.9907 ± 0.0098
spectra	$0.6728 \pm 0.0670_{(1)}$	0.9669 ± 0.0196	0.9861 ± 0.0141

Table 8.7: *Classification of grade II vs. grade III tumour. Average performance on the test set from 100 runs of stratified random splittings. For further explanation, see Table 8.2.*

8.5 Discussion

8.5.1 Classification techniques

Tables (8.2)-(8.7) show that all classification techniques perform very well, especially based on the peak integration values and the combination of MRI and peak integration values. Similar results are obtained for classification based on LDA and LS-SVM (linear and RBF kernel). This is in line with our previous classification studies on long [63] and short echo time ^{1}H MRS [22]. However, in several cases a significant difference was found, e.g. for the discrimination of low- and high-grade astrocytomas, based on the image information LS-SVM with an RBF kernel significantly performs better than LDA and LS-SVM with a linear kernel (a z-value of respectively 2.11 and 2.15). For the same binary problem, based on complete magnitude spectra both kernel-based techniques reach a significantly higher performance than PCA/LDA. This latter also holds for the discrimination of astrocytomas tumour and meningiomas, and the grade II- and grade III tumours.

In a few cases, a quite low performance is reached when using LDA directly based on the magnitude spectra. This might be partly due to the use of PCA a prior to LDA.

8.5.2 MRSI versus MRI

Although MRI reveals spatial information, the medical diagnosis of brain tumours is normally not made based on MRI only. According to Just and Thelen [48], the considerable overlapping of tissue parameters in different tumor entities, indicate that a reliable diagnosis cannot be made on the basis of a quantitative evaluation of T_1, T_2, and proton density alone. However, it could be worthwhile to investigate the relative contribution of MRI and MRSI, and the discriminatory value that MRI could add to the classification of brain tissue data.

From our results we observe that, in some cases the use of MRI alone reaches a significantly lower performance, e.g. for the discrimination of low grade and high grade tumours (MRI vs. peak integration values, vs. the combination of MRI and peak integration values and vs. spectra, using any of the classification techniques). This is in correspondence to the conclusion of [48], as MRI alone does not contain enough information to directly discriminate tumours.

But on the other hand, in some cases the use of MRI alone or the combination with MRSI yields a significantly better performance. E.g. when discriminating normal tissue from patients from that of volunteers, a significantly better performance is found when using MRI information with respect to the use of peak integration values and MRSI spectra, and also when using the combina-

tion of MRI and peak integration values with respect to using MRSI spectra directly. However, this could partly be explained by normalization problems of the images. Also when addressing the classification problem of astrocytomas vs. meningiomas with LDA, the use of MRI yields a significantly higher performance then the direct use of spectra. However, this is mainly due to the very poor performance of PCA/LDA based on the magnitude spectra.

In almost all cases (Tables 8.2, 8.3, 8.5, 8.6 and 8.7) the combined use of MRI and MRSI information is increasing the performance. This is especially remarkable when applying LDA. Although no significant evidence is found, information from the spectra could be able to complement the MRI information and the combined use of MRI and MRSI could be also more accurate in discriminating tumour type as well as tumour grade.

8.5.3 Several classification problems

If a classification technique is developed for diagnostic purposes, then the technique should be able to distinguish healthy tissue from tumour tissue. As mentioned in section 8.1, several metabolic differences were found between normal cells and various tumour types. This should be reflected in MRSI spectra. As can be seen from Figures 8.1-8.4, the *NAA*- and *Cho*-levels are lower in tumour spectra, while the *Cr*-level is higher. Table 8.2 shows that the applied techniques are able to extract and exploit this differing pattern available in quantitation values, the combination of quantitation values and image information. Although the performance seems to be a bit lower (not significantly), also the differences in image information and magnitude spectra can be directly exploited.

A possible strategy for diagnostic purposes is to compare MR images and spectra of tumour tissue with those of normal brain tissue from volunteers. Based on the different characteristics, the tissue is identified as normal brain tissue or some type of brain tumour. However, this approach is only valid under the assumption that normal brain tissue from volunteers has the same MR features as normal brain tissue from patients. In this chapter we investigated whether differences in MR features of normal brain tissue from volunteers and patients make it possible to distinguish these two classes. Fig. 8.1 displays no differences between the mean spectral patterns of both types of normal tissue. Although the image variables from the normal tissue of patients reach a lower value than those from normal tissue of volunteers (see Fig. 8.5) and some discrimination could be made based on these image variables (see Table 8.3), this could be partly due to normalization problems during the preprocessing of the images.

As shown in Table 8.4, the applied classifiers are able to distinguish benign and malignant tumours using peak integration values, the combination of image information and peak integration values or magnitude spectra. Using the image

information only, the performance is poor, especially of the linear classifiers. From Figures 8.6-8.8 we remark that the image information for classes 4 and 7 indeed ressembles a lot with that of classes 5 and 6.

Similarly to the previous case, discrimination of all astrocytic tumours from meningiomas as shown in Table 8.6 give a significant improvement from using image information only. Combining the image information with peak integration does not give significant improvement. Peak integration alone has performed well, also in comparison to the use of magnitude spectra. Glial tumours have quite distinct peak patterns in metabolite rather than those captured by imaging.

Table 8.7 shows the classification performance of the classifier discriminating grade II and grade III astrocytomas. Using the image information only, a linear classifier performed poorly. This might be due to the unbalanced distribution of the data (176 of class 4, and 57 of class 5). Although combination of image information and peak integration yields a relatively higher classification performance for this case, we found no statistically significant difference in performance with the corresponding classifier using only peak integration.

8.6 Conclusions

This chapter discusses and investigates the combined use of MRI and MRSI for the discrimination of several brain tissue types. From the results section, it can be noticed that binary classification based on MRSI and MRI is promising. The use of MRI only does not always reach a good performance, but MRI might be able to add valuable information to MRSI for the discrimination of several brain tissue types.

This chapter specifically addresses binary classification, but using these results a strategy could be elaborated to carry out multiclass classification, e.g. by a tree structure. This work can be extended to the implementation of a multiclass classifier.

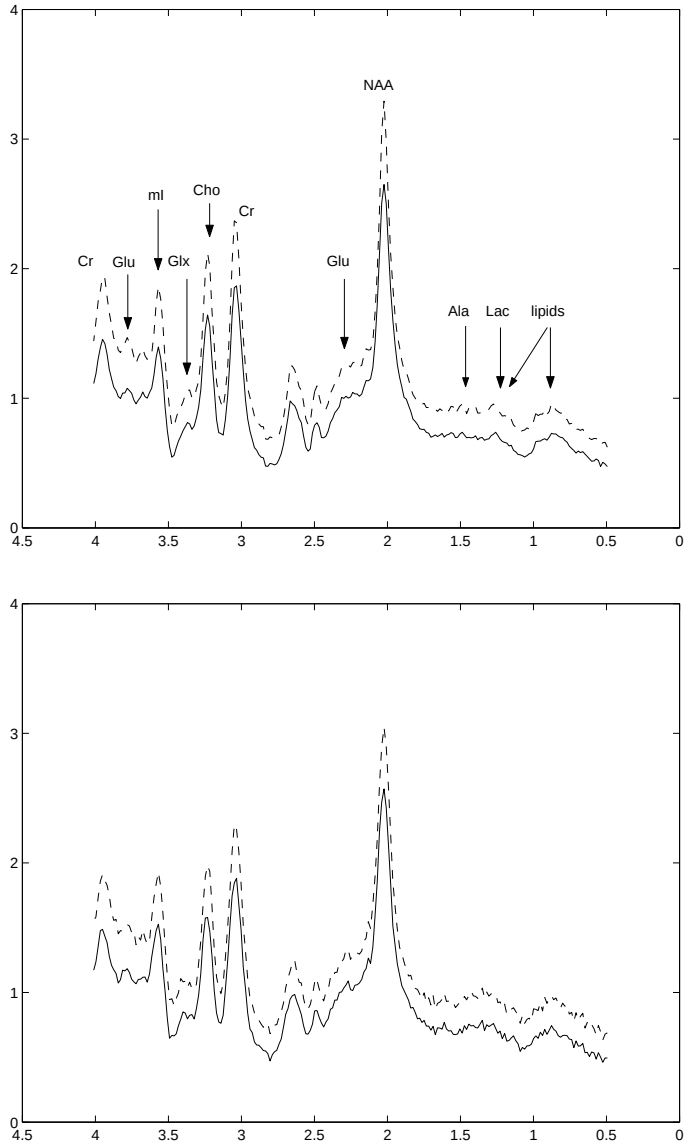


Figure 8.1: *Mean water normalized magnitude ^1H MRSI spectra (TE=20 ms) of the considered classes: class 1 (top), class 2 (bottom) correspond to the normal tissue of volunteers and normal tissue of patients. The solid lines are the means, while the dotted lines are the means plus the standard deviations of each class.*

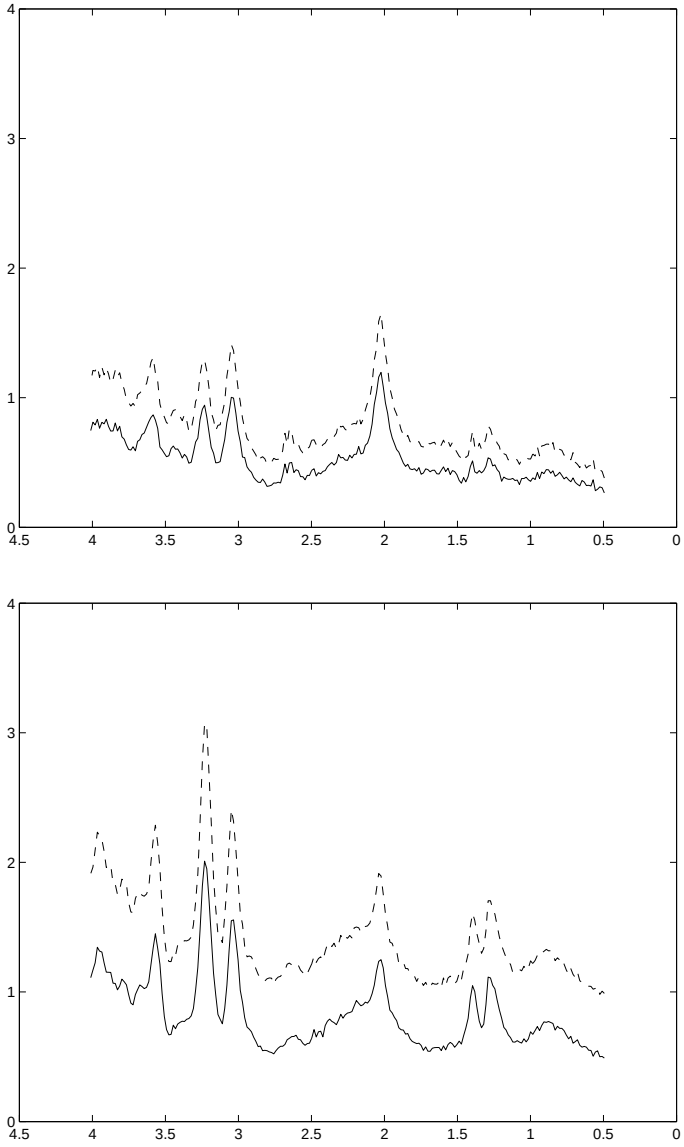


Figure 8.2: *Mean water normalized magnitude ^1H MRSI spectra (TE=20 ms) of the considered classes: class 3 (top), class class 4 (bottom) correspond to the CSF and tumours of grade II. For further explanation, see Figure 8.1.*

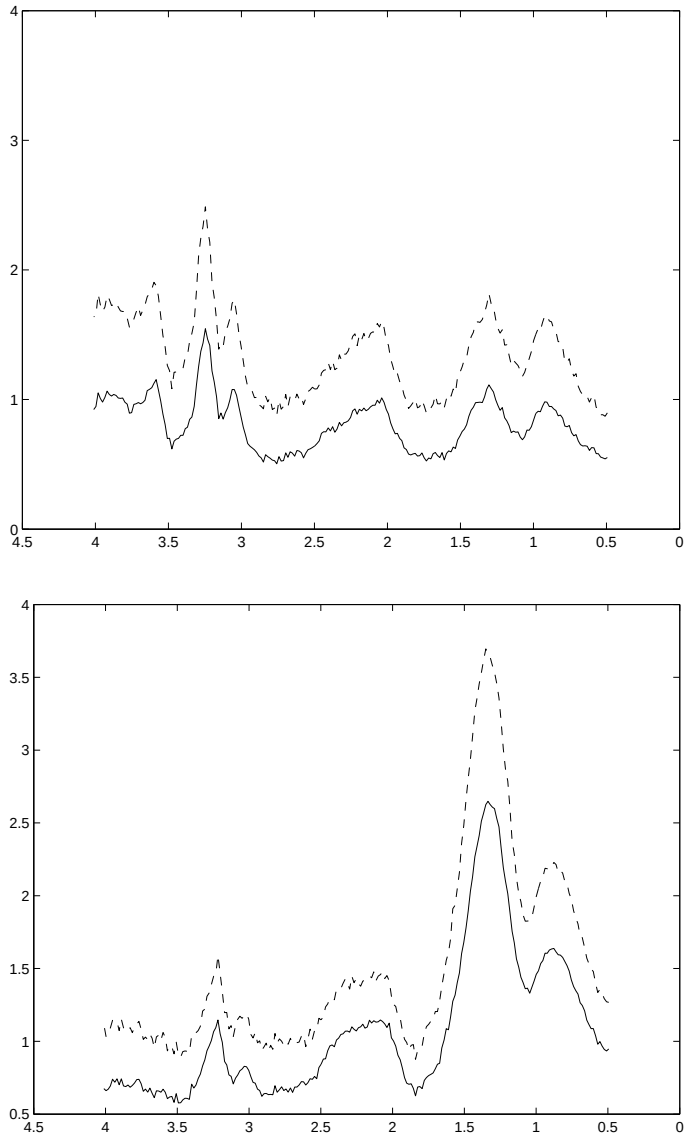


Figure 8.3: *Mean water normalized magnitude* ^1H *MRSI spectra (TE=20 ms) of the considered classes: class 5 (top), class 6 (bottom) correspond to the astrocytomas tumours of grade III and glioblastomas. For further explanation, see Figure 8.1.*

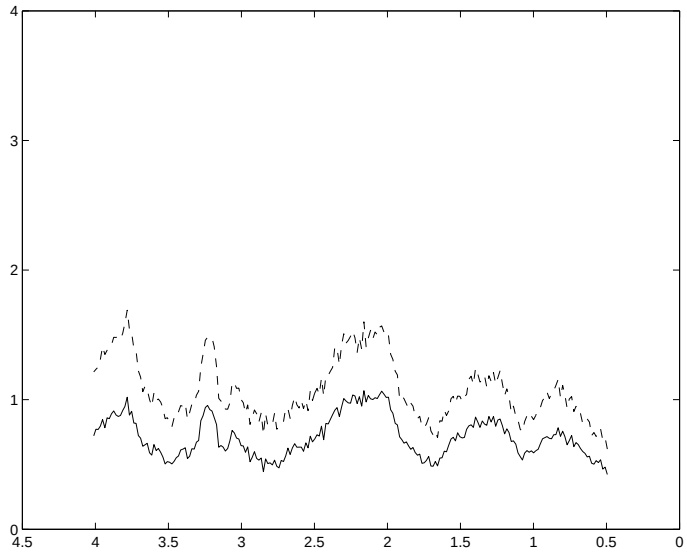


Figure 8.4: *Mean water normalized magnitude* ^{1}*H MRSI spectra (TE=20 ms) of the considered classes: class 7 which corresponds to meningiomas. For further explanation, see Figure 8.1.*

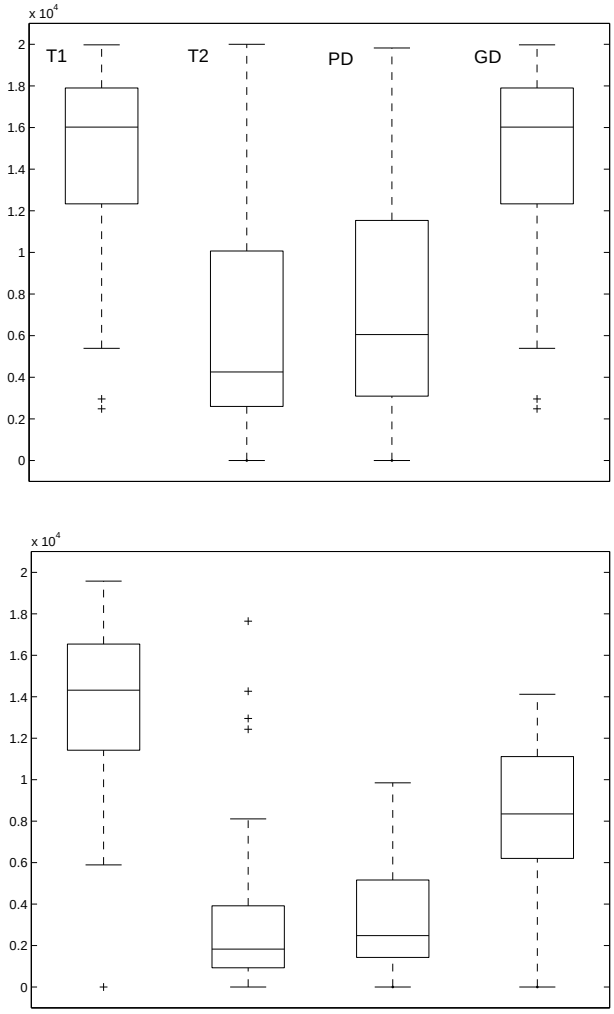


Figure 8.5: *Boxplots of the MRI values of the considered classes: class 1 (top) and class 2 (bottom) correspond to the normal tissue of volunteers, normal tissue of patients. Each box has lines at the lower quartile (lowest line), median (line in the middle), and upper quartile (highest line) values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers.*

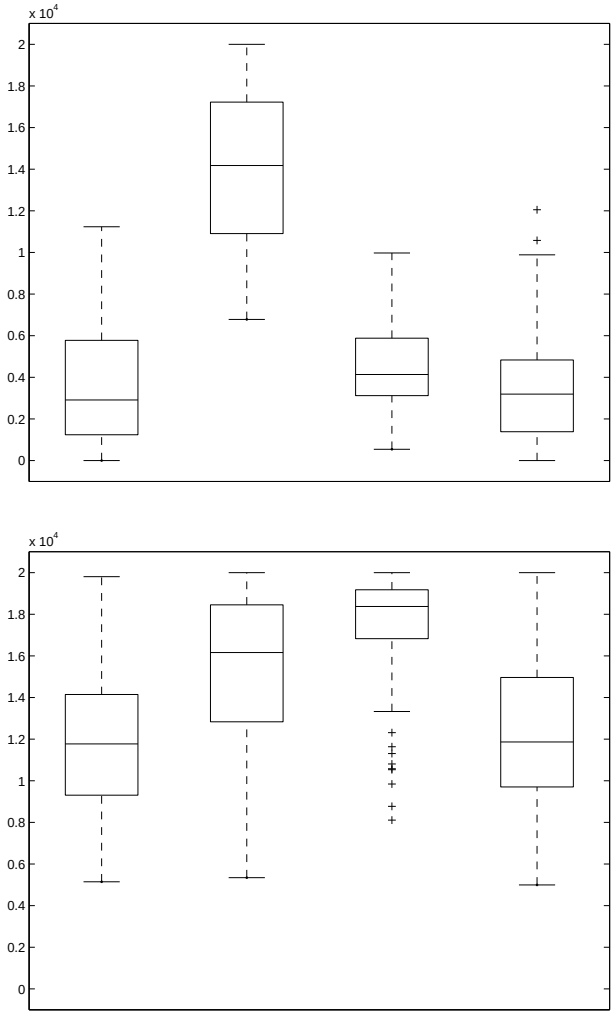


Figure 8.6: *Boxplots of the MRI values of the considered classes: class 3 (top) and class 4 (bottom) correspond to the CSF and astrocytomas tumours of grade II. For further explanation about boxplot, see Figure 8.5.*

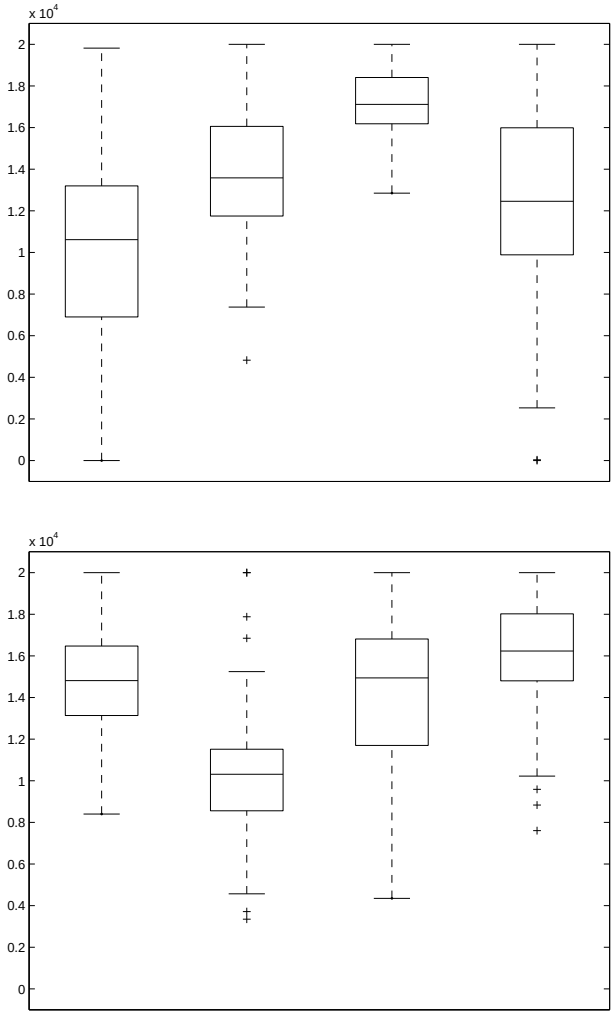


Figure 8.7: *Boxplots of the MRI values of the considered classes: class 5 (top) and class 6 (bottom) correspond to astrocytomas tumours of grade III and glioblastomas. For further explanation about boxplot, see Figure 8.5.*

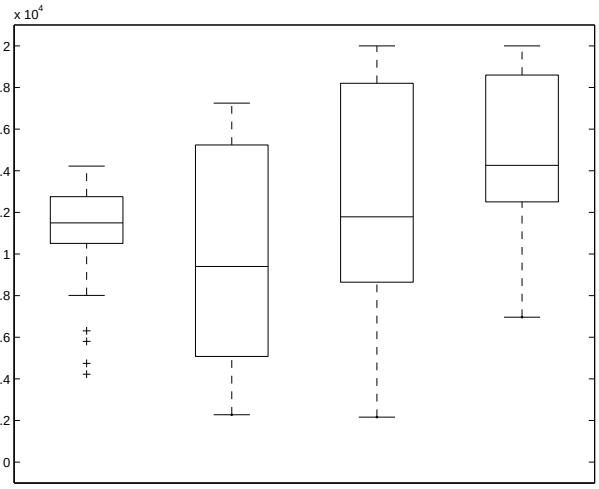


Figure 8.8: *Boxplots of the MRI values of the considered classes: class 7 which corresponds to meningiomas. For further explanation about boxplot, see Figure 8.5.*

Chapter 9

Conclusions

9.1 Summary

Magnetic resonance spectroscopy (MRS) has evolved rapidly over the past 15 years. It has been used specifically in the context of brain tumours and has shown very encouraging correlations between brain tumour type and spectral pattern. *In vivo* MRS enables the acquisition of metabolite concentration non-invasively, avoiding serious risks to brain damage, as in invasive techniques.

With the common use of Magnetic Resonance Imaging (MRI) for identifying the location and size of brain tumours, MRS complements it with the potential to provide detailed chemical information about metabolites present in the brain tissue and enable an early detection of abnormality.

Despite the technological advances in MRS data acquisition, there is a lack of practical, robust and accurate automatic classification tools for classifying brain tumours. The difficulties are partly due to the complexities in processing the high dimensional data, with the limited data available to date. Progresses in pattern recognition classifiers, including kernel-based methods, motivate the application of Least Squares Support Vector Machine classifiers to brain tumour classification problems.

A Least Squares Support Vector Machine is an advanced technique for classification with many useful properties [102]. The LS-SVM problem formulation has a global and unique solution. Moreover it is able to handle efficiently data in a large dimension and find an optimal separation based on a limited amount of data, which might be even smaller than the amount of input variables.

This advantage of kernel-based methods motivates the application to many

kinds of classification problems. In this thesis we report some results of experiments on recognizing brain tumours based on magnetic resonance data.

Within the framework of EU-funded INTERPRET research project, six research centers have acquired a significant number of short and long echo time proton MRS signals for brain tumour classification.

Least Squares Support Vector Machine classifiers have been applied to learn the patterns of several kind of MR data: long echo time MR spectra in Chapter 6, short echo time MR spectra in Chapter 7, and also MRI/MRSI dataa in Chapter 8. LS-SVM with linear kernel together with PCA/LDA and Corr/LDA constitute the linear classifiers, while LS-SVM with RBF kernel is used as nonlinear technique.

In the following are the main conclusions of this thesis:

- Automatic recognition of brain tumours based on MRS spectra has been shown and implemented. Advanced classification technique based on LS-SVM are applied to discriminate between several brain tumour types and grades. The performance LS-SVM classifiers using linear as well as RBF kernels are compared to the LDA classifier and classical Vapnik SVM classifier.
- All classifiers perform well in classifying brain tumours based on MRS spectra (AUC > 0.9 in Chapter 6 and $AUC > 0.95$ in Chapter 7), except for the most difficult binary classification to discriminate glioblastomas from metastases.
- The dimensionality of the spectra, the presence of noise and artefacts and the low amount of data available complicate the classification problem. Before classifying with LDA, the dimension of the data need to be reduced by PCA or using some prior knowledge. In contrary, kernel-based classifiers: SVM with linear kernel, LS-SVMs with linear and RBF kernel are able to solve the large dimensional classification problems, without the need of dimensionality reduction.
- Nonlinear classification using RBF kernel in some cases performs better than a linear classifier by adapting its boundary from linear hyperplanes.
- Dimensionality reduction, by selecting frequency regions or peak integration could possibly reduce the influence of disturbing noise and artefacts in the spectra. The described approach for selecting resonance peaks (see Chapter 6 and Chapter 7) makes it easier to reach a similar performance as using complete spectra. Although peak integration is a rough and inaccurate quantitation method, it yields a relatively high classification performance. This indicates the potential use of absolute quantitation (e.g. by a parametric time-domain model) prior to classification.

- Early results on the relative contribution of MRI in combination with MRSI in the discrimination of several brain tissue types show that binary classification based on MRSI and MRI is promising. The use of image information of MRI with spectral information from MRS results in improvement of the performance.

Application of LS-SVM in real life problems such as brain tumour recognition is a true benchmark of its capabilities. However, there are some factors which hamper the classifiers to perform optimally:

- Research centers within the INTERPRET project have joined their forces to collect the largest possible brain tumour dataset for better treatment of brain tumour problems. However, the amount of data per type is still very limited. Especially the amount of available metastases and astrocytomas grade II is low. This restricts the construction of classifiers with a high generalization capacity.
- Although many algorithms and methods have been developed to filter out noise and artefacts, removing them entirely from the spectra is not always perfect. The remaining artefacts, even after elimination of the dominating water peak, might affect important peaks in the spectra.
- there is always a large variation within each class and overlap between MRS spectra from in-vivo brain tumour acquisition. From Chapter 6 and Chapter 7, for example, we found that the mean spectra of glioblastomas and metastases show a very similar characteristic pattern.

9.2 Open problems

Although a lot of progress has been made, there are still many aspects that need further investigations.

- The choice of hyperparameter(s) of kernel-based classifiers is quite crucial. More research can be done in finding them in optimal way.
- Although LS-SVM has proven to be able to optimally handle high dimensional data, an automatic model order detection is expected. Therefore, fixed size LS-SVMs might be applied as an alternative.
- Up till now, feature selection in the spectral points only devised with some prior knowledge or manual selection. Therefore, a Bayesian framework of LS-SVM could be explored to automate the selection of the important spectral points.

- As alternative to the linear classifiers studied in this thesis, correlation based feature extraction methods [109], using e.g. Correlation Analysis or Canonical Correlation Analysis (CCA), can be applied. However care should be taken for the spectra alignment, baseline correction, prior to feature extraction.
- The use of prior knowledge, in combination with an advanced parametric model [119], may give a better performance than the simple quantification method applied in this thesis.

Appendix

A Articles in international journals

1. J.A.K. Suykens, J. De Brabanter, L. Lukas, J. Vandewalle, "Weighted least squares support vector machines : robustness and sparse approximation", Neurocomputing, Special issue on fundamental and information processing aspects of neurocomputing, vol. 48, no. 1-4, pp. 85-105, Oct. 2002.

2. L. Lukas, A. Devos, J.A.K. Suykens, L. Vanhamme, F.A. Howe, C. Majós, A. Moreno-Torres, M. van der Graaf, A.R. Tate, C. Arús, S. Van Huffel, Brain Tumour Classification based on long echo proton MRS signals, Artif Intell Med, accepted for publication (2003).

3. A. Devos, L. Lukas, J.A.K. Suykens, L. Vanhamme, F.A. Howe, C. Majós, A. Moreno-Torres, M. van der Graaf, A.R. Tate, C. Arús, S. Van Huffel, Classification of brain tumours using short echo time ^{1}H MRS spectra, *J. Magn. Reson.*, submitted for publication (2003).

B Articles in conference proceedings

1. J.A.K. Suykens, L. Lukas, P. Van Dooren, B. De Moor, J. Vandewalle, "Least squares support vector machine classifiers : a large scale algorithm", in Proc. of the European Conference on Circuit Theory and Design (ECCTD'99), Stresa, Italy, Sep. 1999, pp. 839-842.

2. J.A.K. Suykens, L. Lukas, J. Vandewalle, "Sparse approximation using least squares support vector machines", in Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS 2000), Geneva, Switzerland, May 2000, pp. II757-II760.

3. J.A.K. Suykens, L. Lukas, J. Vandewalle, "Sparse Least Squares Support Vector Machine Classifiers", in Proc. of he European Symposium on Artificial Neural Networks (ESANN'2000), Bruges, Belgium, 2000, pp. 37-42.

4. L. Lukas, J.A.K. Suykens, J. Vandewalle, "Least squares support vector machines classifiers : a multi two-spiral benchmark problem", Proc. of the ISSM 2001, ISSN 0855-8692, pp. 289-292.

5. L. Lukas, A. Devos, J.A.K. Suykens, L. Vanhamme, S. Van Huffel, A.R. Tate, C. Majós, C. Arús, "The use of LS-SVM in the classification of brain

tumors based on Magnetic Resonance Spectroscopic signals", in Proc. of the European Symposium on Artificial Neural Networks (ESANN'2000), Bruges, Belgium, 2000, pp. 131-136

6. L. Ameye, C. Lu, L. Lukas, J. De Brabanter, J.A.K. Suykens, S. Van Huffel, H. Daniels, G. Naulaers, H. Devlieger, "Prediction of Mental Development of Preterm Newborns at Birth Time using LS-SVM" in Proc. of the European Symposium on Artificial Neural Networks (ESANN'2000), Bruges, Belgium, 2000, pp. 167-172

7. L. Lukas, A. Devos, J.A.K. Suykens, L. Vanhamme, S. Van Huffel, A.R. Tate, C. Majós, C. Arús, The use of LS-SVM in the classification of brain tumors based on ^{1}H-MR Spectroscopy signals. *in* "Proceedings of the IEE Symposium on Medical applications of signal processing", Savoy Place, London, UK, October 7, 15/1-5 (2002).

C Abstracts for conferences

1. J.A.K. Suykens, L. Lukas, J. Vandewalle, Sparse Least Squares Support Vector Machines. 19th Benelux Meeting on System and Control, Mierlo, The Netherlands, March 1-3, 2000.

2. J.A.K. Suykens, J. De Brabanter, L. Lukas, Weighted Least Squares Support Vector Machines: robust nonlinear estimation and sparse approximation. 20th Benelux Meeting on System and Control, Houffalize, Belgium, March 1-3, 2000.

3. A. Devos, L. Vanhamme, L. Lukas, S. Van Huffel, J.A.K. Suykens, Least Squares Support Vector Machines for pre-operative pattern recognition of brain tumors using Magnetic Resonance Spectroscopy. 9th annual meeting of the Belgian Statistical Society, Oostende, Belgium, October 12-13, 2001.

4. L. Lukas, A. Devos, J.A.K. Suykens, L. Vanhamme, F.A. Howe, C. Majós, A. Moreno-Torres, M. van der Graaf, A.R. Tate, C. Arús, S. Van Huffel, Classification of brain tumours using short echo time ^{1}H MRS spectra, 3rd Belgian Day on Biomedical Engineering. October 17, 2003.

5. A. Devos, L. Lukas, J.A.K. Suykens, L. Vanhamme, S. Van Huffel, Classification of brain tumours using short echo time ^{1}H MRS spectra, 1st Flanders PhD symposium. December 11, 2003.

Bibliography

[1] M.A. Aizerman, E.M. Braverman, L.I. Rozonoer, "Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning", *Autom. Remote Control* 25 (1964).

[2] H. Barkhuijsen, R. De Beer, D. Van Ormondt, Improved algorithm for noniterative time-domain model fitting to exponentially damped Magnetic Resonance Signals, *J. Magn. Reson.* **73**, 553-557 (1987).

[3] R. Bartha, D.J. Drost, P.C. Williamson, Factors affecting the quantification of short echo in-vivo ^{1}H MR spectra: prior knowledge, peak elimination, and filtering, *NMR Biomed.* **12**, 205-216 (1999).

[4] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherface: recognition using class specific linear projection, *IEEE Trans. Pattern Analysis and Machine Intelligence* **19**(7), 711-720 (1997).

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

[6] F. Bloch, "Nuclear Induction", *Phys. Rev.*, 70:460-474 (1946).

[7] F. Bloch, W.W. Hansen, M. Packard, "The nuclear induction experiment", *Phys. Rev.*, 70:474-485 (1946).

[8] British Brain and Spine Foundation, M. Alexander (ed.), Brain Tumour: a guide for patients and carers. London, UK, 2001.

[9] H. Bruhn, J. Frahm, M.L. Gyngell, K.D. Merboldt, W. Hanicke, R. Sauter, C. Hmburger, Noninvasive differentiation of tumours with use of localised H-1 MR spectroscopy in vivo: initial experience in patients with cerebral tumours, *Radiology*, **172**, 541-548 (1989).

[10] I.D. Campbell, C.M. Dobson, R.J.P. Williams, A.V. Xavier, Resolution enhancement of protein PMR spectra using the difference between a broadened and a normal spectrum, *J. Magn. Reson.* **11**, 172-181, 1973.

[11] D. Canet, "Nuclear Magnetic Resonance: concepts and methods", John Wiley & Sons, Chichester, UK (1996).

[12] Comprehensive Cancer Monitoring Programme in Europe, http://www-dep.iarc.fr/hmp/camon.htm

[13] V. Cherkassky, F. Mulier, *Learning from data: concepts, theory and methods*, John Wiley and Sons, 1998.

[14] Y.D. Cho, G.H. Choi, S.P. Lee, J.K. Kim, ^{1}H-MRS metabolic patterns for distinguishing between meningiomas and other brain tumors, *Magn. Reson. Imag.* **21**, 663-672 (2003).

[15] P. Christiansen, P.B. Toft, P. Gideon, G.E.R. Danielsen, P. Ring, O. Henriksen, MR-visible water content in human brain: a proton MRS study, *Magn. Reson. Imag.* **12** (8), 1237-1244 (1994).

[16] C. Cortes, V. Vapnik, "Support Vector Networks", *Machine Learning*, **20**, 273-297 (1995).

[17] R. Courant, D. Hilbert, "Methods of Mathematical Physics", Interscience Publishers, New York (1953).

[18] J.P. Cousins, Clinical MR spectroscopy: fundamentals, current applications, and future potential, *Am J Roentgenol* **164**, 1337-1347 (1995).

[19] C. Daumas-Duport, B. Scheithauer, J. O'Fallon, P. Kelly, Grading of astrocytomas. A simple and reproducible method, *Cancer* **62** (10), 2152-2165 (1988).

[20] L. Debnath, P. Mikusiński, "Introduction to Hilbert spaces with applications (2nd ed.)", Academic Press (1999).

[21] P. Demaerel, Clinical applications of localized ^{1}H-NMR spectroscopy in intracranial tumours. Acta Biomedica Lovaniensia 112. Leuven Univ. Press. Belgium, 1995.

[22] A. Devos, L. Lukas, J.A.K. Suykens, L. Vanhamme, F.A. Howe, C. Majós, A. Moreno-Torres, M. van der Graaf, A.R. Tate, C. Arús, S. Van Huffel, Classification of brain tumours using short echo time ^{1}H MRS spectra, *J. Magn. Reson.*, submitted for publication (2003).

[23] R.O. Duda, P.E. Hart, D.G. Stork, "Pattern Classification", 2nd ed. John Wiley & Sons, (2001).

[24] R.R. Ernst, G. Bodenhausen, A. Wokaun, *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford University Press, Oxford, 1994.

[25] T. Ernst, J. Hennig, Coupling effects in volume selective 1H spectroscopy of major brain metabolites. *Magn Reson Med* **21**, 82-96 (1991).

[26] T. Ernst, R. Kreis and B.D. Ross, Absolute quantitation of water and metabolites in the human brain. I. Compartments and water, *J. Magn. Reson. B* **102**, 1-8 (1993).

[27] S. Fidler, A. Leonardis, Robust LDA classification by subsampling. http://www.cse.lehigh.edu/ rjm2/SACV/

[28] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179-188 (1936).

[29] R. Fletcher, *Practical methods of optimization*, Chichester and New York: John Wiley and Sons, 1987.

[30] M.J. Fulham, A. Bizzi, M.J. Dietz, H.H. Shih, R. Raman, G.S. Sobering, J.A. Frank, A.J. Dwyer, J.R. Alger, G. Di Chiro, Mapping of brain tumor metabolites with proton MR spectroscopic imaging: clinical relevance, *Radiology* **185** (3), 675:686 (1992).

[31] D.G. Gadian, Nuclear magnetic resonance and its applications to living systems. Clarendon Press, Oxford (1982).

[32] Geriatric Education for Emergency Medical Service. http://www.gemssite.com/

[33] V. Govindaraju, K. Young, A.A. Maudsley, Proton NMR chemical shifts and coupling constants for brain metabolites, *NMR Biomed.* **13**, 129-153 (2000).

[34] S.R. Gunn, Support Vector Machines for Classification and Regression. Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton, 1997.

[35] S.F.S. Halpin, MRS develops central role in clinical practice, Diagnostic Imaging Europe, Dec 2002/Jan 2003, 16-20, (2003).

[36] J.A. Hanley, B.J. McNeil, A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases, *Radiology* **148**, 839-843 (1983).

[37] Harrison's principles of internal medicine. http://harrisons.accessmedicine.com

[38] O. Henriksen, *In vivo* quantitation of metabolite of concentrations in the brain by means of proton MRS. NMR Biomed. **8**, 139-148 (1995).

[39] S. Herminghaus, T. Dierks, U. Pilatus, W. Möller-Hartmann, J. Wittsack, G. Marquardt, C. Labish, H. Lanfermann, W. Schlote, F.E. Zanella, Determination of histopathological tumor grade in neuroepithelial brain tumors by using spectral pattern analysis of in vivo spectroscopic data, *J. Neurosurg.* **98**, 74-81 (2003).

[40] J.R. Hesselink, "MR Spectroscopy of the brain", http://spinwarp.ucsd.edu/NeuroWeb/Text/mrs-TXT.htm

[41] J.C. Hoch, A.S. Stern, "NMR Data Processing", Wiley-Liss (1996).

[42] L. Hofmann, J. Slotboom, C. Boesch, R. Kreis, Model fitting of 1H-MR spectra of the human brain: incorporation of short-T1 components and evaluation of parametrized vs. non-parametrized models, *in* "Proceedings of the 7th Scientific Meeting of the International Society of Magnetic Resonance in Medicine (ISMRM)", Philadelphia, USA, May 22-28, 586 (1999).

[43] L. Hofmann, J. Slotboom, B. Jung, P. Maloca, C. Boesch, R. Kreis, Quantitative ^{1}H-Magnetic Resonance Spectroscopy of human brain: influence of composition and parameterization of the basis set in Linear Combination Model-fitting, *Magn. Reson. Med.* **48**, 440-453 (2002).

[44] F.A. Howe, S.J. Barton, S.A. Cudlip, M. Stubbs, D.E. Saunders, M. Murphy, P. Wilkins, K.S. Opstad, V.L. Doyle, M.A. McLean, B.A. Bell, J.R. Griffiths, Metabolic Profiles of Human Brain Tumors Using Quantitative In Vivo ^{1}H Magnetic Resonance Spectroscopy, *Magn. Reson. Med.*, **49**, 223-232 (2003).

[45] International network for Pattern Recognition of Tumours Using Magnetic Resonance (IST-1999-10310), http://carbon.uab.es/INTERPRET/.

[46] H. Ishimaru, M. Morikawa, S. Iwanaga, M. Kaminogo, M. Ochi, K. Hayashi, Differentiation between high-grade glioma and metastastic brain tumor using single-voxel proton MR spectroscopy. *Eur. Radiol.* **11**, 1784-1791 (2001).

[47] M. Julia-Sape, D. Acosta, C. Majós, P. Wesseling, J. Pujol, W. Gajewicz, B.A. Bell, J.R. Griffiths, C. Arús, Comparison between radiological and histopathological diagnoses in a multicentre international brain tumour magnetic resonance database. 19th Annual Meeting of the European Society for Magnetic Resonance in Medicine and Biology (ESMRMB02), Cannes, France, August 22-25, 2002, pp. 92-93.

[48] M. Just, M. Thelen, Tissue characterization with T1, T2, and proton density values: results in 160 patients with brain tumors, *Radiology*, **169**(3), 779-785 (1988).

[49] U. Klose, In vivo proton spectroscopy in presence of eddy currents, *Magn. Reson. Med.* **14**, 26-30 (1990).

[50] R. Kreis, T. Ernst, B.D. Ross, Absolute quantitation of water and metabolites in the human brain. II. Metabolite concentrations, *J. Magn. Reson. B* **102**, 9-19 (1993).

[51] H. Kugel, W. Heindel, R.-I. Ernestus, J. Bunke, R. du Mesmil, G. Friedmann, Human brain tumors: spectral patterns detected with localized H-1 MR spectroscopy, *Radiology*, **183**(3), 701-709 (1990).

[52] C. Ladroue, F.A. Howe, J.R. Griffiths, A.R. Tate, Independent component analysis for automated decomposition of in vivo magnetic resonance spectra, *Magn. Reson. Med.* **50**, 697-703 (2003).

[53] M. Law, S. Cha, E.A. Knopp, G. Johnson, J. Arnett, A.W. Litt, High-grade gliomas and solitary metastases: differentiation by using perfusion and proton spectroscopic MR imaging. *Radiology*, **222**, 715-721 (2002).

[54] J.A. Lazareff, R.K. Gupta, J. Alger, Variation of post-treatment H-MRSI choline intensity in pediatric gliomas, *J. Neuro-Oncol.* **41**(3), 291-298 (1999).

[55] X. Lecrerc, T.A.G.M. Huisman, A.G. Sorensen, The potential of proton magnetic resonance spectroscopy (^{1}H) in the diagnosis and management of patients with brain tumors, *Curr. Opin. Oncol.* **14**, 292-298 (2002).

[56] P.L. Lee, R.G. Gonzalez, Magnetic resonance spectroscopy of brain tumors, *Curr. Opin. Oncol.* **12**, 199-204 (2000).

[57] V. Lefournier, F. Szabo De Edelenyi, F. Estève, S. Grand, P. Bessou, K. Boubagra, J. Bosson, C. Segebarth, J. Le Bas, C. Rémy, Nosologic images for classification of brain tumors with ^{1}H MRSI: clinical performance. 19th Annual Meeting of the European Society for Magnetic Resonance in Medicine and Biology (ESMRMB02), Cannes, France, August 22-25, 2002, pp. 91-92.

[58] Z.P. Liang, P.C. Lauterbur, Principles of Magnetic Resonance Imaging, IEEE Press, New York (2000).

[59] J.C. Lindon, E. Holmes, J.K. Nicholson, Pattern recognition methods and applications in biomedical magnetic resonance, *Prog. Nucl. Mag. Res. Sp.* **39**, 1-40 (2001).

[60] MATLAB/C LS-SVMlab toolbox, http://www.esat.kuleuven.ac.be/sista/lssvmlab.

[61] L. Lukas, A. Devos, J.A.K. Suykens, L. Vanhamme, S. Van Huffel, A.R. Tate, C. Majós, C. Arús, The use of LS-SVM in the classification of brain tumors based on magnetic resonance spectroscopy signals, *in* "Proceedings of the European Symposium for Artifical Neural Networks (ESANN02)", Bruges, Belgium, April 24-26, 131-135 (2002).

[62] L. Lukas, A. Devos, J.A.K. Suykens, L. Vanhamme, S. Van Huffel, A.R. Tate, C. Majós, C. Arús, The use of LS-SVM in the classification of brain tumors based on ^{1}H-MR Spectroscopy signals. *in* "Proceedings of the IEE Symposium on Medical applications of signal processing", Savoy Place, London, UK, October 7, 15/1-5 (2002).

[63] L. Lukas, A. Devos, J.A.K. Suykens, L. Vanhamme, F.A. Howe, C. Majós, A. Moreno-Torres, M. van der Graaf, A.R. Tate, C. Arús, S. Van Huffel, Brain Tumour Classification based on long echo proton MRS signals, Artif Intell Med, accepted for publication (2003).

[64] C. Majós, J. Alonso, C. Aguilera, M. Serrallonga, J.J. Acebes, C. Arús, J. Gili, Adult primitive neuroectodermal tumor: proton MR spectroscopic findings with possible application for differential diagnosis, *Radiology*, **225**, 556-566 (2002).

[65] P. Mansfield, P. Morris, NMR imaging in medicine. *Adv. Mag. Res.*, S2:1-343 (1982)

[66] D.J. Manton, M. Lowry, S.J. Blackband, A. Horsman, Determination of proton metabolite concentrations and relaxation parameters in normal human brain and intracranial tumours, *NMR Biomed.* **8**, 104-112 (1995).

[67] I. Marshall, J. Higinbotham, S. Bruce, A. Freise, Use of Voigt lineshape for quantification of in vivo ^{1}H spectra, *Magn. Reson. Med.*, **37**, 651-657 (1997).

[68] A.M. Martinez, A.C. Kak, PCA versus LDA, *IEEE Trans. Pattern Analysis and Machine Intelligence* **23**(2), 228-233 (2001).

[69] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations", *Philos. Trans. Roy. Soc. London*, **209**, 415-446 (1909).

[70] Merriam-Webster Dictionary. http://www.m-w.com.

[71] Š. Mierisová, A. van den Boogaart, I. Tkáč, P. Van Hecke, L. Vanhamme and T. Liptaj, New approach for quantitation of short echo time *in-vivo* ^{1}H MR spectra of brain using AMARES, *NMR Biomed.* **11**, 32-39 (1998).

[72] V. Mlynárik, S. Gruber, E. Moser, Very short echo time proton MR spectroscopy of human brain with a standard transmit/receive surface coil, *Magn. Reson. Med.* **44**, 964-967 (2000).

[73] D.C. Montgomery and G.C. Runger, "Applied statistics and probability for engineers", John Wiley & Sons, Inc. (1994).

[74] S.K. Mukherji (editor), "Clinical Applications of Magnetic Resonance Spectroscopy", Wiley-Liss (1998).

[75] M. Murphy, A. Loosemore, A.G. Clifton, F.A. Howe, A.R. Tate, S.A. Cudlip, P.R Wilkins, J.R. Griffiths, B.A. Bell, The contribution of proton magnetic resonance spectroscopy (^{1}H MRS) to clinical brain tumour diagnosis, *Br. J. Neurosurg.* **16**(4), 329-334 (2002).

[76] National Cancer Institute.
http://cancer.gov/dictionary/

[77] National Institute on Aging.
http://www.healthandage.net/html/min/nih/content/hyperthermia.htm

[78] S.J. Nelson, Multivoxel magnetic resonance spectroscopy of brain tumors, *Mol. Cancer Ther.* **2**, 497-507 (2003).

[79] N.A. Obuchowski, Receiver operating characteristic curves and their use in radiology, Radiology, **229**, 3-8 (2003).

[80] K.S. Opstad, J.R. Griffiths, B.A. Bell, F.A. Howe, In vivo lipid T2 relaxation time measurements in high-grade tumors: differentiation of glioblastomas and metastases. Proceedings of the Eleventh Scientific Meeting & Exhibition (ISMRM 03), Toronto, Canada, May 10-16 2003, submitted.

[81] J.E. Parisi, B.W. Scheithauer, Glial tumors. In: "Principles and practice of neuropathology", J.S. Nelson, J.E. Parisi and S.S. Shochet (eds.), St Louis, Mosby (1993).

[82] K. Pelckmans, J.A.K. Suykens, T. Van Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor, J. Vandewalle, LS-SVMlab Toolbox User's Guide, Internal Report 02-145, ESAT-SISTA, K.U.Leuven, Leuven, Belgium (2002).

[83] H. Poptani, J. Kaartinen, R.K. Gupta, M. Niemitz, Y. Hiltunen, R.A. Kauppinen, Diagnostic assessment of brain tumours and non-neoplastic brain disorders in vivo using proton nuclear magnetic resonance spectroscopy and artificial neural networks, *J. Cancer Res. Clin. Oncol.* **125**, 343-349 (1999).

[84] M.C. Preul, Z. Caramanos, D.L. Collins, J.-G. Villemure, R. Leblanc, A. Olivier, R. Pokrupa, D.L. Arnold, Accurate, noninvasive diagnosis of human brain tumors by using magnetic resonance spectroscopy, *Nature Medicine*, **2**(3), 323-325 (1996).

[85] S. Provencher, Estimation of metabolite concentrations from localized *in-vivo* proton NMR spectra, *Magn. Reson. Med.*, **30**, 672-679 (1993).

[86] E.M. Purcell, H.C. Torrey, R.V. Pound, Resonance absorption by nuclear magnetic moments in a solid. *Phys. Rev.*, **69**, 37-38 (1946).

[87] A.C. Rencher, "Methods of multivariate analysis", Wiley series in Probability and Mathematical Statistics (1995).

[88] J.A. Rice, "Mathematical Statistic and Data Analysis", Duxbury Press (1995).

[89] B.D. Ripley, "Pattern Recognition and Neural Networks", Cambridge University Press (1996).

[90] R. Schalkoff. "Pattern Recognition: statistical, structural and neural approaches", John Wiley & Sons, Inc. (1992).

[91] D. Schiffer, "Brain tumors", 2nd edition, Springer-Verlag (1997).

[92] U. Seeger, U. Klose, I. Mader, W. Grodd, T. Nägele, Parametrized evaluation of macromolecules and lipids in proton MR spectroscopy of brain diseases, *Magn. Reson. Med.*, **49**, 19-28 (2003).

[93] A.W. Simonetti, W.J. Melssen, M. van der Graaf, A. Heerschap, L.M.C. Buydens, Automated Correction of Unwanted Phase Jumps in Reference Signals Which Corrupt MRSI Spectra after Eddy Current Correction, *J Magn Reson*, **159**, 151-157 (2002).

[94] A.W. Simonetti, W.J. Melssen, M. van der Graaf, A. Heerschap, L.M.C. Buydens, A New Chemometric Approach for Brain Tumor Classification using Magnetic Resonance Imaging and Spectroscopy, *Anal. Chem.*, **75**(20), 5352-5361 (2003).

[95] J. Slotboom, C. Boesch, R. Kreis, Versatile frequency domain fitting using time domain models and prior knowledge, *Magn. Reson. Med.*, **39**, 899-911 (1998).

[96] I.C.P. Smith, L.C. Stewart, Magnetic resonance spectroscopy in medicine: clinical impact, *Prog. Nucl. Mag. Reson. Spectrosc.* **40**, 1-34 (2002).

[97] A.J. Smola and B. Schollköpf, A tutorial on support vector regression, NeuroCOLT2 Technical Report Series NC2-TR-1998-030, ESPRIT working group on Neural and Computational Learning Theory, NeuroCOLT2 (1998).

[98] B.J. Soher, K. Young, A.A. Maudsley, Representation of strong baseline contribution in ^{1}H MR spectra, *Magn. Reson. Med.*, **45**, 966-972 (2001).

[99] A. Srinivasan, Note on the location of optimal classifiers in n-dimensional ROC space, Technical Report PRG-TR-2-99 Oxford University Computing Laboratory, Oxford, England (1999).

[100] MATLAB Support Vector Machines toolbox.
http://www.isis.ecs.soton.ac.uk/isystems/kernel.

[101] J.A.K. Suykens, J. Vandewalle, Least Squares Support Vector Machine Classifiers, *Neur. Proc. Lett.* **9**(3), 293-300 (1999).

[102] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, "Least Squares Support Vector Machines", World Scientific Publishing Co., Singapore (2002).

[103] J.A. Swets, ROC analysis applied to the evaluation of medical imaging techniques, *Invest. Radiol.* **14**(2), 109-121 (1979).

[104] D.L. Swets, J.J. Weng, Using discriminant eigenfeatures for image retrieval, *IEEE Trans. Pattern Analysis and Machine Intelligence* **18**(8), 831-836 (1996).

[105] F. Szabo De Edelenyi, C. Rubin, F. Estève, S. Grand, M. Décorps, V. Lefournier, J.-F. Le Bas, C. Rémy, A new approach for analyzing proton magnetic resonance spectroscopic images of brain tumors: nosologic images, *Nature Medicine* **6**, 1287-1289 (2000).

[106] F. Szabo De Edelenyi, F. Estève, C. Rémy, L. Buydens, Application of independent component analysis to ^{1}H MR spectroscopic imaging exams of brain tumors, *in* "Proceedings of the 19th Annual Meeting of the European Society for Magnetic Resonance in Medicine and Biology (ESMRMB02)", Cannes, France, August 22-25, 91 (2002).

[107] A.R. Tate, J.R. Griffiths, I. Martínez-Pérez, A. Moreno, I. Barba, M.E. Cabañas, D. Watson, J. Alonso, F. Bartumeus, F. Isamat, I. Ferrer, F. Vila, E. Ferrer, A. Capdevila, C. Arús, Towards a method for automated classification of ^{1}H MRS spectra from brain tumours, *NMR Biomed.*, **11**, 177-191 (1998).

[108] A.R. Tate, J.R. Griffiths, F.A. Howe, J. Pujol, C. Arús, Differentiating Types of Human Brain tumours by MRS. A Comparison of Pre-processing Methods and Echo Times, *in* "Proceedings of the Ninth Scientific Meeting & Exhibition (ISMRM01)", Glasgow, Scotland, April 21-27, 2284 (2001).

[109] A.R. Tate, C. Majós, A. Moreno, F.A. Howe, J.R. Griffiths, C. Arús, Automated classification of short echo time in in vivo ^{1}H brain tumor spectra: a multicenter study, *Magn. Reson. Med.* **49**, 29-36 (2003).

[110] The Brain Tumor Society.
http://www.tbts.org.

[111] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Academic Press, San Diego (1999).

[112] M.R. Tosi, R. Ricci, G. Bottura, V. Tugnoli, In vivo and in vitro magnetic resonance spectroscopy investigation of an intracranial mass. *Oncology Report*, **8**, 1337-1339 (2001).

[113] J. Tou, R. Gonzalez, Pattern Recognition Principles, Addison Wesley, Reading Mass. (1979).

[114] J.P. Usenius, P. Vainio, J. Hernesniemi, R.A. Kauppinen, Choline-containing compounds in human astrocytomas studied by 1H NMR spectroscopy in vivo and in vitro, *J. Neurochem.*, **63**(4), 1538-1543 (1994).

[115] T. Usenius, J.P. Usenius, M. Tenhunen, P. Vainio, R. Johansson, S. Soimakallio, R. Kauppinen, Radiation-induced changes in human brain metabolites as studied by ^{1}H nuclear magnetic resonance spectroscopy in vivo. *Int. J. Radiat. Oncol. Biol. Phys.* **33**(3), 719-724 (1995).

[116] T. Van Gestel, J.A.K. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, J. Vandewalle, Benchmarking least squares support vector machines. Machine Learning (in press).

[117] L. Vanhamme, A. van den Boogaart and S. Van Huffel, Improved Method for Accurate and Efficient Quantification of MRS Data with Use of Prior Knowledge, *J Magn Reson*, **129**, 35-43 (1997).

[118] L. Vanhamme, "Advanced time-domain methods for nuclear magnetic resonance spectroscopy data analysis", PhD thesis, K.U. Leuven (1999).

[119] L. Vanhamme, T. Sundin, P. Van Hecke and S. Van Huffel, MR spectroscopy quantitation: a review of time-domain methods, *NMR Biomed.*, **14**, 233-246 (2001).

[120] L. Vanhamme, ESAT-SCD K.U. Leuven, personal communication, (2003).

[121] V. Vapnik and A. Lerner, Pattern recognition using generalized portrait method, *Automation and Remote Control*, **24** (1963).

[122] V. Vapnik and A. Chervonenkis, A note on class of perceptron, *Automation and Remote Control*, **24** (1964).

[123] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York (1995).

[124] V. Vapnik, Statistical Learning Theory, John Wiley & Sons (1998).

[125] V. Vapnik, "The support vector method of function estimation," In *Nonlinear Modeling: advanced black-box techniques*, Suykens J.A.K., Vandewalle J. (Eds.), Kluwer Academic Publishers, Boston, pp.55-85, 1998.

[126] C.-Z. Ye, J. Yang, D.-Y. Geng, Y. Zhou, N.-Y. Chen, Fuzzy rules to predict degree of malignancy in brain glioma, Med. Biol. Eng. Comput. **40**, 145-152 (2002).

[127] T.Y. Young, T.W. Calvert, Classification, Estimation and Pattern Recognition, Elsevier, New York (1974).

Lukas was born on July 6, 1973 in Purwokerto (Banyumas), Indonesia. He received the bachelor degree in Electrical Engineering (major in computer engineering) from the Institut Teknologi Bandung, Indonesia in April 1995. Afterwards, he worked two years as lecturer at the Universitas Katolik Atma Jaya, Jakarta, Indonesia. In 1997 he continued his study at the Katholieke Universiteit Leuven, Belgium on a K.U. Leuven scholarship programme. He obtained his Master degree in Artificial Intelligence in September 1998. After completing one year predoctoral study, he started working as a researcher at the Electrical Engineering Department ESAT of the K.U. Leuven first on the extension of K.U. Leuven scholarship, and from October 2002 to November 2003 on a grant from the National Fund for Scientific Research FWO projects.