NOVEL APPLICATIONS OF MACHINE LEARNING IN PIPELINE INSPECTION AND NEUROSCIENCE

NOVEL APPLICATIONS OF MACHINE LEARNING IN PIPELINE INSPECTION AND NEUROSCIENCE

By

AHMAD KHODAYARI-ROSTAMABAD, B.Sc., M.Sc.

A Thesis
Submitted to the Department of Electrical and Computer Engineering
and to the School of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2010) MCMASTER UNIVERSITY
(Electrical and Computer Engineering) Hamilton, Ontario

TITLE:	**Novel Applications of Machine Learning in Pipeline Inspection and Neuroscience**
AUTHOR:	Ahmad Khodayari-Rostamabad B.Sc. (Electrical Engineering, majoring in Electronics), Iran University of Science and Technology, Tehran, Iran. M.Sc. (Electrical Engineering, majoring in Signal Processing and Telecommunications), Tarbiat Modares University, Tehran, Iran.
SUPERVISOR:	James P. Reilly, Professor, Ph.D., P.Eng.
CO-SUPERVISOR:	Gary M. Hasey, Associate Professor, M.D., FRCP(C)
NUMBER OF PAGES:	xix, 158

To my wife Leila,
to my son Daniel,
and to my parents

Abstract

In this thesis we develop and evaluate automated "expert systems" for two applications: (i) gas/oil pipeline inspection using magnetic flux leakage information, (ii) treatment efficacy prediction and medical diagnosis using electroencephalograph (EEG) and clinical information. Both applications share the same methodology and procedure as they employ machine learning methods which learn their decision models using the training data (or past examples in real life/environment).

The magnetic flux leakage (MFL) technique is commonly used for non-destructive testing (NDT) of oil and gas pipelines which are mostly buried underground. This testing involves the detection of metal defects and anomalies in the pipe wall, and the evaluation of the severity of these defects. The difficulty with the MFL method is the extent and complexity of the analysis of the MFL images. In this thesis we show how modern machine learning techniques can be used to considerable advantage in this respect.

The problem of identifying in advance the most effective treatment agent for various psychiatric conditions remains an elusive goal. To address this challenge, an automated medical expert system is designed and then evaluated. The system is capable of predicting the treatment response for each individual patient at the outset of a therapy (i.e., using pre-treatment information) thus improving therapeutic efficiency and reducing personal and economic costs. Our experiments are focused on treatment planning and diagnosis of mood disorders and psychiatric illnesses. Through different experiments, we have shown that it is possible to predict treatment efficacy of a 'selective serotonin reuptake inhibitor' (SSRI) antidepressant and 'repetitive transcranial magnetic stimulation' (rTMS) therapies for patients with treatment-resistant major depressive disorder (MDD) or major depression. The predictions are based on pre-treatment quantitative EEG measurements. Also, prediction of post-treatment schizophrenia symptomatic scores, using pre-treatment EEG data, showed significant performance in patients treated with the drug clozapine. Clozapine is an antipsychotic medication of superior effectiveness in treating Schizophrenia but has several potentially severe side effects.

Medical diagnosis is the second problem we consider in the neuroscience aspects of this thesis. In this research, an automated digital medical diagnosis methodology is developed to estimate/detect the type of a disease or illness that a patient is suffering. This intelligent diagnostic system can assist the physician/clinician by offering a second opinion on diagnosis. Several complex psychiatric illnesses may have many common symptoms and accurate diagnosis can, at times, be very difficult. Efficient diagnosis helps by avoiding prescription of wrong therapy/treatment to a patient. In our limited experiments, EEG data is used to make a diagnosis for distinguishing between various psychiatric illnesses including MDD, schizophrenia, and the depressed phase of bipolar affective disorder (BAD).

In all problems considered in this thesis, specifically the neuroscience problem, a large number of candidate features are extracted from measurement data but most candidate features are found to be irrelevant and have little or no discriminative power. Finding a few most discriminating features that guarantee numerical efficiency and obtain a smooth and generalizable decision function, is a major challenge in this research. In this thesis, feature selection methods based on mutual information or Kullback-Leibler (KL) distance is employed to find the most statistically relevant features. For the multi-class diagnosis problem, to improve performance, a feature selection procedure denoted as *feature combination* feature selection is used which first finds discriminating features in all binary classification combinations, and then combines them into a larger feature subset to make a final multi-class decision. The two–dimensional (2D) representation of the feature data is also found to be useful for clustering analysis. The overall method was evaluated using a nested cross–validation procedure for which over 80% average prediction performance is obtained in all experiments. The results indicate that machine learning methods hold considerable promise in solving the challenging problems encountered in the two applications of concern.

Acknowledgements

All deepest thanks and gratitude are to Almighty God, the Most Merciful, the Most Compassionate.

During my PhD studies, I was very lucky to work with a great supervisor, Prof. James P. (Jim) Reilly, who taught me more than the academic work. I highly appreciate him for everything I learned from him. My co-supervisor, Dr. Gary M. Hasey provided the motivation and inspiration for me for the biomedical research and I really appreciate his continuous support. Apart from Jim and Gary, I also was lucky to work with three other great professors: Prof. Maung Min-Oo, Dr. Duncan MacCrimmon and Dr. Hubert deBruin. I am thankful and I greatly acknowledge the time and energy that Jim, Gary, Maung, Duncan and Hubert spent with me during several discussions and meetings on my research work and for the generous and friendly support they provided to me.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). For the pipeline inspection project, the Intratech Inline Inspection Services Ltd. (Mississauga, ON, Canada) supported the research, provided the data and the technical information. I would like to thank them, particularly James R. (Jim) Hare and Sabir Pasha of Intratech and the president of the company, Ron Thompson.

During my graduate studies at ECE department of McMaster University, I received useful advice and comments from Prof. Natalia Nikolova, Prof. Thia Kirubarajan, Dr. Timothy R. Field, Prof. Simon Haykin and Prof. Timothy (Tim) Davidson, which I am grateful for.

I also would like to thank my friends and colleagues in the research group: Nazanin Samavati, Amin Zia, Reza Khalaj Amineh, Siamak Salari, Nasser A. Mourad, Derek Yee, Michael Daly, Krzysztof Maryan, and Mark Archambeault.

During the last few years, I had great friends who were kindly supportive and helpful when I needed them. I had many happy and joyful moments with them and I am greatly thankful for their help and friendship: Juan Rodriguez Hernandez, Patrick Fayard, Pavel (Paul) Abumov, Peyman Setoodeh, Mehdi

Ranjbaran, Seyed Vahid Azhari, Ghodratollah (Hamzeh) Hashemi-Motlagh, Abbas Ebrahimi Moghadam, Mohammad-Reza Heydarian Firouzab, Mehdi Hosseinifar, Fathollah (Hanif) Mahmoodzadeh, Morteza Amoorezaei, Mehran Nekuii, Mojtaba Hodjat Shamami, James Anderson, Dimitry Akselrod, Ratnasingham Tharmarasa (Thamas), and Ienkaran Arasaratnam (Haran).

In the neuroscience and biomedical engineering project, I greatly acknowledge the help of Rose Marie Mueller, Cathy Ivanski, Jackie Boyd, Sandra Chalmers and several other nurses in the Mood Disorders Program of St. Joseph Hospital (Hamilton, ON, Canada) who were involved in doing the clinical experiments and in collecting the clinical data. Also, from March 2010, I am involved in a new antidepressant medication study where prediction of response to 3 antidepressant drugs will be investigated. Even though this study is in the data collection phase and is not part of my thesis, but from March 2010, my salary is paid through this study which is supported by The Magstim Company Ltd., Carmarthenshire, Wales, UK, which I am grateful for.

Finally, I would like to thank my family and my parents for their continuous and generous support and for their patience during several years of my study.

List of Acronyms

2D	Two Dimensional
ANN	Artificial Neural Network
BAD	Bipolar Depressive Disorder (or Bipolar Depression)
BD	Bipolar Depression(or Bipolar Depressive Disorder)
CV	Cross–Validation
ECG	Electrocardiography
EC	Eyes-Closed (condition when recording EEG signals)
ECT	Electroconvulsive Therapy
EEG	Electroencephalography
EM	Expectation Maximization (method)
EMG	Electromyography
EO	Eyes-Open (condition when recording EEG signals)
fMRI	Functional Magnetic Resonance Imaging
GMM	Gaussian Mixture Model
ILI	Inline Inspection (tool)
KL	Kullback-Leibler Distance
KPCA	Kernel Principal Component Analysis
KPLS	Kernel Partial Least Squares
KPLSR	Kernel Partial Least Squares Regression
LnO	Leave–n–Out Cross-Validation Procedure
MAP	Maximum *A-Posteriori* (estimation)
MDD	Major Depressive Disorder, (or Major Depression)
MEG	Magnetoencephalography
MFA	Mixture of Factor Analysis
MFL	Magnetic Flux Leakage
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MRI	Magnetic Resonance Imaging
NN	Nearest Neighbor (classification method)
NDT	Non-Destructive Testing
PCA	Principal Component Analysis

PDF	Probability Density Function
PIG	Pipeline Inspection Gauge
PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
PSD	Power Spectral Density
QEEG	Quantitative Electroencephalography
rTMS	Repetitive Transcranial Magnetic Stimulation (or TMS)
RBF	Radial Basis Function
RBFN	Radial Basis Function Network
RLS	Regularized Least Squares
SNR	Signal to Noise Ratio
SSRI	Selective Serotonin Reuptake Inhibitors
SVM	Support Vector Machines (classification method)
SVR	Support Vector Regression
TMS	Transcranial Magnetic Stimulation (or rTMS)

List of Notations and Symbols

$\|\| \cdot \|\|$	Euclidian norm of a vector
$\lceil x \rceil$	the smallest integer larger or equal to x
$\lfloor x \rfloor$	the largest integer smaller or equal to x
$\langle \mathbf{x}, \mathbf{y} \rangle$	inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$
$\mathcal{N}(\mu, \sigma)$	Gaussian probability distribution with mean μ and standard deviation σ
$\mathrm{E}[\cdot]$	expectation operator
$D_{\mathrm{KL}}(\mathcal{U}\|\|\mathcal{V})$	Kullback-Leibler (KL) distance between random variables $\mathcal{U}$ and $\mathcal{V}$
$\mathfrak{M}(\mathcal{U}\|\|\mathcal{V})$	Mutual information (MI) of two random variables $\mathcal{U}$ and $\mathcal{V}$
$\mathbb{R}$	set of all real numbers
$\mathbb{R}^{N_r}$	N_r-dimensional Euclidean space
N_c	number of candidate (and possibly) discriminating features
N_r	number of reduced-dimensional vector, $N_r < N_c$
$\tilde{\boldsymbol{x}}$	feature or measurement vector (after feature extraction), $\tilde{\boldsymbol{x}} \in \mathbb{R}^{N_c}$
$\mathbf{x}$	reduced-dimensional feature vector (i.e., after feature selection), $\mathbf{x} \in \mathbb{R}^{N_r}$
y	target variable (or output variable)
N_p	number of classes (or patterns)
M_t	number of training samples in the available *training set* denoted by $\mathcal{D}$
M_t^j	number of training samples corresponding to class j
M	number of subjects (or patients)
Θ	set of parameters

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The modern version of machine learning[1], which is alternatively denoted by pattern recognition, artificial intelligence or data mining, was introduced after 1955. See e.g. [2]. From the beginning, the two major research topics of the field were: (i) artificial neural network methods (such as multilayer perceptron neural network with back error propagation learning) and (ii) statistical pattern recognition methods (such as Bayesian classifiers). One of main approaches was to model the mechanisms of human learning. Therefore, the starting point was to understand the brain, the neural system and how the neurons interact with each other, then to build similar computational models using electronic computers. Physicians, neuroscientists and psychologists like Donald O. Hebb (a Canadian psychologist) played important roles in this line of research. See [3] for a review. The development of mathematical analysis of learning and inference processes, and the empirical learning theories further focused on solving the practical problems using computational models and algorithms which might not be directly related to observations in biological neural systems, see e.g., [4–8]. A further research topic in this field is developing expert systems (or knowledge-based and intelligent systems) that extract the information hidden in the measured data and perform automatic processes in complex real-life applications.

The focus of this thesis is developing *machine learning* and intelligent inference methods to solve two problems:

1. Gas/Oil pipeline inspection, in which the problem is to use magnetic flux leakage (MFL) data to detect metal-losses and cracks in pipelines.

2. Cognitive medical expert (CME) system, in which the problem is to

[1] By the term *modern*, we mean what is implemented using electronic computers. The invention of computers allowed the implementation of machine learning theories.

develop a system to properly use the clinical and laboratory information (measured pre-treatment) to diagnose psychiatric illness, and then to predict whether or not a set of candidate treatments might be effective for the patient.

In the following we discuss each problem introduced above, and review the previous literature related to each topic.

1.1 Overview of Machine Learning Procedure

We now present a brief overview of the machine learning process used to solve the problems of concern in this thesis. Details will be discussed in Chapter 2. A necessary component of this process is the collection of a *training set.* The machine learning process learns the knowledge from the information hidden in the training set. The training set consists of M_t input or measured data samples. The training set also includes the set of output or target variables y_i, $i = 1, \ldots, M_t$ corresponding to each input feature sample $\tilde{\boldsymbol{x}}_i$, to be explained later. In the automatic pipeline inspection system, the input data is the measured magnetic flux leakage (MFL) image data in addition to some other measurements from the pipeline, and the output variable is metal defect class associated with the MFL image, for example. In the cognitive medical expert system, the input data is the clinical information from the subject (which includes the electroencephalography (EEG) information, for example).For the medical diagnosis problem, the output variable is the class of psychiatric illness that the subject is suffering. For treatment-response prediction problem, the output variable is the indicator of response (or no response) to the treatment administered to the subject. The response to treatment is measured by an expert clinician after the subject completes a course of treatment. Machine learning can also be used to construct models that do regression or interpolation, where the target or output variable is continuous.

There are three phases in a machine learning procedure. These are the design, operational and evaluation phases. In the following, we briefly describe each phase. The details will be further described in Chapter 2.

The design phase, alternatively denoted by learning phase, which consists of the feature extraction, feature selection and classification components, is now described. The first step is to extract numerical candidate features or quantitative attributes from measured data. These features include statistical attributes, spatial and temporal data model attributes, time-series model attributes, dynamic process model attributes, etc., of the input data. The number N_c of such candidate features can be quite large. The feature extraction process is applied over all input data in the training set. The result of the feature extraction process is a set of vectors $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{N_c}$, $i = 1, \ldots, M_t$, where M_t is the number of training samples. After extracting candidate features, the second step in the design phase is 'dimensionality reduction' in which a small set of the most relevant features are calculated and stored for further processing. *Feature selection* is one kind of dimensionality reduction in which we select the most discriminative and relevant features from several candidates. Typically, only a relatively small number of the extracted candidate features bear any significant statistical relationship with the output or target variable. We therefore identify those features which share the strongest statistical dependencies with the target variable. The result of the feature selection process is to reduce the number N_c of candidate features to a much smaller number N_r of most relevant features. The output of the feature selection process is a set of indices that identify which of the N_c candidate features are to be included in the set of N_r most relevant features. The feature selection process yields a set of vectors, $\mathbf{x}_i \in \mathbb{R}^{N_r}$, $i = 1, \ldots, M_t$. Each of these vectors correspond to a point in an N_r–dimensional feature space. The selection of "good" features; i.e., features with greater statistical dependence on the outcome variable, results in improved performance.

The next step in the design phase of the prediction process is the specification of the classifier or regressor. The specification (i.e., the "learning") of the classifier or regressor involves determining a function $f : \mathbb{R}^{N_r} \longmapsto \mathbb{R}$ which inputs a reduced feature vector $\mathbf{x}$ and outputs the corresponding target value y.

The Operational Phase is where the designed model is tested. Once the machine learning process is designed, it may be applied e.g., in an operational mode in solving real life problems. Here, the input data is collected from the test samples, and the set of reduced features identified in the design phase are computed from the measured data, to give a sequence of feature vectors. These feature vectors are fed into the classifier or regression function which is specified from the classifier parameters determined in the design phase. The classifier outputs the predicted target variables associated with the input data.

In the current situation however, we are interested in evaluating the performance of the machine learning procedure resulting from the design phase, using the available data. This is the Evaluation Phase. One of the popular methods in this respect, using the whole available training set, is the leave–n–out cross-validation procedure, where n samples at a time are sequentially removed from the available training set. The feature selection and classifier design processes are then executed using all remaining $M_t - n$ data samples.

The resulting machine learning structure is then tested using the omitted samples. The classifier output is then compared to the known target values of the omitted n samples, and a performance tally is recorded. The process repeats, each time omitting a different set of n samples, until all samples have been omitted once. The overall performance figure for the prediction process is then the aggregate performance over all iterations (or folds) of the leave–n–out cross-validation process. With this method, we test over all available data and in each trial we use the largest possible training set. Further, the method is "fair", since the tested data is not part of the training set used in the design phase. The other alternative when we have a large training set is to divide the available training set into two independent and non-overlapping subsets: a single training subset and a test subset. The machine learning system is then designed using the training subset and then tested on the independent test subset.

1.2 Using MFL Data for Pipeline Inspection

Due to the adverse effect of environmental damage that could result from a pipeline leak or catastrophic failure, pipelines must be routinely evaluated for integrity. The logistics and cost of shutting down a pipeline for inspection is prohibitive, so inspection devices, referred to as *pipeline inspection gauges* (PIGs) in the trade, are designed for autonomous operation in the pipeline, and are propelled along the pipe by normal transport flow. PIGs are also referred to as *in-line-inspection* (ILI) tools.

The ILI tool magnetizes the pipe wall as it travels down the pipe. Hall effect or coil sensors measure the localized magnetic flux leakage intensity along the pipe wall. Defects in the pipe wall cause irregularities in the magnetic field, that are detected by an array of these sensors, placed at regular intervals around the circumference of the inside of the pipe wall. Thus, *magnetic flux leakage* (MFL) testing is based on detecting the magnetic field that leaks from a pipe's wall in the vicinity of surface and subsurface flaws. See [9–12] for reviews of history, problems and concerns.

The stored magnetic image data is then analyzed off-line. Of special interest to pipeline operators is the extent and location of various defects that can adversely affect the integrity of the pipeline and its operation. Some of these defects include corrosion[2], deformations, fatigue, hairline cracks, dents, buckles, de-laminations, and faulty welds. The results are used to determine repair

[2] Corrosion turns steel into non-ferromagnetic iron oxide that locally reduces permeability.

and replacement priorities for a pipeline operator. Typically, the image analysis portion of the inspection procedure is done exclusively by human operators referred to as *non–destructive testing* (NDT) technicians. As a result of the manual nature of the inspection process, it is inherently slow and error–prone. Thus there is a strong motivation in the industry to automate the inspection process using methods such as machine learning as we propose.

Once defects have been identified, an equally important problem is the assessment of the size or severity of the defect (sizing). Estimated defect depths are used to determine the safety of the pipe, and to calculate accurate *maximum allowable operating pressure* (MAOP) of the oil/gas flowing through the line. In this work, we propose machine learning techniques for defect detection and sizing using real MFL images recovered from actual pipeline inspections.

There have been several previous works in defect detection and sizing using the MFL technique. Machine learning has been used previously in this context in [13] which evaluates the use of multilayer Perceptrons (MLP) for pattern recognition of MFL signals in weld joints of pipelines. Inverse modelling approaches are used in [14] for defect-shape construction. In [15], an iterative inversion method is proposed, using a multi-resolution wavelet transform and a radial basis function neural network (RBFN), for the prediction of 3-D defect geometry from MFL measurements.

The inversion procedure in [16] employs the 'space mapping' (SM) methodology. Space mapping shifts the optimization burden from a computationally expensive accurate or fine model like a finite-element method (FEM) simulation, to a less accurate or coarse but fast model, like analytical formulas.

The works of [17,18] present a modified wavelet transform domain adaptive finite impulse response (FIR) filtering algorithm for removing seamless pipe noise (SPN) in the MFL data. Papers [19,20] give wavelet based approaches to this problem for both de-noising and classification. In [21], an adaptive method for channel equalization (to compensate the mismatch between sensors) in MFL inspection is presented using the finite impulse response filter. Reference [22] presents a model based probability of detection (POD) evaluation technique for MFL inspection of natural gas transmission pipelines.

Comparison of the MFL method to ultrasonic methods is given in [23]. See also [24–26] for some other related topics.

The work of [27] presented a fast direct method that provides estimation of the crack parameters (orientation, length, and depth) in rectangular surface-breaking cracks, based on measurements of one tangential component of the magnetic field. The difficulty with this and other defect depth estimation methods is that the technique is limited to regular rectangular cracks. Real defects occurring naturally in pipelines virtually never exhibit this simple form of geometry.

A traditional approach for crack depth estimation is to use 'calibration curves or surfaces'. For example, [27] constructed a calibration surface that shows MFL signal strength versus two parameters of crack– depth and crack– length. In [28], three two-dimensional curves are drawn to show MFL signal sensitivity, in which each curve shows the effect of a specific test parameter on the magnitude of the MFL signal. In a different approach, [29] presented an analytic method for depth estimation in rectangular slots. In general, calibration curves/surfaces are 2D (or 3D) plots showing the effect of one measured parameter (or at most 2 parameters/features) on the magnitude of the measured MFL signal, or on the defect depth. However, the problem in real life is that cracks (and metal-losses) exhibit irregular and complex geometries and therefore cannot be characterized with just a few simple parameters. Therefore methods which assume a specific geometry or require representation in terms of a a few parameters will not perform well with real defects.

In this thesis, we extend the above works to develop high performance machine learning methods (as discussed in Chapter 2) for both defect detection and sizing of MFL images. We show that these methods, if properly executed, can give very satisfactory performance for both defect detection and depth estimation for real defects with irregular (or arbitrary) geometric configurations. The details of pipeline inspection and experimental results will be described in Chapter 3.

1.3 Treatment Efficacy Prediction and Treatment Planning

The problem in treating complex psychiatric illnesses like major depression or schizophrenia is that although patients may appear to have similar clinical characteristics, a treatment that works for one patient, may not work well for others. This suggests that current medical diagnostic systems and testing procedures may not be sufficiently sensitive to detect subtle but highly relevant differences between patients presenting with similar complaints.

The wide array of psychological, physical, hematological, radiological and other laboratory tests and assessments generate a very large information set that the busy clinician may find challenging to compile and process. The vast amounts of data that may be generated are typically viewed in isolation or as part of simple syndromatic clusterings. It is possible that a great deal of salience with respect to diagnosis and treatment that is embedded in these data may not be extracted using these basic analytic methods. The abundance and complexity of this information requires a new approach to data management and analysis methods to assist the physician/clinician to make diagnosis and treatment decisions with greater accuracy and efficiency. For example, mental and neurological illnesses such as mood disorders, depression and schizophrenia are common and debilitating conditions for which current treatment algorithms lack precision. The process for assessing an effective therapy (specifically, medication therapy) is poorly defined at best. In short, the basic procedure in the treatment of such complex illnesses is to prescribe various therapies on a trial and error basis until one is found that is effective.

Furthermore patients must often wait lengthy periods of time before seeing the clinical experts who possess the skill to effectively treat these conditions, particularly in rural areas where such specialists are few in number. As a result family physicians often initiate treatment themselves without the benefit of the extensive experience and knowledge possessed by psychiatrists and neurologists. Even among the clinical experts it is acknowledged that patients meeting the diagnostic criteria for most psychiatric and neurological conditions are not uniformly responsive to the same treatment. Some patients respond well to a given treatment while others, with very similar clinical features, do not.

Treatment failure may be a function of extraneous factors such as treatment adequacy (in terms of medication dose and duration of treatment), poor absorption of oral medication, unusual medication metabolism or inadequate patient adherence to prescribed treatment. However, individual patients often fail to respond to a particular treatment whose efficacy has been demonstrated in large clinical trials. This suggests that biological subtypes may exist within a given syndrome or diagnostic category. Patients afflicted with a particular biological subtype of a condition or diagnosis may respond preferentially to only some of the many medication treatments available to treat that condition or diagnosis. Often, even expert clinicians cannot readily distinguish these illness subtypes using current methodology. This suggests that current diagnostic systems and testing procedures may not be efficiently exploring the information to detect subtle but highly relevant differences between patients presenting with similar complaints.

1.3.1 Treatment Planning for Major Depressive Disorder

Major depressive disorder (MDD) or *major depression* is a serious mental disorder and is now the third largest cause of workplace disability. By the year 2020, depression is expected to account for about 15% of total global disease burden, second only to ischemic heart disease [30,31]. In industrialized countries mental illnesses may account for about 16% of total health care costs [32] and for about 30% of disability claims [33].

Despite the prevalence of MDD, objective procedures for selecting optimal treatments are lacking. The choice of antidepressant therapy is currently based on personal preference, weighted by clinical factors such as family history, symptom clustering and previous medication history. An effective algorithm for selecting the optimal antidepressant treatment on the basis of symptomatic presentation and other clinical data has proven to be an elusive objective [34,35], probably because the same collection of depressive symptoms may be produced by several different neurobiological pathologies as discussed previously. Typically, 60 to 70% of subjects do not remit after the first antidepressant medication trial [36]. Although 67% of those treated for MDD will eventually reach remission, up to 4 different antidepressant treatment trials may be required, each taking 6 weeks or longer [37, 38]. The personal and economic cost of delayed or ineffective therapy is substantial [38]. Clearly, choosing an effective treatment during the initial trial would be of immense clinical and economic value.

In this research, we tackled this problem by proposing a machine learning procedure which uses the pre-treatment clinical information to predict whether or not a particular therapy for MDD will be effective. In our clinical experiments, we considered two therapies for MDD:

1. Medications, in the form of selective serotonin reuptake inhibitors (SSRIs)

2. Repetitive transcranial magnetic stimulation (rTMS), (The details of this form of therapy are discussed later in Chapter 4).

The clinical data used in the experiments are electroencephalography (EEG) data collected prior to administering the SSRI or rTMS treatment to the patient; however, the approach is general and other clinical and laboratory data can be used. These experiments are discussed in Chapter 4.

1.3.2 Treatment Planning for Schizophrenia

Schizophrenia is a chronic, disabling brain disorder that occurs in about 0.5% of the population. It is a psychiatric diagnosis that describes a mental illness characterized by impairments in the perception or expression of reality, most commonly manifesting as auditory hallucinations, paranoid or bizarre delusions or disorganized speech and thinking in the context of significant social or occupational dysfunction.

For schizophrenia and schizoaffective disorder, the principle treatment is pharmacological, with drugs from the antipsychotic class, though mood stabilizers and antidepressant medications can also be used when mood symptoms are prominent.

The atypical antipsychotic medication *clozapine* is particularly effective in patients with schizophrenia, schizoaffective disorder and bipolar disorder, even after failed trials of other antipsychotic medications. However, clozapine is expensive and may produce life threatening side-effects such as bone marrow suppression in some who take it [39]. Also, a considerable number of patients treated with clozapine are still nonresponsive or only partially responsive. Though it would be highly advantageous to determine, in advance, whether a given patient would benefit from this dangerous medication there is currently no accepted method of determining potential response to clozapine short of an actual clinical trial. In this thesis, the proposed machine learning methodology shows significant promise in predicting the response of schizophrenic patients to this drug using only the EEG data collected from the patient before the onset of treatment. The results are discussed in Chapter 4.

1.3.3 What is the EEG?

Electroencephalography (EEG), or quantitative EEG (QEEG) is the measurement of electrical activity generated by the brain and recorded from electrodes placed on the scalp. See [40–42] and references therein for a more extensive review. "*EEG signals recorded on the scalp surface arise from large dendritic currents generated by the quasi-synchronous firing of a large number of neurons. At a finer spatial scale, these same currents are also responsible for local field potentials recorded extracellularly in-vivo in both humans and animals. The local field potential is generated by extracellular currents that pass through the extracellular space in a closed loop. These currents induce voltage changes (in the micro-Volts range) that are smaller than action potentials but that last longer and extend over a larger area of neural tissue. The local field potential reflects the linear sum of overlapping sources (current flows from the intracellular to the extracellular space) and sinks (current flows from the extracellular to the intracellular space). Scalp EEG arises from the passive conduction of currents produced by the summation of local field potentials over large neuronal aggregates. The columnar structure of the neocortex facilitates the summation of electrical activity distributed among multiple neuronal groups. EEG activity recorded on a scalp electrode corresponds to the sum of* *activity from regions near the electrode, but large signals originating from more distal cortical sites can make a significant contribution to the activity observed at a given point on the scalp*" [40]. Most popular uses of EEG signals are in epilepsy, seizure, sleep-related disorders, and cortical mapping.

The EEG is a commonly used method for collecting electrical activity of the brain that can help in diagnosis and selection of treatment in psychiatric disorders. See [42–44] for excellent reviews.

1.4 Medical Diagnosis

Diagnosis is defined as *the recognition of a disease or condition by its outward signs and symptoms*[3]. Currently, most clinicians make a diagnosis of psychiatric illness based upon a standard set of diagnostic criteria such as the Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association (DSM) [45] or the International Statistical Classification of Diseases and Related Health Problems (ICD) of World Health Organization [46]. The symptoms and signs of a neuro-psychiatric disease or condition are ordinarily reviewed and the critical information is discovered as follows: the clinician hears the presenting complaint, elicits subjective symptoms and, in some cases, conducts a physical examination of the patient. Based upon the information available at the time, a range of diagnostic possibilities is considered. The most likely diagnosis is designated the "preferred diagnosis". The other diagnostic possibilities are then listed in decreasing order of probability to form a "differential diagnosis". The preferred and differential diagnoses then suggest further analysis, including using laboratory and other clinical tests that will help to rule in or rule out the various entities in the diagnostic list. The idea in this thesis is to make this diagnosis procedure 'automatic' through use of a cognitive system that processes the data and extracts the critical information from clinical and laboratory measurements using machine learning methodologies.

The first step in obtaining an efficient treatment for a mental illness or disorder is obtaining a correct diagnosis. This can be a more difficult task than it might seem. Though the diagnostic criteria of different conditions are designed to differentiate patients with this condition from those with other conditions requiring other forms of treatment, often specific symptoms can appear in more than one diagnostic category and diagnostic criteria can overlap to the point where confident differentiation is impossible. Furthermore, current

[3] Diagnosis is also alternatively defined as *the analysis of the underlying physiological/biochemical cause(s) of a disease or condition.*

diagnostic systems are imperfect; i.e., all patients meeting diagnostic criteria for an illness such as major depressive disorder (MDD) do not all respond to the same treatment; e.g., antidepressant medication. This observation provides compelling clinical evidence for very substantial biological heterogeneity within a single diagnostic category.

In our clinical experiments to be discussed in Chapter 5, EEG data is used to suggest a diagnosis for psychiatric and neurological disorders and illnesses including MDD, schizophrenia, and bipolar affective disorder (BAD) or bipolar depression (BD).

Although the physician's estimated diagnosis (when available) can be used as prior knowledge, findings suggestive of an alternate diagnosis to the one preferred by the physician can be identified and this information conveyed to the physician. Also, while a medical diagnostic system proposed in this thesis may be useful to the family practitioner as well as the expert specialist physician, it will be of particular utility in circumstances where expert specialists or family physicians may not be readily available, and care must be administered by other clinically trained personnel such as nurse practitioners or other health-care providers. Therefore as part of this research, a patent application [47] is prepared and filed in which a cognitive medical expert (CME) system is proposed to solve this problem.

The design of the cognitive medical expert (CME) was a main outcome of this research. The proposed system and methodology is capable of predicting the treatment response for each individual patient at the outset of a therapy (i.e., using pre-treatment information) thus improving therapeutic efficiency and reducing personal and economic costs. Our experiments however, are focused on treatment planning and diagnosis of mood disorders and psychiatric illnesses.

1.5 Contributions

The contributions in this thesis are as follows:

- An automated method and procedure for gas/oil pipeline inspection using MFL data is developed.

- An automated method and procedure for treatment efficacy prediction or treatment planning based on the pre-treatment clinical data is proposed. The experimental results are obtained using pre-treatment clinical data for two psychiatric disorders: major depressive disorder and schizophrenia; however, the proposed methodology is more general and can be used for other illnesses.

- An automated method and procedure for medical diagnosis is proposed. This medical diagnosis can be used as a second opinion reported to the physician. Statistical methods using mixture of factor analyzers showed a promising performance in this application.
- A regularized feature selection based on Kullback-Leibler (KL) distance, is proposed.

1.6 Publications

The main material from this thesis has already been published, submitted, or in preparation to be submitted to journals and conferences, as reflected in papers [48–56]. The solution to the neuroscience application (treatment-response prediction as well as medical diagnosis) is also published as a PCT international patent application [47] as a medical expert system. The PCT is filed after filing a US provisional patent application [57]. The corresponding patent is to be filed in the national phases in US, Europe and Canada.

1.6.1 Journal Papers

1). The paper titled: "Machine learning techniques for the analysis of magnetic flux leakage images in pipeline inspection," published in *IEEE Transactions on Magnetics*, vol. 45, no. 8, pp. 3073–3084, Aug. 2009:
This paper describes automatic gas/oil pipeline inspection using real MFL data provided by the Intratech Inline Inspection company, (Mississauga, ON). Detection of metal defects as well as estimation of crack depth are investigated using machine learning procedure.

2). The paper titled: "A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy," *Clinical Neurophysiology*, 2010, DOI:10.1016/j.clinph.2010.05.009:
In this paper, the treatment-response prediction for the antipsychotic clozapine for treating schizophrenia is studied using machine learning methods.

3) The paper titled: "A Machine Learning Approach for Distinguishing Age of Infants Using Audio Evoked Potentials," which is submitted to: *IEEE Transactions on Information Technology in Biomedicine*, 2010:
This paper proposes using the auditory evoked potential data in response to a 4-note melody, for classifying the infants based on their age.

1.6.2 Conference Papers

1). The paper titled "Using pre-treatment EEG data to predict response to SSRI treatment for MDD", accepted at the *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2010.
In this paper, the prediction of response to SSRI antidepressant for major depression is studied. A regularized feature selection based on Kullback-Leibler (KL) distance is also described in this paper.

2). The paper titled "Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model", accepted at the *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2010.
In this paper, the medical diagnosis using EEG data and employing a maximum likelihood decision approach is studied. The *mixture of factor analysis* (MFA) model is used to build a probabilistic generative data model. The proposed system can perform diagnosis among four classes: major depression, schizophrenia, bipolar depression and normals.

1.6.3 Papers in Preparation

1). The paper titled: "A Machine Learning Approach Using EEG Data to Predict Response to SSRI Treatment for Major Depressive Disorder," which is to be submitted to: *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2010:
This paper describes a machine learning process (including automatic identification of discriminating features, classification and performance evaluation). The problem is to use the EEG data to find whether or not an antidepressant drug will be beneficial to a patient suffering major depressive disorder(MDD). Data clustering performance is also studied.

2). The paper titled: "A statistical decision model for the diagnosis of psychiatric disorders," which is to be submitted to: *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2010:
This paper describes solving the medical diagnosis problem using EEG data and applying a statistical decision model implemented using mixture of factor analyzers. The experiments described recognizing MDD from schizophrenia and from bipolar disorder. A 'multi-binary' feature selection and multi-class classification model is presented to improve performance. Data clustering performance is also studied.

3). The paper titled: "Analyzing pre-treatment EEG to predict the response to repetitive transcranial magnetic stimulation therapy in patients with

treatment-resistant MDD," which is in preparation and to be submitted to: *Journal of Psychiatric Research*, 2010:
This paper describes whether or not the EEG data has discriminating value in predicting the response to rTMS therapy for depression.

1.6.4 Patent Applications

1). The patent application was filed with the title: "Expert system for determining patient treatment response," as an International Patent Application in WIPO, PCT/CA2009/000195, Feb. 2009:
This PCT describes the system and methodology for a complete medical data analysis system which provides medical diagnosis as well as treatment planning and treatment response prediction. The cognitive medical expert system is like a digital clinician/physician which explores the data available from variety of sources.

2). Before the above PCT, a US provisional patent application was filed with the title: "Digital expert and neuro-psycho-biological signal processing as a predictor of response to treatment," as US Provincial Patent Application 61/064177, Feb. 2008:
This is a preliminary, short version of PCT described above.

Chapter 2

Methods

In this chapter, the signal processing and machine learning methods employed in this research are explained. These methods are commonly used in both pipeline inspection and neuroscience applications which both use similar mathematical methods through which we predict or estimate a target variable corresponding to a measured data sample, based on the information in the training set.

The outline of this chapter is as follows. First, the problem and notation will be defined in Section 2.1. The machine learning process starts with feature extraction. Features are a set of quantitative measurements taken from the object under test that allow us to discriminate which class the object belongs to. In Section 2.2, the supervised dimensionality reduction methods will be described. These methods select relevant features out of the all candidate features extracted from the measured data in the first step. These features are then used to build classification or regression models which will be described in Section 2.3. Finally in Section 2.4, the methods to measure the prediction performance and to find the design parameters of the classification/regression models will be described.

2.1 Problem Definition

The data analysis procedure and the corresponding variables and notations are described here. First, input data (*a.k.a* measured attributes), denoted by $\boldsymbol{E}_i$ and the corresponding target or output variable y_i, $i = 1, \ldots, M_t$ where M_t is the number of training samples are collected.

In the pipeline inspection problem, for example, the magnetic flux leakage (MFL) image segment is the input data, measured by the inline inspection tool traveling through pipeline, and the corresponding class or defect type (e.g. metal defect, weld, benign noise, etc.) is the target variable or output data.

In the neuroscience application, pre-treatment resting or spontaneous EEG signals after being divided into epochs (or segments), are collected from M available subjects who suffer from a psychiatric disorder. Here, M_t is the number of training epochs. Note that the parameter M_t is not the number of subjects in this case since we have typically 12 or more EEG epochs from each subject (to be discussed later in Chapter 4). In the treatment response prediction problem, the corresponding response outcome y_i of the patient to the treatment, after a suitable period, is the target variable. The possible values for the y_i are either "R" (responder), or "NR" (non–responder). In the medical diagnosis problem, y_i denotes the diagnostic class (MDD, schizophrenia, BAD, normal, etc.) and the input data is the resting EEG data.

The set of input data and the corresponding target variables is referred to as a *training set*, denoted by $\mathcal{D}$ as follows

$$\mathcal{D} = \Big\{(\boldsymbol{E}_i, y_i), \quad i = 1, \dots, M_t\Big\}. \tag{2.1}$$

These measured input data for each sample are pre-processed to extract a large number N_c of candidate features $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{N_c}$ that might be relevant for prediction of the target variable. By feature extraction process, the measured data is transformed into a high-dimensional space. The set of possibly effective candidate features depend greatly on the underlying problem, but typically they include an assortment of various attributes characterizing the statistical (e.g. first and second-order statistics on the observed data, histograms, coefficients of Fourier transform, etc.), geometric, temporal and dynamic model properties of the measured data. Examples of typical features used for the prediction problem will be discussed in detail later for each application. These set of candidate features are then reduced to a set of most relevant features $\mathbf{x}_i \in \mathbb{R}^{N_r}$, where $N_r \ll N_c$, to extract those features which are most indicative of the target variable. These reduced-dimensionality features are then fed into a classifier or regressor which outputs the corresponding value y_i.

Most of this thesis is about classification, which is a discrete process, in which the target variable y_i corresponds to class of the input data $\mathbf{x}_i$. The number of patterns or classes are denoted by N_p. The number of training samples corresponding to class j is shown with M_t^j, and we have $M_t = \sum_{j=1}^{N_p} M_t^j$ and $j = 1, \dots, N_p$.

In the sequel, the following notations are used. The matrices $\boldsymbol{X} \in \mathbb{R}^{M_t \times N_r}$ and $\tilde{\boldsymbol{X}} \in \mathbb{R}^{M_t \times N_c}$ are defined whose ith rows are $\mathbf{x}_i^T$ and $\tilde{\boldsymbol{x}}_i^T$, respectively, where superscript ^{T} denotes the transpose operation. The column vectors $\boldsymbol{\chi}_j$, $j =$

$1, \ldots, N_r$ and $\tilde{\chi}_\ell$, $\ell = 1, \ldots, N_c$ denote columns of $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ respectively. These vectors contain all values of the jth (ℓth) feature over the entire training set.

2.2 Supervised Dimensionality Reduction and Feature Selection

The extraction and selection of features is a critical issue to obtain optimum performance in most machine learning application. We are considering supervised[1] dimensionality reduction (or supervised feature selection) in this thesis. Given a training set $\mathcal{D}$, the goal in feature selection is to reduce the candidate set containing N_c features into a *reduced* subset containing only N_r most relevant features which are most indicative of the target variable. It is desirable to reduce the redundancy in the feature set $\mathbf{x}_i$ to the maximum extent possible, yet choose N_r large enough so that the set remains highly predictive of the y_i.

A simple approach for feature selection is ranking the features based on their *correlation* with the target variable (or the label) y. See [58] and references therein. In this method, the square of 'cross-correlation coefficient' (which is a positive number between 0 and 1) is used as the feature-ranking measure.

There are other alternative methods called 'feature-subset selection', in which instead of ranking the features one-by-one, subsets of features are ranked based on their predictive power. See [58] and references therein. In other words, the selected subset of features is the subset which is most useful for the estimator/predictor method. This criterion combines feature selection with the classification/prediction method and the goal is to generate the best final result with the greatest possible efficiency. This includes using *wrapper* and variable subset selection methods, nested subset selection, direct objective optimization, sequential feature selection, or similar methods or combinations of methods that achieve this goal. See [4,58–65] for a review of feature selection methods.

An optimal method for feature reduction is to find the subset of N_r discriminating features from the available set of N_c features which results in the best

[1] In contrast, in unsupervised dimensionality reduction methods, the target variable is not used in the process. An example of unsupervised feature selection is the principal component analysis method where the major principal vectors are used as the low-dimensional representation of the input data.

overall classification/prediction performance over the training data. Unfortunately, this method has combinatorial complexity and hence is intractable. We therefore must resort to sub–optimal methods.

2.2.1 Feature Selection Using Maximum Mutual Information

A prominent approach to feature reduction involves the use of mutual information. The mutual information of two discrete random variables $\mathcal{U}$ and $\mathcal{V}$ whose samples are u_i and v_i respectively, is defined as follows

$$\mathfrak{M}(\mathcal{U};\mathcal{V}) = \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} p(u,v)\,\log\frac{p(u,v)}{p(u)\,p(v)} \tag{2.2}$$

where $p(u,v)$ is the joint probability distribution function of $\mathcal{U}$ and $\mathcal{V}$, and $p(u)$ and $p(v)$ are the marginal probability distribution functions of $\mathcal{U}$ and $\mathcal{V}$, respectively. When at least one of the variables is continuous, mutual information is conveniently evaluated using e.g., Parzen windows. For details, see [66].

Intuitively, mutual information measures the information that $\mathcal{U}$ and $\mathcal{V}$ share: it measures how much knowing one of these variables reduces our uncertainty about the other. Mutual information quantifies the distance between the joint distribution of $\mathcal{U}$ and $\mathcal{V}$ and what the joint distribution would be if $\mathcal{U}$ and $\mathcal{V}$ were independent. It is a measure of dependence in the following sense: $\mathfrak{M}(\mathcal{U};\mathcal{V}) = 0$ iff $\mathcal{U}$ and $\mathcal{V}$ are independent random variables. Moreover, mutual information is nonnegative (i.e. $\mathfrak{M}(\mathcal{U};\mathcal{V}) \geq 0$) and symmetric (i.e. $\mathfrak{M}(\mathcal{U};\mathcal{V}) = \mathfrak{M}(\mathcal{V};\mathcal{U})$).

For feature selection, one can find the N_r indices (out of N_c), whose corresponding features yield the maximum mutual information with the target variable y. A better method which considers the relation between features is to find a subset of discriminating features of size N_r so that the mutual information between the joint feature variables in a selected subset and the response variable y is maximum. Unfortunately, this is a sub-optimal method considering the fact that the overall classification performance is not taken into account when selecting features, yet it has combinatorial complexity and hence is intractable.

2.2.2 Feature Selection Using Maximum Kullback-Leibler Distance

A further information theoretic criterion of value in this application is the Kullback–Leibler (KL) divergence [67]. This is a means of measuring "distance" or separation between two probability density functions (pdf's). The KL divergence between two random variables with probability distributions $p(u), u \in \mathcal{U}$ and $p(v), v \in \mathcal{V}$ is defined as follows, assuming that the random variables have discrete values indexed by i:

$$\mathcal{A}_{\text{KL}}(\mathcal{U}\|\mathcal{V}) = \sum_i p(u_i) \log \frac{p(u_i)}{p(v_i)}. \tag{2.3}$$

Mutual information can also be written in the form of a Kullback-Leibler divergence, as follows

$$\mathfrak{M}(\mathcal{U};\mathcal{V}) = \mathcal{A}_{\text{KL}}\Big(p(u,v)\|p(u)p(v)\Big). \tag{2.4}$$

KL divergence is not a symmetric measure; i.e., $\mathcal{A}_{\text{KL}}(\mathcal{U}\|\mathcal{V}) \neq \mathcal{A}_{\text{KL}}(\mathcal{V}\|\mathcal{U})$. For this reason we consider KL distance, which is a symmetric quantity, and is defined as

$$D_{\text{KL}}(\mathcal{U}\|\mathcal{V}) = 0.5\,\mathcal{A}_{\text{KL}}(\mathcal{U}\|\mathcal{V}) + 0.5\,\mathcal{A}_{\text{KL}}(\mathcal{V}\|\mathcal{U}). \tag{2.5}$$

The KL distance is strictly a function of the *distributions* of the variables; however, in this thesis, for notational convenience, the KL distance is defined in terms of the variables themselves.

2.2.3 Regularized Feature Selection

One sub-optimal (greedy) method as discussed in previous subsections is to evaluate each $\mathfrak{M}(y, \tilde{\chi}_\ell), \ell = 1, \ldots N_c$ for each candidate feature individually and choose the reduced features as those corresponding to the largest N_r such values. Another alternative is to find features with maximum KL distance between R and NR groups as described in previous subsection. The difficulty with these approaches is that the selected features are likely to have rich redundancy, with the result that the number of features could be reduced without degrading performance. To address this problem, [68] proposes a greedy algorithm, which at the jth step, $j = 1, \ldots, N_r$, selects that feature which corresponds to an optimal weighted combination of *(1)* maximum mutual information with the response vector y, and *(2)* minimum mutual information with the set of features already chosen in previous steps.

Here, a novel sub–optimal approach will be proposed which is a modification of the method of [59, 68]. This approach to feature selection considers the *pdf* of the ℓth candidate feature given that the subject is a responder, and the *pdf* of the ℓth feature given that the subject is a non-responder. "Good" features are those for which the KL distance between these distributions is large, yet the redundancy between selected features is minimum. To form these distributions, the column vector $\tilde{\chi}_\ell \in \mathbb{R}^{M_t}$, $\ell = 1, \ldots, N_c$ is divided into two subsets $U_\ell^{(\mathrm{R})}$ and $U_\ell^{(\mathrm{N})}$ (using the known response y_i corresponding to each element). These subsets contain the values of respective feature over the responder and non-responder groups, respectively. Approximations to the *pdf*'s of these subsets can be evaluated using histograms or Parzen windows. The idea of the proposed method is to choose the feature at the jth step, $j = 1, \ldots, N_r$ so that the selected feature is a combination of maximum relevance (i.e., the KL distance between $U_\ell^{(\mathrm{R})}$ and $U_\ell^{(\mathrm{N})}$ is maximum), and minimum redundancy (i.e., the combined KL distances between the distributions of the proposed feature at the current step and the corresponding distributions of features already chosen at previous steps is also maximum).

More precisely, the first column of $\boldsymbol{X}$ is the vector whose corresponding feature has maximum KL distance between responders and non-responders. Then, at the jth step, we already have $\boldsymbol{X}(j-1) \in \mathbb{R}^{M_t \times (j-1)}$, the matrix corresponding to the previously selected most relevant features. Let us define sets $\mathcal{L} = \{\ell \mid \ell = 1, \ldots, N_c\}$, $\mathcal{J}_1 = \{n_q \mid q = 1, \ldots, j-1\}$ (the set of indexes already chosen in previous steps) and $\bar{\mathcal{J}} = \{\mathcal{L} - \mathcal{J}_1\}$, the set of remaining indexes. The task is to select the jth feature vector whose index is $n_j \in \bar{\mathcal{J}}$ (i.e., the jth column $\chi(n_j)$ of $\boldsymbol{X}(j)$). In a manner similar to [68], this can be done by solving the following "regularized" optimization problem which implements a tradeoff between maximum relevance and minimum redundancy. The index n_{opt} corresponding to the optimal feature is then given by

$$\begin{aligned} n_{\mathrm{opt}} &= \arg\max_{n \in \bar{\mathcal{J}}} \Bigg\{ D_{\mathrm{KL}}\left(U_n^{(\mathrm{R})} \| U_n^{(\mathrm{N})}\right) \\ &+ \xi \frac{0.5}{j-1} \sum_{n_q \in \mathcal{J}_1} D_{\mathrm{KL}}\left(U_n^{(\mathrm{R})} \| U_{n_q}^{(\mathrm{R})}\right) + D_{\mathrm{KL}}\left(U_n^{(\mathrm{N})} \| U_{n_q}^{(\mathrm{N})}\right) \Bigg\}. \end{aligned} \qquad (2.6)$$

In the above, $\xi > 0$ is a regularization parameter[2] which controls the relative weighting between the relevance (first term) and the redundancy (second term) of each feature. The sets $U_n^{(\cdot)}$ (indexed by n) represent the R and NR subsets

[2] The default value for the regularization parameter is $\xi = 1$, which is used in all experiments in this thesis.

formed from the feature column $\chi(n)$, $n \in \bar{\mathcal{J}}$ under test, and subsets $U_{n_q}^{(\cdot)}$ (indexed by $n_q \in \mathcal{J}_1$) are the R and NR subsets corresponding to the features already chosen in previous iterations. The resulting column $\chi(n_{\text{opt}})$ from (2.6) is appended to $\boldsymbol{X}(j-1)$, j is incremented, and the process repeats until all N_r features are found. This procedure is denoted as the 'maxKLD' method and is the technique of choice for the experimental results in this thesis.

It is worthy of note that compared to Eq. (2.6), the method of Peng et al. [68] selects n_{opt} according to

$$n_{\text{opt}} = \arg\max_{n \in \bar{\mathcal{J}}} \left\{ \mathfrak{M}(\tilde{\chi}_n; \mathbf{y}) - \eta \frac{1}{j-1} \sum_{n_q \in \mathcal{J}_1} \mathfrak{M}(\tilde{\chi}_n; \chi_{n_q}) \right\} \tag{2.7}$$

where again, $\eta > 0$ is a regularization parameter, (with the default value of $\eta = 1$, as used in our experiments). This criterion[3] uses mutual information as a criterion of relevance and redundancy, rather than the KL distance, as in (2.6).

The feature selection method of [68], as described in Eq. (2.7), was the main method of choice in the pipeline inspection as well as neuroscience applications to be discussed in Chapters 3–5. The proposed feature selection method based on KL distance, as described in Eq. (2.6), as well as its simplified version to be described next, were used as alternative procedures.

2.2.4 Simplified Models for Supervised Feature Selection

An advantage of the proposed method for feature reduction based on KL distance is that it enables considerable simplification with little or no apparent performance degradation in our applications, by imposing the assumption that the two sets $U_\ell^{(\mathrm{R})}$ and $U_\ell^{(\mathrm{N})}$ are each univariate Gaussian distributed, with means μ_R, μ_N and variances σ_R^2, σ_N^2 respectively. With this assumption the KL distances in (2.6) can be evaluated using the following closed form expression (see [4], for example)

$$\begin{aligned} D_{\text{KL}}\left(U_l^{(\mathrm{R})} \| U_l^{(\mathrm{N})}\right) &= \frac{1}{4}\left(\frac{\sigma_R^2}{\sigma_N^2} + \frac{\sigma_N^2}{\sigma_R^2} - 2\right) \\ &+ \frac{1}{4}(\mu_R - \mu_N)^2 \left(\frac{1}{\sigma_R^2} + \frac{1}{\sigma_N^2}\right). \end{aligned} \tag{2.8}$$

[3] The method in [68] is also referred to as the "minimum-redundancy maximal relevance" (mRMR) feature selection method.

The means and variances required above are estimated from the respective subsets at each iteration. This procedure circumvents the evaluation of histograms or Parzen windows and the numerical evaluation of KL distances as required by (2.6).

Additionally, based on the above assumption, the relative contributions of the two terms above can be controlled by introducing a weighting factor, as follows:

$$\begin{aligned} D_{\text{KL}}\left(U_l^{(\text{R})} \| U_l^{(\text{N})}\right) &= (1-\delta)\frac{1}{4}\left(\frac{\sigma_R^2}{\sigma_N^2}+\frac{\sigma_N^2}{\sigma_R^2}-2\right) \\ &+ \delta\,\frac{1}{4}(\mu_R-\mu_N)^2\left(\frac{1}{\sigma_R^2}+\frac{1}{\sigma_N^2}\right) \end{aligned} \tag{2.9}$$

where $0 \leq \delta < 1$ is a weighting parameter. By using a δ value close to 1, more significance will be given to the difference in the mean values of the R and NR *pdf*s. Then, assuming the variances are not too different, the most individually discriminating features can be found.

Using multivariate Gaussian assumptions for the R and NR groups with corresponding N_r-dimensional pdfs $\mathcal{N}(\boldsymbol{\mu}_R; \boldsymbol{Q}_R)$ and $\mathcal{N}(\boldsymbol{\mu}_N; \boldsymbol{Q}_N)$, the KL distance for the h-th subset (of size N_r selected from total of N_c candidate features) is

$$\begin{aligned} D_{\text{KL}}\left(U_h^{(\text{R})} \| U_h^{(\text{N})}\right) &= \frac{1}{4}\left\{\text{trace}\left(\boldsymbol{Q_R}^{-1}\boldsymbol{Q_N}+\boldsymbol{Q_N}^{-1}\boldsymbol{Q_R}\right)-2N_r\right\} \\ &+ \frac{1}{4}\left\{(\boldsymbol{\mu}_R-\boldsymbol{\mu}_N)^T\left(\boldsymbol{Q_R}^{-1}+\boldsymbol{Q_N}^{-1}\right)(\boldsymbol{\mu}_R-\boldsymbol{\mu}_N)\right\} \end{aligned} \tag{2.10}$$

where trace$(\cdot)$ denotes the *trace* function of a square matrix, which is equal to sum of the diagonal elements of the matrix. The superscript $^{-1}$ denotes the inverse of the matrix. Note that here, $\boldsymbol{\mu}_R$ and $\boldsymbol{\mu}_N$ are vectors of size N_r, and $\boldsymbol{Q_R}$ and $\boldsymbol{Q_N}$ are both matrices of size $N_r \times N_r$. As a further simplification, assuming $\boldsymbol{Q_R} = \boldsymbol{Q_N} = I$, Eq. (2.10) reduces to $D_{\text{KL}}\left(U_h^{(\text{R})} \| U_h^{(\text{N})}\right) = \frac{1}{2}\|\boldsymbol{\mu}_R - \boldsymbol{\mu}_N\|^2$, where I denotes the identity matrix and $\|\cdot\|$ denotes the Euclidean norm. In a manner similar to Eq. (2.9), by weighting the contribution of each term in the above definition, one can use the following criterion instead for feature selection

$$\begin{aligned} D_{\text{KL}}\left(U_h^{(\text{R})} \| U_h^{(\text{N})}\right) &= (1-\delta)\frac{1}{4}\left\{\text{trace}\left(\boldsymbol{Q_R}^{-1}\boldsymbol{Q_N}+\boldsymbol{Q_N}^{-1}\boldsymbol{Q_R}\right)-2N_r\right\} \\ &+ \delta\frac{1}{4}\left\{(\boldsymbol{\mu}_R-\boldsymbol{\mu}_N)^T\left(\boldsymbol{Q_R}^{-1}+\boldsymbol{Q_N}^{-1}\right)(\boldsymbol{\mu}_R-\boldsymbol{\mu}_N)\right\} \end{aligned} \tag{2.11}$$

2.2.5 Multiclass Feature Selection

First, a simple multiclass feature selection procedure using the maximum KL distance criterion is discussed. Assuming an N_p-class classification problem, and considering values of feature $l, l = 1, \ldots, N_c$, in all M_t samples in the training set, let us denote the subset of l-th feature value for class j by $U_l^{(j)}$. Then, for feature selection, we find the indices of N_r features which yield the maximum average KL distance between feature values in all class pairs, that is defined as follows

$$\bar{D}_{\mathrm{KL}}(l) = \sum_{j=1}^{N_p-1} \sum_{h=j+1}^{N_p} D_{\mathrm{KL}} \left(U_l^{(j)} \| U_l^{(h)} \right) \tag{2.12}$$

where in the above, $\bar{D}_{\mathrm{KL}}(l)$ is the multiclass average KL distance for the feature indexed by l.

We also propose an alternative multiclass feature selection procedure. Assume that we have N_p classes. All possible binary feature selections are performed in which each binary feature selection is done independently for the corresponding binary classification problem. The total number of binary classifiers are $N_p(N_p - 1)/2$. The solution is to concatenate the relevant feature lists from the all possible binary feature selection processes into a single list, and then construct a single multi-class classifier using the overall collection of distinct relevant features as the input. This is referred to as the *feature index collection* or the 'feature combination' method.

2.3 Techniques for Classification and Regression

Even though classification and regression are often seen as two separate entities in the machine learning literature, in fact regression can be readily modified to perform classification tasks. This is accomplished in the regression case simply by quantization, i.e., fitting the regression output, denoted by $f(\mathbf{x})$, to a discrete–valued function, rather than a continuously–valued one as suggested previously. Because the algorithms used for regression are found to give better performance than classification algorithms in our applications, most of the classification results in this thesis are obtained using regression methods.

The specification of the classifier is equivalent to determining a function $f : \mathbb{R}^{N_r} \longmapsto \mathcal{C}$, where $\mathcal{C}$ is the the set of classes, e.g., $\mathcal{C} = 1, 2, \ldots, N_p$ for an N_p classification problem. In the regression case, f takes the form

$f : \mathbb{R}^{N_r} \longmapsto \mathbb{R}$. Thus we have $y = f(\mathbf{x}; \Theta)$, where the variable Θ represents the parameters of the model function f which are to be determined. There are many methods of determining the function f. In this section, a summary of various classification and regression methods will be given that were found to give good performance in the two applications under consideration (i.e., pipeline inspection and neuroscience).

First a brief discussion on the kernelization technique will be given. Three methods including support vector machine regression (SVR), regularized least squares (RLS) regression, and the partial least squares (PLS) regression method will be discussed. SVR and RLS methods [3–5, 69] are based on solving a regularization problem using the general principle of minimizing the expected discrepancy between the target value y and $f(\mathbf{x}; \Theta)$ for a given input vector $\mathbf{x}$, by their specific loss functional $\mathcal{L}(y, f(\mathbf{x}; \Theta))$, to be described later. The PLS regression method [70] is a modelling procedure based on dimensionality reduction and is related to the idea of principal component analysis (PCA). Finally, statistical classification methods based on a maximum likelihood decision approach or on a minimum Bayesian decision risk approach will be described.

Both SVR and PLSR regression methods were used as the main methods of choice in all pipeline inspection as well as treatment-response prediction applications to be discussed in Chapters 3, 4. The RLS method was also used as an alternative procedure. For the medical diagnosis application to be discussed in Chapter 5, the statistical classification procedure to be described in Subsection 2.3.6 was used as the classification method of choice.

2.3.1 Kernelization

Kernelization is an efficient method for improving performance by introducing nonlinearity into the feature space. Many techniques (such as the basic support vector machine below) result in linear decision boundaries and require only inner–product operations on the input data. It is possible to produce nonlinear decision boundaries in the feature space by transforming a feature vector $\mathbf{x} \in \mathbb{R}^{N_r}$ into another space $\mathcal{U}$ by defining a new vector $\boldsymbol{U} = \boldsymbol{\phi}(\mathbf{x})$ for some nonlinear function $\boldsymbol{\phi}(\cdot)$. However, in the case at hand, a more efficient method of introducing nonlinearity is to compute inner products of the form $\boldsymbol{\phi}(\boldsymbol{u})^T\boldsymbol{\phi}(\boldsymbol{v})$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are input feature vectors in $\mathbb{R}^{N_r}$ using kernel representations [71], as follows

$$\boldsymbol{\phi}(\boldsymbol{u})^T\boldsymbol{\phi}(\boldsymbol{v}) = \mathcal{K}(\boldsymbol{u}, \boldsymbol{v}). \tag{2.13}$$

This procedure allows us to compute the value of the inner product in $\mathcal{U}$ without explicitly carrying out the transformation $\phi(\cdot)$. Examples of commonly used kernels are the dth–order polynomial, the Gaussian and the sigmoid functions, given respectively as

$$\mathcal{K}_1(\boldsymbol{u},\boldsymbol{v}) = (\gamma\,\boldsymbol{u}^T\boldsymbol{v}+1)^d = \left(\gamma\sum_{i=1}^{N_r} u_i v_i + 1\right)^d \tag{2.14}$$

$$\mathcal{K}_2(\boldsymbol{u},\boldsymbol{v}) = \exp\left(\frac{-\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2\sigma^2}\right) \tag{2.15}$$

$$\mathcal{K}_3(\boldsymbol{u},\boldsymbol{v}) = \tanh\left(\kappa\,\boldsymbol{u}^T\boldsymbol{v}+\delta\right) \tag{2.16}$$

where $\gamma, d, \sigma, \kappa$ and δ are design parameters, and determine some sort of input scaling in the corresponding kernel functions. The linear kernel is defined as $\mathcal{K}_o(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{v}$. The relationship between the kernels above and their corresponding function ϕ is not always straightforward to determine. Nevertheless, for some kernels satisfying relatively benign conditions such as the Mercer condition [69, 71], a unique ϕ can be found. Some examples are given in [5, 71]. The dimension of the transformed space $\mathcal{U}$ can be much larger than N_r, which can result in better dimensionality reduction and better classification performance.

The (Euclidean) distance between examples $\mathbf{x}_i$ and $\mathbf{x}_j$ in the feature space of the kernel is, by definition

$$r_{ij} = \text{dist}(\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)) = \sqrt{(\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2)} \tag{2.17}$$

Distances can be computed directly from kernel values, as follows

$$r_{ij} = \sqrt{K_{ii} - 2K_{ij} + K_{jj}} \tag{2.18}$$

where $K_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ is the (i,j)-th element of the kernel matrix $\boldsymbol{K}$.

2.3.2 Regularized Least-Squares Regression

In the specific case of least-squares regression, the loss functional assumes the form $\mathcal{L}(y_i, f(\mathbf{x}_i;\Theta)) = \|y_i - f(\mathbf{x}_i;\Theta)\|_2^2$, where Θ denotes the set of design parameters of the regression function, to be discussed later. In any machine learning problem, there is always a tradeoff between model complexity on the one hand (which can lead to over–fitting) and model accuracy, which is essential for prediction, on the other. Regularization theory is a means of regulating this tradeoff. The goal is to minimize the following regularized risk functional

$$\mathcal{R}(f) = \frac{1}{M_t}\sum_{i=1}^{M_t} \mathcal{L}(y_i, f(\mathbf{x}_i;\Theta)) + \beta\,\|f\|_{\mathcal{K}}^2 \tag{2.19}$$

where the second term, $\|f\|_{\mathcal{K}}^2$, is a stabilizer to penalize model complexity of the function f, and the positive constant $\beta > 0$ is called the regularization parameter, which represents the relative importance of the model complexity with respect to the performance measure, (or model accuracy). The function $\|f\|_{\mathcal{K}}^2$ is a norm in a 'reproducing kernel Hilbert space' (RKHS) $\mathcal{H}$ defined by the positive definite function $\mathcal{K}$ [72]. Instead of $\|f\|_{\mathcal{K}}^2$, other stabilizer functions like $\|D(f)\|^2$ can be used in (2.19), where D is a linear differential operator [3]. This makes the function smooth, thereby satisfying the property of continuity.

Under some general conditions, the function $f(\cdot)$ in (2.19) in an RKHS is assumed to be of the form

$$f(\mathbf{x};\Theta) = \sum_{i=1}^{M_t} v_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) \tag{2.20}$$

where the coefficients v_i are to be determined. In this thesis, for the regularized least-squares (RLS) regression method, the Gaussian kernel is used, as defined in Eq. (2.15).

The solution $\mathbf{v}_*$ minimizing (2.19) where $f(\cdot)$ assumes the form of (2.20) and $\|f\|_{\mathcal{K}}^2 = \mathbf{v}^T \boldsymbol{K}\, \mathbf{v}$, satisfies

$$(\boldsymbol{K} + \beta\,\mathbf{I})\,\mathbf{v} = \mathbf{y} \tag{2.21}$$

where $\mathbf{I}$ denotes the $M_t \times M_t$ identity matrix, $\mathbf{y} = [y_1, y_2, \ldots, y_{M_t}]^T$, $\mathbf{v} = [v_1, v_2, \ldots, v_{M_t}]^T$, and

$$\boldsymbol{K} = \begin{bmatrix} \mathcal{K}(\mathbf{x}_1,\mathbf{x}_1) & \ldots & \mathcal{K}(\mathbf{x}_1,\mathbf{x}_{M_t}) \\ \vdots & \ddots & \vdots \\ \mathcal{K}(\mathbf{x}_{M_t},\mathbf{x}_1) & \ldots & \mathcal{K}(\mathbf{x}_{M_t},\mathbf{x}_{M_t}) \end{bmatrix}. \tag{2.22}$$

In summary, for the RLS method described above, the kernel matrix $\boldsymbol{K}$ is calculated from all the training data, and the coefficient vector $\mathbf{v}$ is then obtained from (2.21) for some specified values of design parameters $\Theta = \{\beta, \sigma\}$. The RLS method is alternatively referred to as the kernel ridge regression, e.g. in [4].

2.3.3 Support Vector Machine for Regression

The original support vector machine (SVM) was proposed by Vapnik [5] for use as a classifier. As an extension to SVM, the support vector machine regression (SVR) method [5, 69] uses the so-called ϵ-insensitive loss function defined as follows

$$\mathcal{L}_\epsilon(y, f(\mathbf{x};\Theta)) = \begin{cases} 0, & \text{if } |y - f(\mathbf{x};\Theta)| \le \epsilon \\ |y - f(\mathbf{x};\Theta)| - \epsilon, & \text{otherwise} \end{cases} \tag{2.23}$$

where $\epsilon > 0$.

First the *linear* regression case will be described where f is modelled as the linear function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w} \in \mathbb{R}^{N_r}$ is a weight vector and $b \in \mathbb{R}$ is a bias term.

The SVR problem with a linear kernel and given constants C and ϵ can be formulated as the following optimization problem

$$\min_{\{\mathbf{w},\boldsymbol{\xi},\boldsymbol{\xi}^*,b\}} \left\{ C\sum_{i=1}^{M_t}(\xi_i + \xi_i^*) + \frac{1}{2}\|\mathbf{w}\|^2 \right\} \tag{2.24}$$

$$\begin{aligned} &\text{subject to} \\ &y_i - \mathbf{w}^T\mathbf{x}_i - b \le \epsilon + \xi_i, \quad \xi_i \ge 0 \\ &\mathbf{w}^T\mathbf{x}_i + b - y_i \le \epsilon + \xi_i^*, \quad \xi_i^* \ge 0 \end{aligned} \tag{2.25}$$

Deviations larger than ϵ are measured by the variables ξ and ξ^*. The term $\|\mathbf{w}\|^2$ measures the flatness of f. Therefore, the constant $C > 0$ controls a trade-off between the flatness of f and the amount up to which deviations larger than ϵ are tolerated. If C is chosen too large, then it will fit the training data well, but may suffer from "over–training"; i.e., inability to generalize to new observations. On the other hand, if it is chosen too small, it will likely generalize, but may suffer from a lack of accuracy. It is straightforward to show that the problem defined by (2.24) and (2.25) is convex; therefore there is a unique minimum for the parameters defining f.

To extend f to the nonlinear case, the kernel function $\mathcal{K}(\mathbf{x}_i, \mathbf{x})$ is employed. Here, f assumes the form

$$f(\mathbf{x}) = \sum_{i=1}^{M_t}(\alpha_i - \alpha_i^*)\mathcal{K}(\mathbf{x}_i, \mathbf{x}) + b \tag{2.26}$$

where α and α^* are the dual Lagrange variables [69]. They are the solution to the following Lagrangian dual optimization problem

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & -\frac{1}{2}\sum_{i,j=1}^{M_t}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\,\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \\ & -\epsilon\sum_{i=1}^{M_t}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{M_t} y_i(\alpha_i - \alpha_i^*) \end{aligned} \tag{2.27}$$

$$\begin{aligned} &\text{subject to} \\ &\sum_{i=1}^{M_t}(\alpha_i - \alpha_i^*) = 0, \quad \alpha_i, \alpha_i^* \in [0, C] \end{aligned} \tag{2.28}$$

One method to solve the above optimization problem, is to use the Platt's 'sequential minimal optimization' (SMO) method which decomposes the main problem into subproblems of size 2, and then solves for the two analytically. So, in the SMO method, the whole problem is solved analytically without using numerical convex optimization routines, [69].

2.3.3.1 Multi-class SVM Classification

Some of the applications of interest in this thesis are multi-class problems, but the SVM classification is in principle based on binary classification. There are some methods for generalization of the SVM classifier to multiple class pattern recognition, (see [1,73,74], for example). However, efficient extension of the SVM to multi-class classification is still a challenge.

Assume that the number of classes is N_p. A method to generalize to the multi-class case is to train a separate binary SVM for each class, (each time considering all others as the second class). So, a total of N_p decision functions are constructed that separate each distinct class from the ensemble of the remaining classes. This is also referred to as the "one-versus-rest" classification technique. Then in the test phase, the winning class is the one with the largest margin.

Another alternative is the "one-versus-one" classification technique, in which SVM training is applied to all possible pairs of classes. Having N_p classes, a total of $M_B = \frac{N_p(N_p-1)}{2}$ decision functions are constructed. The "one-versus-one" can be implemented in two ways: The hierarchical tree approach uses a tree structure to recognize the test data $\mathbf{x}$ through M_B binary classification possibilities, as shown in Fig. 2.1. Based on this decision graph, in the test phase, $N_p - 1$ decision nodes will be evaluated to derive the final answer. This is also referred to as DAG-SVM, where DAG stands for 'directed acyclic graph'. The "one-versus-one" parallel approach is based on running all M_B binary classifiers in parallel, and then using a scoring (or decision averaging rule) to find the winning class. For example, let us denote the binary SVM discriminant function between classes Ω_i and Ω_j, by $f_{i,j}(\mathbf{x})$, and their corresponding class labels by $y_{i,j} \in \{+1, -1\}$, and $i, j = 1, \ldots, N_p$, also $i \neq j$. In this approach, the decision is made according to

$$\Omega^* = \arg \max_{i=1,\ldots,N_p} \left\{ \sum_{j=1, j\neq i}^{N_p} \operatorname{sign}\{y_{i,j}\, f_{i,j}(\mathbf{x})\} \right\} \tag{2.29}$$

where Ω^* denotes the final winning class label.

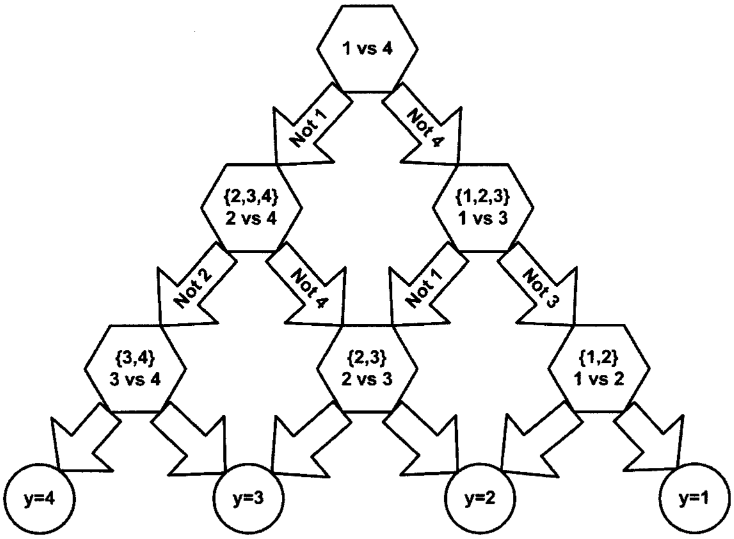

Figure 2.1: An example of DAG-SVM: One-versus-one multiclass classification based on the use of multiple binary classifiers to construct a four-class decision platform [1].

2.3.4 Partial Least Squares Regression

Partial least squares (PLS) is an extension of the principal component analysis (PCA) technique, and is an efficient method for reduced dimensional regression. The following is a summary of this method and its extensions borrowed from [70]. The feature vectors are arranged into a matrix $\boldsymbol{X} \in \mathbb{R}^{M_t \times N_r}$, where each row is a set of features (i.e., $\mathbf{x}^T \in \mathbb{R}^{N_r}$) extracted from one training sample. Also the y_i is similarly arranged into a vector $\mathbf{y} \in \mathbb{R}^{M_t}$.

The goal in PLS is to predict[4] $\boldsymbol{Y} \in \mathbb{R}^{M_t \times N_o}$ using components from $\boldsymbol{X}$. The dependent variables y are referred to as *response variables,* and the input variables $\mathbf{x}$ are the *predictors.* It is assumed that $N_r > N_o$. The prediction is done by extracting a set of orthogonal factors called *latent variables* (not directly observed or measured) which have the best predictive power of $\boldsymbol{Y}$. PLS regression (PLSR) is particularly useful when the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are highly correlated, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ exhibit strong cross–correlations. The objective criterion for constructing components in PLS is to sequentially maximize the covariance between the response variable and a linear combination of the predictors [70], [75].

[4] In the general PLS formulation, the response variables $\boldsymbol{Y}$ can have multiple columns and thus is represented as an $M_t \times N_o$ matrix rather than a vector, as it is in this application.

PLS regression searches for a set of components that perform a simultaneous decomposition of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with the constraint that these components explain as much of the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ as possible. It is this step which generalizes PCA. PLS is followed by a regression step where the decomposition of $\boldsymbol{X}$ is used to predict $\boldsymbol{Y}$.

More specifically, $\boldsymbol{X}$ is decomposed as

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{P}^T \qquad \text{such that} \quad \boldsymbol{S}^T\boldsymbol{S} = \boldsymbol{I}. \tag{2.30}$$

The matrix $\boldsymbol{S} \in \mathbb{R}^{M_t \times N_r}$ is referred to as the 'score matrix', while $\boldsymbol{P} \in \mathbb{R}^{N_r \times N_r}$ is the 'loading matrix' (in PLS regression the loadings are not orthogonal). Likewise, the prediction $\hat{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ is given as

$$\hat{\boldsymbol{Y}} = \boldsymbol{S}\boldsymbol{B}\boldsymbol{C}^T = \boldsymbol{X}\boldsymbol{B}_{\text{PLS}} \tag{2.31}$$

where $\boldsymbol{B}$ is a diagonal matrix with the regression weights as diagonal elements and $\boldsymbol{C}$ is the weight matrix of the dependent variables. Their elements will be defined later. The columns of $\boldsymbol{S}$ are the latent vectors. The matrix $\boldsymbol{B}_{\text{PLS}} = (\boldsymbol{P}^T)^+\boldsymbol{B}\boldsymbol{C}^T$ where $(\boldsymbol{P}^T)^+$ denotes the pseudo-inverse of $\boldsymbol{P}^T$.

In order to achieve dimensionality reduction, only $m \leq N_r$ major latent vectors are used to build the PLS regressor. In that case, $\boldsymbol{S}$ is an $M_t \times m$ matrix, $\boldsymbol{P}$ is a $N_r \times m$ matrix, $\boldsymbol{B}_{\text{PLS}}$ is an $N_r \times N_o$ matrix, $\boldsymbol{B}$ is an $m \times m$ matrix, and $\boldsymbol{C}$ is a $N_o \times m$ matrix.

Because PLS (or kernel PLS) can predict y, PLS can be used directly as a regressor or a classifier. In this case, $f = \boldsymbol{X}\boldsymbol{B}_{\text{PLS}}$, from (2.31). Thus, given a set of test feature vectors $\boldsymbol{X}_t$, the corresponding y-values can be determined using $\boldsymbol{B}_{\text{PLS}}$, which is calculated using the training set.

Any set of orthonormal vectors spanning the column space of $\boldsymbol{X}$ could be used to play the role of $\boldsymbol{S}$. In order to uniquely specify $\boldsymbol{S}$, additional constraints are required. For PLS regression this amounts to finding two sets of weights $\mathbf{w}$ and $\mathbf{c}$ in order to create (respectively) a linear combination of the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ such that their covariance is maximum. Specifically, at the ith iteration, the goal is to obtain the ith pair of vectors $\mathbf{s}_i = \boldsymbol{X}\mathbf{w}_i$ and $\mathbf{u}_i = \boldsymbol{Y}\mathbf{c}_i$, where $\mathbf{c}_i$ is the ith column of $\boldsymbol{C}$, with the constraints that $\mathbf{w}_i^T\mathbf{w}_i = 1$, $\boldsymbol{S}_i^T\boldsymbol{S}_i = \boldsymbol{I}$ and $b_i = \mathbf{s}_i^T\mathbf{u}_i$ is maximal. The matrix $\boldsymbol{S}_i \triangleq [\mathbf{s}_1, \ldots, \mathbf{s}_i]$. In each iteration, $\mathbf{w}_i$ is calculated by $\mathbf{w}_i = \boldsymbol{X}^T\mathbf{u}_{i-1}/(\mathbf{u}_{i-1}^T\mathbf{u}_{i-1})$. Then the vector $\mathbf{w}_i$ is normalized to unity norm. After calculating $\mathbf{s}_i$, the vector $\mathbf{c}_i = \boldsymbol{Y}^T\mathbf{s}_i/(\mathbf{s}_i^T\mathbf{s}_i)$, is normalized to unity norm, before calculating $\mathbf{u}_i$. After getting convergence in calculating $\mathbf{s}_i$ (by repeating the above calculation loop), the scalar $b_i = \mathbf{s}_i^T\mathbf{u}_i$ which constructs the ith diagonal element of $\boldsymbol{B}$, is calculated.

After the extraction of the score vectors $\mathbf{s}_i$ and $\mathbf{u}_i$, the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ at the ith iteration are deflated by subtracting the rank-one approximations based on $\mathbf{s}_i$ and $\mathbf{u}_i$. The matrix $\boldsymbol{U}_i \triangleq [\mathbf{u}_1, \ldots, \mathbf{u}_i]$ is not explicitly used in the PLS procedure. Other forms of deflation are discussed in [70]. The process repeats until $\boldsymbol{X}$ becomes a null matrix.

Alternatively, it can be shown that the ith weight vector $\mathbf{w}_i$ is the ith right singular vector of the matrix $\boldsymbol{X}^T\boldsymbol{Y}$. Similarly, the ith weight vector $\mathbf{c}_i$ is the ith left singular vector of $\boldsymbol{X}^T\boldsymbol{Y}$. The same argument shows that the ith vectors $\mathbf{s}_i$ and $\mathbf{u}_i$ are the ith eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T$ and $\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{X}^T$, respectively.

PLS can also be kernelized. Kernel-based PLS [70] is an extension of PLS and employs kernels to model nonlinear data relations. In the kernel PLS regression (also denoted by kernel PLSR, as well as KPLSR) method that is used in the experiments in this thesis, a linear PLS regression model in a nonlinear feature space $\mathcal{U}$ is considered. First an output kernel Gram matrix, $\boldsymbol{K}_o \in \mathbb{R}^{M_t \times M_t}$ is defined as

$$\boldsymbol{K}_o = \boldsymbol{Y}\boldsymbol{Y}^T. \tag{2.32}$$

Also an input kernel Gram matrix $\boldsymbol{K} \in \mathbb{R}^{M_t \times M_t}$ is defined, with elements $K_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ where $\mathcal{K}(\cdot, \cdot)$ is a selected kernel function. Using these definitions, the estimates of $\mathbf{s}_i$ and $\mathbf{u}_i$ in $\mathcal{U}$ can be obtained by reformulating the problem into a nonlinear kernel variant form as

$$\begin{aligned} \boldsymbol{K}\boldsymbol{K}_o\mathbf{s}_i &= \lambda_i\,\mathbf{s}_i && (2.33)\\ \mathbf{u}_i &= \boldsymbol{K}_o\mathbf{s}_i, \quad i = 1, \ldots, M_t. && (2.34) \end{aligned}$$

In a manner similar to the ordinary PLS method, a zero-mean nonlinear kernel PLS model is assumed. At each step, after the extraction of the new score vectors $\mathbf{s}_i$ and $\mathbf{u}_i$, the matrices $\boldsymbol{K}$ and $\boldsymbol{K}_o$ are deflated by subtracting their rank-one approximations based on the estimated $\mathbf{s}_i$ and $\mathbf{u}_i$. The process continues until the desired number m of latent variables is extracted. As in the linear PLS case, the effectiveness of the method results from the fact that the score variables $\{\mathbf{s}_i\}_{i=1}^m$ are good predictors of $\boldsymbol{Y}$. See [70,76,77] for further details.

2.3.5 Statistical Classification Methods

In the statistical machine learning approach, using the information in the training set, the goal is to assign each pattern (or feature vector) to one of N_p categories or classes $\{\Omega_1, \ldots, \Omega_{N_p}\}$ denoted by discrete status or target variable y, so that a statistical criterion of optimality (such as obtaining maximum probability of correct classification, minimum probability of error, etc.) is met. In the Bayesian classification approach, this is implemented based on the

Bayes decision criterion, to be described later. In this thesis, we alternatively use the the event Ω_j to denote the class with value $y = j$.

Assume that we estimated the probability models for all classes,

$$\left\{p(\mathbf{x}|\Omega_j, \hat{\Theta}^{(j)}),\ y = j = 1, \ldots, N_p\right\} \tag{2.35}$$

where $\hat{\Theta}^{(j)}$ denotes the estimated parameters for the probabilistic generative model of class $y = j$ corresponding the probabilistic event Ω_j. Based on the 'maximum a-posteriori' (MAP) classification rule[5], given a test data $\mathbf{x}$, the decision is

$$\begin{aligned} \hat{y} &= \arg\max_j \ p(\Omega_j|\mathbf{x}, \hat{\Theta}) \\ &= \arg\max_j \ p(\mathbf{x}|\Omega_j, \hat{\Theta}^{(j)})\, p(\Omega_j) \end{aligned} \tag{2.36}$$

where $p(\Omega_j) = p(y = j)$ is the a-priori probability of class $y = j$, and $p(\Omega_j|\mathbf{x}, \hat{\Theta})$ denotes the posterior pdf. The set of all estimated parameters is denoted by $\hat{\Theta} = \{\hat{\Theta}^{(j)}\}_{j=1}^{N_p}$. If we have some a-priori information (i.e., in the form of initial guess, or primary estimate) about the type/class of the test pattern, then the MAP decision rule is the statistically optimal decision criterion to be used. Otherwise, we can assume that classes are equiprobable, i.e., $p(\Omega_j) = \frac{1}{N_p}$ for all j, and therefore can use the maximum likelihood (ML) classification rule instead,

$$\hat{y} = \arg\max_j \ p(\mathbf{x}|\Omega_j, \hat{\Theta}^{(j)}) \tag{2.37}$$

In a more general and more statistically efficient Bayesian classification or identification approach, when relative cost values for various decision alternatives are given, the goal is to minimize the average decision cost (or average decision risk) $\bar{C}$, defined as follows

$$\bar{C} = \sum_{l=1}^{N_p} \sum_{j=1}^{N_p} C_{l,j}\ p(\text{decide}\, \Omega_l|\Omega_j)\, p(\Omega_j) \tag{2.38}$$

where $C_{l,j} \geq 0$ is the decision cost associated with deciding (or choosing, or reporting) class $y = l$ when $y = j$ is the true class (or actual class), and the corresponding probability is denoted by $p(\text{decide}\, \Omega_l|\Omega_j)\, p(\Omega_j)$.

[5] Alternatively this is referred to as the MAP detection or identification rule in the literature.

Rearranging, and after some simple mathematical operations, the decision function based on the above *Bayesian detection* approach is

$$f_l(\mathbf{x}) \triangleq \sum_{j=1}^{N_p} C_{l,j}\, p(\mathbf{x}|\Omega_j, \hat{\Theta}^{(j)})\, p(\Omega_j) \tag{2.39}$$

and the decision rule is

$$\hat{y} = \arg\min_{l} \; f_l(\mathbf{x}), \qquad l = 1, \dots, N_p \tag{2.40}$$

In other words, the statistical classification rule is: report the class of feature vector $\mathbf{x}$ as Ω_k when for $\forall l = 1, \dots, N_p$, $l \neq k$, we have $f_k(\mathbf{x}) < f_l(\mathbf{x})$. Also since $f_l(\mathbf{x}) \geq 0$, and $\log(\cdot)$ is a monotonic function, we can compare $\log f_l(\mathbf{x})$ for all indices l, instead. This is the test/operational phase of the statistical decision method. An efficient method for learning the probability model using the training data set will be described next.

There are also more specific methods in probabilistic data modeling including the naive-Bayesian method, and the Gaussian mixture model (GMM). See [4,7,78–81], for example.

2.3.6 Statistical Data Modeling by Mixture of Factor Analysis Models

Factor analysis (FA) is a method for modeling correlations in high-dimensional data, by correlations in a lower-dimensional, oriented subspace, [78,82,83]. In the following, we summarize the discussion from [82]. For the training data set of each class indexed by j, the model assumes that each N_r-dimensional data vector $\mathbf{x}_i^{(j)}$ was generated by first linearly transforming a $m < N_r$ dimensional vector of unobserved independent zero-mean unit-variance Gaussian sources (factors) $\mathbf{z}_i^{(j)} = [z_{i1}, \dots, z_{im}]$, translating by a fixed amount $\boldsymbol{\mu}^{(j)}$ in the data space, followed by adding N_r-dimensional zero-mean Gaussian noise, $\mathbf{n}_i$, with diagonal covariance matrix $\Psi^{(j)}$. This means that the data model is

$$\mathbf{x}_i^{(j)} = \mathbf{A}^{(j)}\, \mathbf{z}_i^{(j)} + \boldsymbol{\mu}^{(j)} + \mathbf{n}_i, \qquad i = 1, \dots, M_t^j \tag{2.41}$$

where $\mathbf{z}_i^{(j)} \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$, $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}; \Psi^{(j)})$. The matrix $\mathbf{A}^{(j)}$ with size of $N_r \times m$, is the linear transformation known as the 'factor loading matrix', and $\boldsymbol{\mu}^{(j)}$ is the mean vector of the analyzer. The symbol $\mathcal{N}(\cdot\,;\cdot)$ denotes the multidimensional Gaussian probability distribution in which the first argument is its average or mean vector and the second argument is its covariance matrix. The symbol $\mathbf{I}$ denotes the identity matrix. Statistically, we have

$$p(\mathbf{x}_i^{(j)}|\mathbf{z}_i^{(j)}, \mathbf{A}^{(j)}, \boldsymbol{\mu}^{(j)}, \Psi^{(j)}) = \mathcal{N}(\mathbf{x}_i^{(j)}|\mathbf{A}^{(j)}\, \mathbf{z}_i^{(j)} + \boldsymbol{\mu}^{(j)}; \Psi^{(j)}). \tag{2.42}$$

Note that because of the fact that $m < N_r$, in comparison with the unknown $N_r \times N_r$ covariance matrices associated with the traditional Gaussian mixture model (GMM) model, the factor analysis model can be considered as a more compact method with a smaller unknown parameter size [78].

Let us denote the data model for the j-th class by the corresponding parameter set $\Theta^{(j)} = \{\mathbf{A}^{(j)}, \boldsymbol{\mu}^{(j)}, \Psi^{(j)}\}$, where the superscript (j) refers to class j. Integrating out $\mathbf{z}_i^{(j)}$, it is simple to show that the marginal density of $\mathbf{x}_i^{(j)}$ will be Gaussian about the displacement $\boldsymbol{\mu}^{(j)}$,

$$\begin{aligned} p(\mathbf{x}_i^{(j)}|\Omega_j, \mathbf{A}^{(j)}, \boldsymbol{\mu}^{(j)}, \Psi^{(j)}) &= \int p(\mathbf{x}_i^{(j)}|\mathbf{z}_i^{(j)}, \mathbf{A}^{(j)}, \boldsymbol{\mu}^{(j)}, \Psi^{(j)})\, p(\mathbf{z}_i^{(j)})\, d\mathbf{z}_i^{(j)} \\ &= \mathcal{N}(\mathbf{x}_i^{(j)}|\boldsymbol{\mu}^{(j)}; \mathbf{A}^{(j)}\mathbf{A}^{(j)T} + \Psi^{(j)}), \end{aligned} \tag{2.43}$$

and assuming independence of the training samples, the joint probability of an *i.i.d.* data set $\boldsymbol{X}^{(j)}$ (i.e., the training data set for class j) is given by

$$p(\boldsymbol{X}^{(j)}|\Omega_j, \mathbf{A}^{(j)}, \boldsymbol{\mu}^{(j)}, \Psi^{(j)}) = \prod_{i=1}^{M_t^j} p(\mathbf{x}_i^{(j)}|\Omega_j, \mathbf{A}^{(j)}, \boldsymbol{\mu}^{(j)}, \Psi^{(j)}). \tag{2.44}$$

Given a training data set belonging to class j which has covariance matrix Σ^* and mean $\boldsymbol{\mu}^*$, factor analysis finds the $\mathbf{A}^{(j)}$, $\boldsymbol{\mu}^{(j)}$ and $\Psi^{(j)}$ for which the covariance of $\boldsymbol{X}^{(j)}$ optimally fits Σ^* in the maximum likelihood sense. The diagonal entries of the matrix $\Psi^{(j)}$ concentrate on fitting the axis-aligned (measurement) noise in the data, leaving the factor loadings in $\mathbf{A}^{(j)}$ to model the remaining (assumed-interesting) covariance structure.

A problem in FA is determining the dimensionality of the latent space, m. If too low a value of m is chosen, then the model has to discard some of the covariance in the data as noise, and if m is given too high a value this causes the model to fit spurious correlations in the data. An upper bound on proper values for m can be obtained by comparing the number of degrees of freedom in the covariance specification of the data set and the degrees of freedom that the FA parameterisation has in its parameters. The approximate bound is [84]

$$m_{\max} = N_r + \frac{1}{2}\left(1 - \sqrt{1 + 8N_r}\right) \tag{2.45}$$

For example, for $N_r = 30$, we have $m_{\max} = 22$.

In factor analysis, each factor dictates the amount of each linear transformation on the data set. However, with factor analysis we are restricted to linear transformations, and so any one analyzer can only explain well a small region of the manifold in which it is locally linear, even though the data manifold maybe globally non-linear.

One way to overcome this is to use mixture models to tile the data manifold. A mixture of factor analyzers (MFA) models the density for a data point $\mathbf{x}_i$ as a weighted average of factor analyzer densities, as follows

$$p(\mathbf{x}_i^{(j)}|\Omega_j, \Theta^{(j)}) = \sum_{k=1}^{K} \alpha_k^{(j)} p\left(\mathbf{x}_i^{(j)}|s_k^{(j)}, \mathbf{A}^{(j,k)}, \boldsymbol{\mu}^{(j,k)}, \Psi^{(j)}\right), \tag{2.46}$$

where K is the number of mixture components in the model, $\alpha_k^{(j)} = p(s_k^{(j)}|\boldsymbol{\alpha}^{(j)})$ is the mixing proportion[6], $s_k^{(j)} \in \{1, 2, \ldots, K\}$ represents a discrete random variable indicating the component from which $\mathbf{x}_i$ has been generated, $\boldsymbol{\alpha}^{(j)} = [\alpha_1^{(j)}, \ldots, \alpha_K^{(j)}]^T$ such that $\sum_{k=1}^{K} \alpha_k^{(j)} = 1$ and $\alpha_k^{(j)} > 0$. The factor loading matrix for analyzer $s_k^{(j)}$ is denoted by $\mathbf{A}^{(j,k)}$, and $\boldsymbol{\mu}^{(j,k)}$ are the corresponding analyzer means. The last term in the above probability is just the single analyzer density, given in Eq. (2.43). Based on the MFA model, the parameter set corresponding to the j-th class of data is

$$\Theta^{(j)} = \left\{\{\mathbf{A}^{(j,k)}, \boldsymbol{\mu}^{(j,k)}\}_{k=1}^{K}, \boldsymbol{\alpha}^{(j)}, \Psi^{(j)}\right\}. \tag{2.47}$$

In a MFA model, each Gaussian cluster has intrinsic dimensionality m, or m_s if the dimensions are allowed to vary across mixture components. Consequently, the mixture of factor analyzers simultaneously addresses the problems of clustering and local dimensionality reduction.

A 'maximum likelihood' (ML) procedure for fitting MFA model to the training data can be derived from the *expectation-maximization* (EM) algorithm. There are a few options in how to implement the EM algorithm. See [78, 82–86] for more details on implementation of the conventional EM algorithm, the 'variational EM' method, as well as the 'expectation conditional maximization' algorithm. In the experiments of this thesis, the learning method in [82] is used.

In learning the MFA model, the number of mixture components (K) and number of factors (m) is assumed the same in the *pdf* models of all classes. However, the best numbers for K and m (denoted as the *design parameters*) are found using a cross-validation procedure in which we perform a two-dimensional grid search and pick the best values (from a set of candidates) which result in best classification performance. However, if one uses the variational EM algorithm [83] instead, only a maximum value for the factor dimensionality m needs to be determined.

[6] The value of $p(s_k^{(j)}|\boldsymbol{\alpha}^{(j)})$ is considered as a prior probability distribution on the mixture components, defined by vector $\boldsymbol{\alpha}^{(j)}$.

2.3.6.1 Inference by the MFA model

After the MFA model is learned, then during the inference or operational phase, the likelihood that a test data vector $\mathbf{x}$ belongs to class j is given by

$$p(\mathbf{x}|\Omega_j, \hat{\Theta}^{(j)}) = \sum_{k=1}^{K} \hat{\alpha}_k^{(j)} \mathcal{N}\left(\mathbf{x}|\hat{\boldsymbol{\mu}}^{(j,k)}; \hat{\mathbf{A}}^{(j,k)} \, \hat{\mathbf{A}}^{(j,k)^T} + \hat{\Psi}^{(j)}\right) \tag{2.48}$$

Given the prior probability values $\{p(\Omega_j)\}_{j=1}^{N_p}$, the posterior probability of belonging to class j is

$$p(\Omega_j|\mathbf{x}, \hat{\Theta}) = \frac{p(\mathbf{x}|\Omega_j, \hat{\Theta}^{(j)}) \, p(\Omega_j)}{\sum_{l=1}^{N_p} p(\mathbf{x}|\Omega_l, \hat{\Theta}^{(l)}) \, p(\Omega_l)} \tag{2.49}$$

and as previously described, the class of $\mathbf{x}$ can then be estimated based on MAP or ML criteria.

One of the advantages of using the MFA method is that the posterior probability of a test data sample belonging to a particular class is one of the output variables of the MFA model. This can assist the user of the system in determining the certainty of the diagnosis, for example. These estimated class-conditional likelihood values can be used as an additional outcome of the diagnosis model that helps the user of the system. In using this kind of result, the user may want to put a detection threshold on the estimated likelihoods and then infer his/her own decision in a more controlled way.

2.3.7 Low-dimensional Representation

Apart from the supervised feature selection methods discussed in Section 2.2, another interesting approach for dimensionality reduction is feature transformation in which high-dimensional feature data is mapped into a new low-dimensional feature space. The resultant features may have a nonlinear relationship with the original features. A traditional method for such dimensionality reduction is principal component analysis (PCA) mapping. This can be used for clustering analysis as well. In doing so, we generate a two-dimensional representation (2D) of input data, and investigate how the data points cluster in this space. Kernelized PCA [71] is an efficient method for achieving this goal, and is used in Chapters 3–5 for such analysis.

Kernel PCA (also denoted as KPCA) introduces a nonlinear mapping into the feature subspace, as described in Sect. 2.3.1. Kernel PCA first maps the data into some feature space $\mathcal{U}$ via a (usually nonlinear) function ϕ, and then performs linear PCA on the mapped data. Since PCA is a linear algebra construct, it involves only inner product evaluations, and thus it is possible to perform kernel PCA without explicitly evaluating the transformation ϕ, [71].

First we make the assumption that we are dealing with centered data, i.e.,

$$\frac{1}{k}\sum_{i=1}^{k}\phi(\mathbf{x}_i)=0 \tag{2.50}$$

where k is the total number of data samples. Then the covariance matrix in $\mathcal{U}$ is

$$\underline{\tilde{\mathbf{C}}}=\frac{1}{k}\sum_{i=1}^{k}\phi(\mathbf{x}_i)\,\phi(\mathbf{x}_i)^T \tag{2.51}$$

Covariance matrix $\underline{\tilde{\mathbf{C}}}$ can be diagonalized by eigenvalue decomposition

$$\underline{\tilde{\mathbf{C}}}\,\tilde{\mathbf{v}}=\lambda\,\tilde{\mathbf{v}} \tag{2.52}$$

We have the fact that each eigenvector with $\lambda\neq 0$ can be expressed as

$$\tilde{\mathbf{v}}=\sum_{i=1}^{k}\alpha_i\,\phi(\mathbf{x}_i) \tag{2.53}$$

i.e., each eigenvector lies in the span of Φ-images of the data set. Substituting Eq. (2.51) and (2.53) in (2.52), and defining the $k\times k$ kernel matrix $\boldsymbol{K}$ with elements $K_{ij}=\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)=\phi(\mathbf{x}_i)^T\,\phi(\mathbf{x}_j)$, and by some manipulations, results in the following eigenvalue decomposition problem

$$\boldsymbol{K}\boldsymbol{\alpha}=k\lambda\,\boldsymbol{\alpha} \tag{2.54}$$

where $\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_k]^T$.

Normalizing the solution $\tilde{\mathbf{v}}_j$ in $\mathcal{U}$ (i.e., $\tilde{\mathbf{v}}_j^T\,\tilde{\mathbf{v}}_j=1$), translates into

$$\lambda_j(\boldsymbol{\alpha}_j^T\boldsymbol{\alpha}_j)=1 \tag{2.55}$$

To extract nonlinear principal components, we compute the projection of the Φ-image of a data point $\mathbf{x}_i$ onto the j-th eigenvector in the feature space by

$$\begin{aligned}\beta_{ij}\triangleq\tilde{\mathbf{v}}_j^T\,\phi(\mathbf{x}_i) &= \frac{1}{\sqrt{\lambda_j}}\sum_{h=1}^{k}\boldsymbol{\alpha}_j(h)\,\mathcal{K}(\mathbf{x}_h,\mathbf{x}_i)\\ &= \sqrt{\lambda_j}\,\alpha_j(i)\end{aligned} \tag{2.56}$$

where $\alpha_j(h)$ denotes the h-th element of vector $\boldsymbol{\alpha}_j$.

The low-dimensional representation of a data vector $\mathbf{x}_i$ is computed as follows: First, calculate the kernel matrix using all data set. Second, compute m largest eigenvalues (in decreasing order) of kernel matrix $\boldsymbol{K}$, and normalize them using (2.55). Third, compute the projection of $\phi(\mathbf{x}_i)$ onto corresponding eigenvectors. Finally

$$\mathbf{z}_i = [\beta_{i1}, \ldots, \beta_{im}]^T \tag{2.57}$$

where β_{ij} are defined by Eq. (2.56). The projection of $\mathbf{x}$ onto the first m principal components in $\mathcal{U}$ can be written as follows

$$\Pr\{\phi(\mathbf{x})\} = \sum_{j=1}^{m} \beta_j \, \tilde{\mathbf{v}}_j. \tag{2.58}$$

Note that if m is large enough to take into account all directions belonging to eigenvectors with non-zero eigenvalues, we will have $\Pr\{\phi(\mathbf{x})\} = \phi(\mathbf{x})$.

In general, we cannot use the assumption in Eq. (2.50). By relaxing this assumption, we need to use

$$\tilde{\phi}(\mathbf{x}_i) = \phi(\mathbf{x}_i) - \frac{1}{k}\sum_{h=1}^{k} \phi(\mathbf{x}_h) \tag{2.59}$$

instead of $\phi(\mathbf{x}_i)$. After calculating the kernel matrix $\boldsymbol{K}$ from the data set, the modified version $\tilde{K}_{ij} = \tilde{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j) = \tilde{\phi}(\mathbf{x}_i)^T \tilde{\phi}(\mathbf{x}_j)$ is used in Eq. (2.54) and (2.56):

$$\begin{aligned} \tilde{K}_{ij} &= \left(\phi(\mathbf{x}_i) - \frac{1}{k}\sum_{h=1}^{k}\phi(\mathbf{x}_h)\right)^T \left(\phi(\mathbf{x}_j) - \frac{1}{k}\sum_{l=1}^{k}\phi(\mathbf{x}_l)\right) \\ &= K_{ij} - \frac{1}{k}D_i - \frac{1}{k}D_j + \frac{1}{k^2}\sum_{h=1}^{k} D_h \end{aligned} \tag{2.60}$$

where $D_i = \sum_{j=1}^{k} K_{ij}$. Alternatively, one can use the following equation,

$$\tilde{\boldsymbol{K}} = \boldsymbol{K} - \mathbf{1}_k\boldsymbol{K} - \boldsymbol{K}\mathbf{1}_k + \mathbf{1}_k\boldsymbol{K}\mathbf{1}_k \tag{2.61}$$

where $(\mathbf{1}_k)_{i,j} \triangleq \frac{1}{k}$ for all i and j.

2.4 Performance Analysis

This section discusses issues relating to over-fitting and performance evaluation.

2.4.1 The Over-fitting Issue

A problem that may arise during the machine learning process is *over-fitting* or *over-training*[7]. In theory, if we use a complex and sufficiently large model for classification or regression which over-fits the training set with a relatively small size, then we could achieve high training performance. This means that if we input an already seen data sample from the same training subset to the model, there is a high probability that the classification or regression result will be correct. However, if we test this complex model with a new and unseen data sample, the fitting error will be large and the predictive performance will drop significantly. In other words, fitting and modeling may not generalize well to independent test data samples. This is because the relatively small training subset does not sufficiently represent all possible samples, and therefore random noise and fluctuations in the training data are unnecessarily taken into account.

One way to partly avoid this issue, is to employ a smooth data model for classification/regression which employs a relatively small number of unknown parameters. This is usually implemented by using a regularization or trade-off between model complexity and modeling error, as is done with the RLS and SVR models, as discussed in Sections 2.3.2 and 2.3.3. Using smoother and simpler models usually translates to better generalization capability, meaning that it is less sensitive to small random fluctuations in the input data. However, since the criterion of simplicity and smoothness is not well defined, we still need to measure the test performance of the model. An overly simple model on the other hand usually under-fits the data, a phenomenon which is as undesirable as over-fitting.

A popular and well-established method that avoids the over-fitting (or over-training) issue is to use independent training and test data samples, meaning that the test samples should not be used in the training process, and then to measure only the *test performance* as the predictive performance of the classification/regression model. What ultimately matters is the test performance. The *cross-validation* procedure to be discussed later in this section is a structured method which uses all the available training data samples in a way that this criterion is met, yet the final performance reflects the information in all available data samples. Using cross-validation is the method of choice for performance evaluation when the number of available data samples is small.

[7] This issue can be easily seen when the number of training cycles are excessively large when training a multi-layer perceptron (a popular artificial neural neural network model), see e.g., [3].

2.4.2 Quantification of Performance and Optimization of Design Parameters

All classification methods discussed in this chapter have some form of design parameters that need to be determined. In this subsection, we discuss two issues: First, an optimality criterion for finding the design parameters and second, how a performance evaluation index is determined. The problem addressed here is how we measure the performance of a classification model, using the estimated target values (i.e., the output of classification model) and the corresponding true (or actual) target values.

We denote the set of design parameters associated with the regression or classification procedures by Θ. This set includes various parameters associated with the regression and kernelization procedures as previously described, i.e., parameters of the particular kernel function in use, such as σ in Gaussian kernel of Eq. (2.15), and hyper-parameters of regression or classification method such as C and ϵ parameters in Eq. (2.24). The proposed prediction process is evaluated with the available training set using the well established *leave-n-out* (LnO) testing procedure (see e.g. [87]), to be described in Sect. 2.4.3. The accuracy of the classifier is then conveniently represented by constructing an $N_p \times N_p$ classification table (see e.g. [4]), (also known as a "contingency matrix" or "confusion matrix") $\boldsymbol{T}(\Theta)$, where N_p is the number of classes or patterns. Rows are indexed according to the true class of the test object, and columns are indexed by the class estimated by the machine learning procedure. We denote[8] $P_{i|j}$ as the probability of deciding class i when class j is true. This probability may be estimated as $P_{i|j} = t_{j\,i}/t_j$ where $t_{j\,i}$ is the (j, i)th entry of the contingency matrix (i.e., the entry in row j and column i), and $t_j = \sum_{i=1}^{N_p} t_{j\,i}$ is the j-th row sum. The quantity t_j is equal to the total number of subjects in class j, and the number of subjects $M = \sum_{j=1}^{N_p} t_j$. The question is then how to convert this $N_p \times N_p$ table $\boldsymbol{T}(\Theta)$ into a single performance index $\mu(\Theta)$ which is indicative of overall performance.

For this study, optimal values of Θ are determined using an optimization procedure, such as a simple grid search, in conjunction with this performance index as an objective function. The performance index is also useful on its own as a means of quantifying the performance of our proposed classification/prediction procedure.

We now briefly discuss three methods for selecting such an index [88]. One such technique is to minimize the Bayesian decision cost $\mu_B(\Theta)$. The goal is to select Θ such that the average Bayes decision cost, defined as follows, is

[8] Notational dependence on Θ is suppressed.

minimized

$$\mu_B(\Theta) = \sum_{i=1}^{N_p} \sum_{j\neq i, j=1}^{N_p} C_{i,j}\ P_{i|j}(\Theta)\ P_j \tag{2.62}$$

where $C_{i,j}$ is the decision cost associated with deciding class i when j is the true class. The P_j are the prior probabilities of the classes, which in this work are assigned according to a uniform distribution, i.e., $P_j = \frac{1}{N_p}, j = 1, \ldots, N_p$.

It is often desirable to favour one form of miss–classification result over another. For example, in the treatment-response prediction problem, it may be less desirable to miss–classify an actual responder as a non–responder, rather than vice–versa, so that the actual responder is more likely to receive beneficial treatment, assuming there is less cost involved in prescribing a non–effective treatment to a non–responder. A benefit of using the Bayesian decision cost as an optimization criterion is that such considerations may be realized by appropriate choice of the cost parameters $C_{i,j}$. Even though some methods of interest in this thesis like the RLS, SVR and PLSR models do not inherently incorporate a Bayesian decision cost assignment, they may acquire this property through the use of $\mu_B(\Theta)$ as an objective function in their respective training process, when finding design parameters.

The second method is to maximize the average correct decision probability $\mu_H(\Theta)$, defined as follows

$$\mu_H(\Theta) = \frac{1}{N_p} \sum_{i=1}^{N_p} \gamma_i P_{i|i}(\Theta) \tag{2.63}$$

where $\gamma_i > 0$ is a "balancing" factor to weigh the relative importance of each class. The default value is $\gamma_i = 1,\ i = 1, \ldots, N_p$.

When dealing with multi-class classification and/or with an imbalanced classification problem (i.e., when the number of samples in each class are not of the same order of magnitude), then it is worthy of note that $\mu_H(\Theta)$ is a more fair performance measure than the *total classification accuracy (TCA)* which is defined as the number of correct decisions divided by the total number of data samples [9]. Note that the TCA measure is insensitive to transposing the contingency matrix. The *Cohen's kappa* measure which is used in some clinical literature has the same undesirable property as TCA, (see [88]).

Another assessment is based on the KL divergence between the prior probability value P_i, and the estimated correct decision probability $P_{i|i}(\Theta)P_i$. The

[9] TCA performance measure is defined as: $\mu_{\text{TCA}}(\Theta) = \frac{1}{M}\sum_{i=1}^{N_p} t_{ii}$.

following measure is defined as a function of Θ whose exponent is the negative KL divergence value between these distributions

$$\mu_{\mathrm{KL}}(\Theta) \;=\; \exp\left(\sum_{i=1}^{N_p} P_i \log P_{i|i}(\Theta)\right) = \prod_{i=1}^{N_p} \left(P_{i|i}(\Theta)\right)^{P_i} \tag{2.64}$$

where $P_i = \frac{1}{N_p}$ is the uniform prior, but one can use $P_i = t_i/M$ instead. Note that with this definition, we have $0 \le \mu_{\mathrm{KL}}(\Theta) \le 1$, with the value of one corresponding to the ideal model where the contingency matrix is diagonal.

For the experiments of interest in this thesis, a simple multi–dimensional grid search to find the optimal value of Θ was found to yield adequate results. It would indeed be possible to use a more sophisticated procedure such as a Newton-based method (see e.g., [89]) for accelerated convergence behaviour.

Note that for the pipeline inspection application, instead of multiple epochs for each subject as in the psychiatric examples, we have only one data vector for each data sample, and therefore M_t is equal to the total number of training samples. Considering this simple difference, the same performance evaluation methods can be easily used in both applications considered in this thesis.

2.4.3 The LnO Nested Cross-Validation Procedure

The performance of the machine learning process was evaluated using a "leave–n–out" (LnO) nested cross-validation procedure[10] described as follows (see e.g. [4, 87]).

In the neuroscience application (including treatment-response prediction or medical diagnosis), we have multiple epochs for each subject, and therefore in that case, M_t refers to total number of epochs in the training samples belonging to M participating subjects. Since the main work of this thesis was the neuroscience application, in the following the "leave–n–patients–out" procedure for this application will be described in detail. The use of the procedure in the pipeline inspection application is straightforward, where we have $M = M_t$ samples and there is no averaging on the classification results.

Before the LnO procedure is executed, a set of N_c candidate features $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{N_c}$ is extracted from each available training samples $\boldsymbol{E}_i, i = 1, \ldots, M_t$. Here, assume that available training data set after feature extraction is denoted by $\mathcal{D} = \{(\tilde{\boldsymbol{x}}_i, y_i),\ i \in \mathcal{I}\}$, where $\mathcal{I} = \{1, \ldots, M_t\}$ denotes indexes of all epochs belonging to M participating subjects.

[10] For $n = 1$, the 'leave–one–out cross-validation' procedure is denoted by the acronyms LOO or LOOCV in some of the literature. For clinical experiments, the LnO procedure can be alternatively denoted by the 'leave–n–patients–out" cross-validation method.

Then, the M subjects are divided into $P = \lceil M/n \rceil$ contiguous subsets of size n subjects (except possibly the last subset), where $\lceil \cdot \rceil$ denotes the ceiling operator[11]. In other words, P is the number of folds or iterations of the cross-validation procedure. Then, in the p-th fold, $p = 1 \ldots P$, the indexes of all epochs corresponding to the p-th group of subjects form a subset $\mathcal{I}_p$. The remaining indexes are denoted by $\bar{\mathcal{I}}_p$. We form a test subset $\mathcal{D}_p^{\text{test}} = \{(\tilde{\boldsymbol{x}}_i, y_i),\ i \in \mathcal{I}_p\}$, which is omitted from the training process, and a modified training subset $\mathcal{D}_p^{\text{train}} = \{(\tilde{\boldsymbol{x}}_i, y_i),\ i \in \bar{\mathcal{I}}_p\}$ consisting of all epochs from the remaining $M - n$ subjects. We have

$$\begin{aligned} \mathcal{I} &= \mathcal{I}_p \cup \bar{\mathcal{I}}_p && (2.65) \\ \mathcal{D} &= \mathcal{D}_p^{\text{test}} \cup \mathcal{D}_p^{\text{train}} && (2.66) \end{aligned}$$

where the operator $\cup$ denotes union of two sets. Then, the feature selection process and the classifier design procedure are undertaken using only $\mathcal{D}_p^{\text{train}}$ as described further below. The resulting classifier structure is then tested using the test subset $\mathcal{D}_p^{\text{test}}$ which includes the group of n omitted subjects. As we mostly used regression models to do classification, the test is done by averaging the continuously–varying regression outputs over all epochs corresponding to each subject in the set $\mathcal{D}_p^{\text{test}}$, and then quantizing this average into the appropriate predicted response value "R" or "NR". If we use a classification model (with discrete target value outputs) instead, then the final target value for each subject is the majority vote among the classification results for all associated epochs. For statistical methods (when using MFA model, for example), for all epochs associated with a subject, the posterior probability values for each class is averaged. Then, for example, in the maximum likelihood decision framework, the class with maximum probability will be selected as the final output.

To measure the average test error, a contingency table $\boldsymbol{T}(\hat{\boldsymbol{\theta}})$ is then constructed from these P trials, by comparing the estimated responses to the true responses y_i, from which the overall performance may be estimated and

[11] The subjects are reordered before starting the performance evaluation so that we have an approximately balanced number of each class throughout the list. This is similar to the method referred to as *stratified cross-validation* in [87]. Another possible method is to exhaustively process all possible $\binom{M}{n}$ combinations. The difficulty with the second approach is that, even when M has the rather modest value of 22 subjects, for example, the computational requirements become enormous when n is a reasonable fraction of M. We therefore use the stratified cross-validation scheme which is much more reasonable in terms of computational requirements. In this thesis, when the number of subjects is small, using the L1O cross-validation (i.e., with $n = 1$) is preferred since there is only one fair way of selecting the test subjects.

an objective function value $\mu(\hat{\boldsymbol{\theta}})$ may be evaluated, according to one of the methods described in Subsection 2.4.2. $\hat{\boldsymbol{\theta}}$ denotes the set of best design parameter values selected from a multi-dimensional grid of candidate values as explained later.

In the p-th fold of the outer LnO procedure, an inner LmO loop is executed for feature selection and estimation of the parameters $\boldsymbol{\theta}$ (see [58] and [90]). In each fold (indexed by q) of the inner LmO loop the training subset corresponding to each fold of the outer LnO loop is further divided into contiguous subsets of size of m subjects. This is done by dividing the set of training subjects contained in $\mathcal{D}_p^{\text{train}}$ into $Q = \lceil \frac{M-n}{m} \rceil$ contiguous subsets, each of equal cardinality, except possibly one subset. Therefore in the q-th trial of the inner cross-validation loop, m subjects from the training subset constitute the validation subset (denoted by $\mathcal{E}_{q,p}^{\text{validate}} = \{(\mathbf{x}_j, y_j),\ j \in \mathcal{J}_{q,p}\}$ and the remaining samples (denoted by $\mathcal{E}_{q,p}^{\text{train}} = \{(\mathbf{x}_j, y_j),\ j \in \bar{\mathcal{J}}_{q,p}\}$) are used for building the classifier/regressor. For each $q = 1, \ldots, Q$, we have

$$\bar{\mathcal{I}}_p = \mathcal{J}_{q,p} \cup \bar{\mathcal{J}}_{q,p} \tag{2.67}$$

$$\mathcal{D}_p^{\text{train}} = \mathcal{E}_{q,p}^{\text{validate}} \cup \mathcal{E}_{q,p}^{\text{train}} \tag{2.68}$$

In this thesis, the optimal parameter vector $\hat{\boldsymbol{\theta}}$ is selected using a simple grid search using the nested cross–validation procedure. Let the set of possible candidate values/vectors of $\boldsymbol{\theta}$ be denoted as $\boldsymbol{\Theta}_A = \{\boldsymbol{\theta}_a | a = 1, \ldots, A\}$, where A is the total number of candidate values/vectors. The optimal design parameter set $\hat{\boldsymbol{\theta}}$ corresponds to the value of a that results in the best average validation performance. The vector $\hat{\boldsymbol{\theta}}$ is then used to build the classifier/regressor with $\mathcal{D}_p^{\text{train}}$. The classifier is then tested on $\mathcal{D}_p^{\text{test}}$.

The feature selection process can be sensitive to which data samples (and which subjects) are included in the training set $\mathcal{D}_p^{\text{train}}$. Furthermore, it is possible that the regularized feature selection method (described in Section 2.2) automatically picks some redundant or non-significant features. We wish to obtain a set of statistically persistent discriminating features that is a robust representation of all elements of the training subset. We have adopted the following alternative scheme in this respect. At the pth fold of the outer LnO cross-validation procedure and before building the classifier using $\mathcal{E}_{q,p}^{\text{train}}$, we perform a nested L$m$O procedure similar to the one described in Table 2.1 for parameter selection. We divide the set of subjects contained in $\mathcal{D}_p^{\text{train}}$ into Q contiguous subsets of equal cardinality of m subjects. Then feature selection is repeated Q times, where in each iteration, the epochs corresponding to one subset in turn is selected for omission, whereupon the feature selection process is performed using the remaining subsets. In each round, a list of kN_r, $k > 1$ most relevant feature indexes are selected. Then, the reduced feature set for the pth iteration is selected as the set of N_r most commonly occurring feature indexes amongst the Q available lists. This is referred to as a *feature polling procedure*, in which discriminating feature indices which have a 'maximum vote' among the Q subsets are chosen. The specific details relating to feature selection procedure are shown in Table 2.2, whereas the overall performance evaluation procedure is outlined in Table 2.1.

Also, as shown in Table 2.1, the evaluation process can be improved by repeating the whole nested cross-validation process H times, each time randomly permuting the order of the elements in $\mathcal{D}$ beforehand (i.e., two separate permutations: (i) permuting feature indices and (ii) permuting subject order). This can compensate for any problem that may occur from a non-uniform distribution of the R and NR subjects in the training and test subsets as well as any sensitivity that the feature selection may have on the order of features in the feature vector $\tilde{\boldsymbol{x}}_i$. The contingency matrix resulting from the h-th iteration of the LnO procedure is denoted by $\boldsymbol{T}_h(\hat{\boldsymbol{\theta}})$, $h = 1, \ldots, H$. Then, $\mu(\hat{\boldsymbol{\theta}})$ is evaluated from the final contingency matrix $\boldsymbol{T}(\hat{\boldsymbol{\theta}})$ which is taken as the average of $\{\boldsymbol{T}_h(\hat{\boldsymbol{\theta}})\}$, i.e.

$$\boldsymbol{T}(\hat{\boldsymbol{\theta}}) = \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{T}_h(\hat{\boldsymbol{\theta}}) \tag{2.69}$$

For example, in one of our experiments, i.e. treatment-response prediction in SSRI therapy, where there are $M = 22$ available subjects and $M_t = 262$ epochs, $P = 11$, $Q = 10$ and $m = n = 2$ is used. Based on our experiments, we recommend $H = 50$ number of random permutations to limit the overall performance evaluation time.

Throughout all experiments in this thesis, the value of n in the LnO cross-validation procedure is varied with the number of training samples in the respective experiment, so that the number of folds remains constant at a value of approximately 10 iterations. Using something between 10-fold and 20-fold cross-validation is popular in machine learning applications and [87] suggests this procedure in order to obtain a balance between bias and variance of the performance estimate.

As discussed in Section 2.3.7, for all applications studied in this thesis, we also study the clustering performance by looking at two-dimensional representation of the input feature space. The leave-n-out testing procedure introduces a minor complication with regard to each of these clustering experiments. Recall that at each iteration of the LnO cross-validation procedure, n subjects (or n samples in the pipeline inspection case) are omitted and the feature selection and model-training steps are performed on the remaining data points. The resulting classifier/model is then tested on the omitted subjects/samples.

Table 2.1: A description of the nested LnO cross-validation procedure for performance evaluation in the neuroscience applications.

- The M subjects are divided into $P = \lceil M/n \rceil$ contiguous subsets of size n.
- Iterate over P LnO trials: for $p = 1, \ldots, P$

    1. Select the indexes of all epochs corresponding to the p-th group of subjects to form a subset $\mathcal{I}_p$. The test subset $\mathcal{D}_p^{\text{test}} = \{(\tilde{\boldsymbol{x}}_i, y_i),\ i \in \mathcal{I}_p\}$, which is omitted from the training process. The training subset $\mathcal{D}_p^{\text{train}} = \{(\tilde{\boldsymbol{x}}_i, y_i),\ i \in \bar{\mathcal{I}}_p\}$, consists of all epochs from the remaining $M - n$ subjects.
    2. Identify a set of N_r discriminating features over the set $\mathcal{D}_p^{\text{train}}$ based on the method described in Table 2.2. Indexes of N_r discriminating features are denoted by set $\mathcal{F}_p$. The reduced dimensional vector obtained from $\tilde{\boldsymbol{x}}_i$ is denoted by $\mathbf{x}_i,\ i \in \bar{\mathcal{I}}_p$.
    3. *Optimize* $\boldsymbol{\theta}$: Iterate over the multi-dimensional grid of candidate parameters denoted by the set $\{\boldsymbol{\theta}_a,\ a = 1, \ldots, A\}$, where the subscript a denotes the a-th node of the grid:
        (a) Divide $M - n$ subjects in $\mathcal{D}_p^{\text{train}}$ into $Q = \lceil \frac{M-n}{m} \rceil$ contiguous subsets of size of m subjects.
        (b) Iterate over Q LmO trials: for $q = 1, \ldots, Q$
            i. Select the indexes of all epochs corresponding to the q-th group of subjects to form a subset $\mathcal{J}_{q,p} \subset \bar{\mathcal{I}}_p$. The validation subset at q-th iteration is denoted by $\mathcal{E}_{q,p}^{\text{validate}} = \{(\mathbf{x}_j, y_j),\ j \in \mathcal{J}_{q,p}\}$. The remaining indexes in $\bar{\mathcal{I}}_p$ are denoted by $\bar{\mathcal{J}}_{q,p}$ and corresponds to the subset $\mathcal{E}_{q,p}^{\text{train}} = \{(\mathbf{x}_j, y_j),\ j \in \bar{\mathcal{J}}_{q,p}\}$. $\bar{\mathcal{I}}_p = \mathcal{J}_{q,p} \cup \bar{\mathcal{J}}_{q,p}$.
            ii. Design the classifier (realized as a regression process) using the set of N_r discriminating features, the data set $\mathcal{E}_{q,p}^{\text{train}}$ and parameter $\boldsymbol{\theta}_a$.
            iii. Test the design by feeding $\mathbf{x}_j$, the input variables of the validation subset $\mathcal{E}_{q,p}^{\text{validate}}$ to the regression function, obtaining $\hat{y}_j, j \in \mathcal{J}_{q,p}$. Average the regression outputs over all epochs for each subject and quantize into the estimated target values.
        (c) Construct the contingency matrix $\boldsymbol{T}(\boldsymbol{\theta}_a)$ from the Q trials using the true target values and the corresponding estimated values over all $M - n$ subjects. Then evaluate $\mu(\boldsymbol{\theta}_a)$.
        (d) Select the best parameter set, $\hat{\boldsymbol{\theta}}_p$, which yields the best classification/regression performance.
    4. Design the classifier using the set of N_r discriminating features, the data set $\mathcal{D}_p^{\text{train}}$ and $\hat{\boldsymbol{\theta}}_p$.
    5. Construct the reduced dimensionality input test vectors $\mathbf{x}_i$ from $\tilde{\boldsymbol{x}}_i,\ i \in \mathcal{I}_p$, using the set of discriminating feature indexes $\mathcal{F}_p$.
    6. Test the design by feeding $\mathbf{x}_i$, the input variables of the test subset $\mathcal{D}_p^{\text{test}}$, to the regression function, obtaining $\hat{y}_i, i \in \mathcal{I}_p$. Average these over all epochs for each subject and quantize into the estimated target values.

- Construct the contingency matrix $\boldsymbol{T}(\hat{\boldsymbol{\theta}})$ from the P trials using the true target values and the corresponding estimated values over all M subjects. Then evaluate $\mu(\hat{\boldsymbol{\theta}})$.

Table 2.2: A description of the *feature polling procedure* to find a persistent list of most discriminating feature indices in the inner loop of the nested cross-validation procedure outlined in Table 2.1.

1. Divide $M-n$ subjects in $\mathcal{D}_p^{\text{train}}$ into Q contiguous subsets of roughly equal cardinality of m subjects.
2. Repeat Q times: Sequentially select a subset for omission. From the remaining subsets, identify a list of $kN_r,\ \ k \geq 1$ most relevant features.
3. From the Q lists of kN_r most relevant features obtained in the step above, choose the N_r indexes which occur most frequently.

The issue is that at each iteration, a slightly different set of most relevant features may be selected. To remedy this problem, we choose gN_r features at each iteration, where g is a constant greater than unity, typically $g \geq 2$. The final set of N_r features used for low-dimensional representation consists of the N_r most commonly selected features chosen over all the iterations.

Chapter 3

Application to Pipeline Inspection

This chapter addresses the problem of automated non-destructive testing of installed oil and natural gas pipelines using inline *magnetic flux leakage* (MFL) inspection techniques. The MFL technique provides a high-resolution image of the interior of the pipe wall from which defects and other anomalies can be detected and duly reported.

3.1 The Magnetic Flux Leakage Imaging Method

A simplified generic in-line-inspection (ILI) tool is depicted in Fig. 3.1. Permanent magnets magnetize the pipe to saturation or near saturation flux density, typically in the axial direction. Flux excitation in the circumferential direction is also possible. As shown, the magnetic leakage fields from the pipe wall are detected using uniformly–spaced Hall[1] or coil sensors. The sensor is placed at the center of the poles in the magnetic circuit to ensure that it is located in the most uniform part of the field. Systems with standard resolution measure the leakage flux in one direction of excitation, whereas high-resolution MFL systems employ a large number of such sensors that can record anomalies and changes in the magnetic field in one or two directions with sufficient density to recognize even very small pipeline metal defects. The signals from the sensors are sampled as the ILI moves through the pipe. These samples are then recorded on–board the ILI. Once the ILI is retrieved at the

[1] An example of Hall effect sensor which was used in our experiments is built as a solid-state sensor using quad Hall sensing elements. The detection area of the sensor is less than 2mm by 3mm. The output voltage varies in linear proportion to the strength of magnetic field.

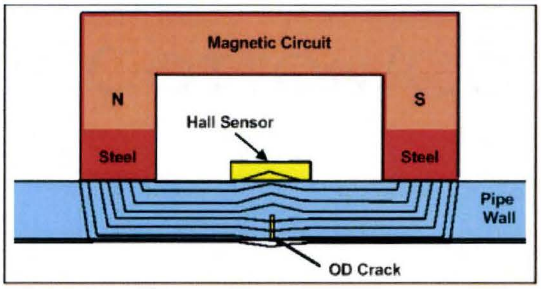

Figure 3.1: Simplified schematic of the magnetic circuit of a typical ILI tool, showing the scattering of the magnetic field due to an external crack. OD stands for 'outer diameter'.

end of the inspection run, the stored signals are then processed and displayed section–by–section on a computer monitor, to render a magnetically–derived 2D or 3D image of the pipe wall.

The magnetic flux from powerful magnetizers is coupled to the pipe wall through steel brushes (or steel protectors). The magnetic sensors are placed in close proximity with the inside surface of the pipe wall for optimum sensitivity to flux variations. The ILI tool body, magnets, steel couplers, and the pipe wall then create a magnetic circuit. A volumetric wall loss in the pipeline wall acts as a region of high magnetic reluctance causing an increased magnetic leakage field. For example, if the wall's thickness is reduced by the presence of a defect (as depicted in the figure), a higher fraction of the magnetic flux leaks from the wall into the space inside and outside the pipe, allowing the defect to be detected by the presence of a corresponding increase in leakage magnetic flux density.

In general, the information that is stored on–board the ILI tool includes measurements of distance, tool speed, temperature, pressure, along with the 2D or 3D samples of the magnetic field through the Hall effect sensors. Other information such as wall–thickness, diameter, properties of the fluid or gas being transported by the pipe, geometry of the pipe, material of the pipe, properties of magnetizers and magnetic sensor arrays are also available to the NDT technician.

The functionality of the ILI tools is divided up into separate modules that are inter–connected by articulating joints. This configuration allows the tool to negotiate elbows and sharp turns that may be encountered in the pipeline. Fig. 3.2 shows two examples of these devices, with each module being identified as 1) the tow section, 2) the main sensor array, and 3) the electronics body.

Other modules (not shown) include battery vessels, odometer modules, sensor arrays for interior/exterior discrimination, and calipers to detect physical damage or protruding artifacts in the pipeline.

ILI tools come in different sizes to fit various pipeline diameters. Outer diameters of pipes used in our experiments are 8.625, 10.75 and 12.75 inches (referred to as 8, 10, and 12 in. pipes in the chapter), with corresponding wall thicknesses equal to 4.77mm, 5.16mm, and 5.6mm, respectively.

A sample of a real MFL image of a metal loss defect in an 8 in. pipe is shown in Fig. 3.3, and its smoothed and de-noised version is shown in Fig. 3.4. The deep blue regions represent a girth weld in the pipe. Fig. 3.5 shows the corresponding contour plot. The de-noising and smoothing method used in our experiments is discussed in Sect. 3.2. The depth of the metal-loss as determined by independent measurement is approximately 60% of wall-thickness, where the wall-thickness is 4.77mm. A single recognizable "bump" corresponding to the defect is apparent in the image. The physical shape of the outer surface of this metal-loss defect, as recorded off–line by an NDT technician is shown in Fig. 3.6.

As may be seen, the MFL image shown in Fig. 3.3 is not well represented by a simple geometric configuration, such as a rectangle. Other real defects may be even more complex. For example, this may happen when the metal-defect is close to or inside a 'long-seam weld' or a 'girth weld'. This justifies our assertion that methods which fit simple geometric models, or characterize defects using a simple low-dimensional parameter estimation methods will not perform well with real measurements. Therefore the benefit of using machine learning in MFL data analysis becomes apparent when we work with real MFL data. In such cases, the machine learning method will automatically find an estimate of the detection/estimation model. However, a drawback of machine learning methods is that they require an adequate and often large quantity of training samples to perform properly. The generation of training data is discussed in Sect. 3.2.

3.2 Experimental Results

In this section, we examine the performance of the machine learning algorithms we have discussed in the analysis of real MFL images. We first investigate the binary detection problem of classifying anomalous image segments into one of two classes: the first class consists of *injurious or non–benign defects* such as various crack-like anomalies and metal-losses in girth welds, long–seam welds, or in the pipe wall itself, which if left untreated, could lead to pipeline rupture. The second class consists of *non–injurious* or *benign* objects such

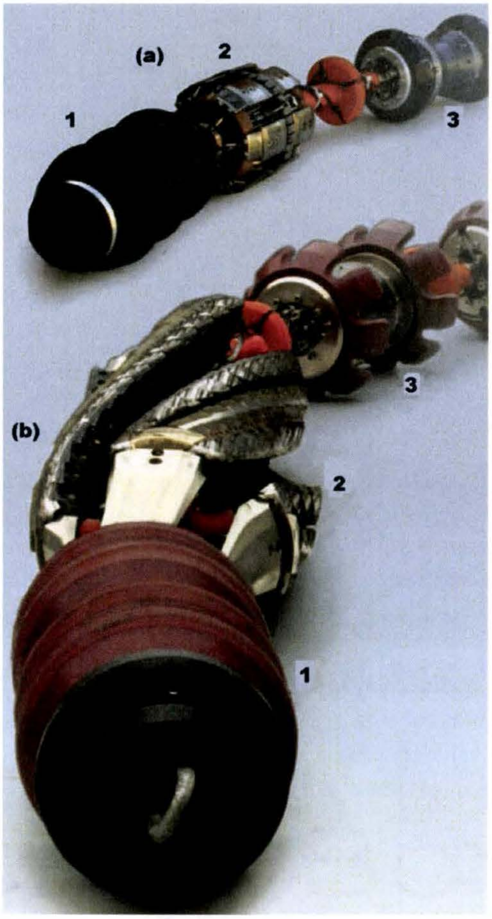

Figure 3.2: Images of frontal parts of two 'in-line inspection tools' or PIGs by Intratech. Printed with permission from Intratech Inline Inspection Services Ltd., Mississauga, ON, Canada.

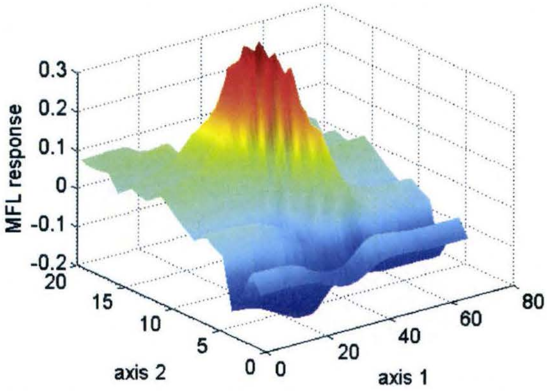

Figure 3.3: A sample real MFL image data for a metal-loss defect in a pipe wall. Printed with permission from Intratech Inline Inspection Services Ltd., Mississauga, ON, Canada. The Z-axis has been scaled.

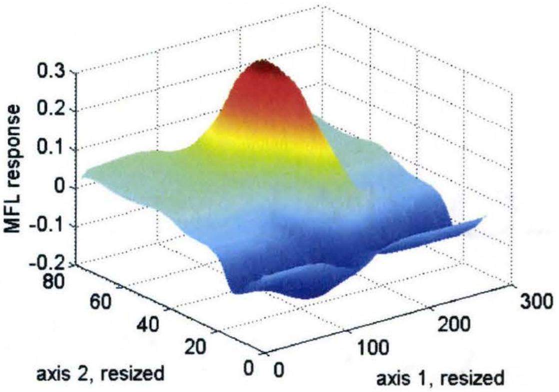

Figure 3.4: De-noised and smoothed version of the sample real MFL image data shown in Fig. 3.3. The Z-axis has been scaled.

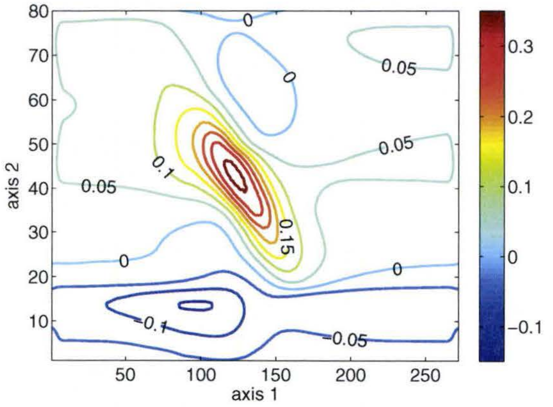

Figure 3.5: Contour plot of de-noised MFL image data of Fig. 3.4.

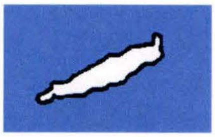

Figure 3.6: Physical shape of the outer surface of the metal-loss defect corresponding Fig. 3.3-3.5. The image is rotated and scaled relative to the images of Figs. 3.3-3.5.

as noise events, safe and non-harmful pipeline deformations, manufacturing irregularities, etc.

The second problem we consider in this thesis is estimation of the severity of defects. In all experiments, we use real MFL data derived from actual pipeline inspection runs provided by Intratech Inline Inspection Services Ltd., Mississauga, ON, Canada. With real data, it is sometimes difficult to guarantee that we have correct labeling for all our training data. For our classification experiments, some target values y_i corresponding to each $\mathbf{x}_i$ may not indicate the true class in some training samples. The reason is that the metal defects are not easily accessible and verifiable; secondly, the labeling of data is done through manual inspection by an NDT expert. Thus, the reference training data are subject to some technical errors and it is likely that we have, for example, some small or complex metal-loss samples mislabeled as non-injurious, and vice-versa. So, in our particular numerical analysis, we do not expect to get 100% performance.

Pipelines used in the experiments are of the 'electric resistance welded' (ERW) long-seam type. The pipe material is standard carbon steel which is used throughout North America for oil and natural gas pipelines, including X42 and X52 pipes as specified in 'API specifications 5L', or the corresponding 290 and 359 grades as specified in 'CSA Z245.1'.

Hall effect sensors are used to measure the circumferential or transverse (B_y) component of magnetic field, where the excitation flux is in the same (y) direction. The output sensitivity of the sensors is 2.5 mV per Gauss. Sensor spacing is 3.32mm in the y direction. Since the sampling rate is time-based and the ILI tool varies in speed due to the variation of transport flow, the sample spacing in the x direction can vary, but is generally in the range of 2mm to 3mm. Due to its compressibility, the flow speed of natural gas varies more than that of crude oil or other liquids.

We now briefly describe the processing steps involved in the following experiments. Initially, the received MFL image is segmented to isolate regions which may contain an object of interest such as some form of metal-defect or injurious event, or some form of benign event. Then we used a two-stage de-noising and smoothing process: First, the level of MFL measurement noise (also referred to as 'seamless pipe noise' or SPN) is reduced, and the MFL image smoothed, by using a pixel-wise 2D adaptive Wiener filtering method based on statistics estimated from a local neighborhood of each pixel [91]. Then to further remove the noise, we used a 2D discrete stationary wavelet transform (with Daubechies wavelets) with soft-thresholding to de-noise the result [92].

Once the image segments have been de–noised, we use one of the feature reduction methods of Sect. 2.2 to automatically pick the N_r most discriminating features. However, the selected set is not unique as some features are statistically dependent. The designer can choose any of these dependent features with only a slight change in performance.

As described above, the initial list of simple candidate features can be large, since at the beginning we do not know which properties of the MFL image are relevant in solving the problem. In the first and second experiments that are explained in this section, a total of $N_c = 81$ simple candidate features are extracted. In the third and fourth experiments, the number of simple candidate features of the MFL images are $N_c = 406$, in which $N_r = 24$, or $N_r = 36$ selected features are used.

The initial set of N_c candidate features consists of quantities commonly used in image classification problems, as explained in [4, 91]. Space does not permit the listing of all specific feature items used for our experiments. However, features of interest include such quantities as statistical properties of average background image information, magnitude, size and scale of the metal defect, orientation, maximum to average ratio of the segment, statistical measurements of the segment, such as means, variances, covariances and coefficients of 2D autocorrelation, aspect ratios, absolute peak values, major principal components of the 2D discrete cosine transform (DCT) of the MFL image segment, various parameters such as skewness relating to the histogram of the segment, various statistical moments (central), normalized integrated optical density, coefficients of the wavelet transform, etc.

In the experiments of this section, we consider a wide range of possible defects that are commonly encountered in practical MFL inspection. These include external corrosion pits, linear crack-like indications for which many are associated with the ERW long seam, and various types of local metal-losses and crack-like defects. Since the defects we wish to process are both internal and external to the pipe wall, the training data sets we employ in the experiments must also include both internal and external defect samples.

For feature selection, the regularized feature selection based on mutual information [68], as discussed in Section 2.2.3, Eq. (2.7), is used.

The following procedures were followed for the implementation of the classification and regression models. For each of the regression (or classification) methods considered in this thesis, the associated design parameter values are determined using a grid search and a cross-validation procedure as described in Section 2.4.3.

3.2.1 Detection of Injurious Defects

Here, we show results for a binary detection case. In this set of experiments, suspicious image segments are classified either into a *injurious* class (including metal loss defects and crack-like anomalies), and a *non–injurious* object class. Then, given that a particular image segment is determined to be an 'injurious' defect, the second step is to estimate the depth of the defect, so that the urgency of repair can be established. This step is discussed in the next subsection.

In the following experiments, we wish to determine the percentage accuracy of our machine learning inspection procedures for classifying defects into their respective classes as described above. Thus, the class of all image segments used in these experiments must be determined in advance. This is done by an NDT technician who examined all the measured MFL image samples considered in this experiment on a computer screen, and based on their experience, manually assigned a classification label (i.e., injurious or non–injurious) to each image segment. Some of these data samples are then randomly selected as training data for the classifier, while the remaining samples are tested by the classifier. This process may be repeated many times, each time drawing a different sample of training samples. The accuracy of our method is then established by comparing the decision made by the classifier to the corresponding value determined by the analyst, for each test case. The actual number of image segments available depend on the experiment and is given later.

In the detection or classification problem, the two classes are

1. The injurious class, including metal loss on a girth weld (labelled as ML-GW), metal loss on a 'long seam weld' (ML-LSW), metal loss on the plain pipe-wall (ML-PW), manufacturing metal-loss samples on a long seam weld (MML-LSW), and a crack-like anomaly on a long seam weld (CL-LSW). The MML-LSW label denotes metal-losses associated with the manufacturing process or any pipeline installation material loss. The corresponding target value is assigned to be $y = 1$.

2. The non–injurious or benign class including measurement noise and benign pipeline manufacturing anomalies (such as dents) and MFL image irregularities that may be confused with metal-loss, but are not harmful to pipeline operation. In general, non–injurious class anomalies do not require replacement or repair of the pipe, and are considered as safe. The corresponding target value is $y = 2$.

In the first experiment, MFL data from an 8-inch pipeline is used. We have a total of $M_t = 1529$ image segments, consisting of 656 injurious, and 873 non-injurious samples. The data in the injurious class include 197 ML-PW, 132 ML-LSW, 58 ML-GW, 116 MML-LSW and 153 CL-LSW samples. A 10-fold cross-validation method is used, as described in Section 2.4.3. After applying a random permutation on the available data set, in each fold of the cross-validation procedure, 90% of the available data samples are used for training, while the remaining 10% are used for testing.

A comparison amongst the different identification techniques is shown in Table 3.3. The parameters for each method were estimated by optimizing the performance using a grid search and cross-validation procedure using the training data, as described in Chapter 2. The linear discriminant analysis (LDA) method shown in the table is a well-known classification technique used as a reference in our comparison [4]. In a manner similar to [13], we also used a multilayer Perceptron network (MLPN) (with 18 hidden neurons, and employing a conjugate gradient method for training). We also evaluate performance using a radial-basis function network (RBFN), [3, 4]. As can be seen, RLS, SVR and kernel PLSR methods with a Gaussian kernel are all powerful methods and have approximately similar performance in this experiment, but are all better than the LDA, MLPN [13] and RBFN methods. For the SVR method, the identification results with $N_r = 36$ corresponds to the following number of misclassifications in each class: 19 misclassification cases in the injurious class (including 12 ML-PW errors, 3 ML-GW, and 4 CL-LSW), and 13 misclassification cases in the non–injurious or benign class. The result by the kernel PLSR method corresponds to using 9 major latent vectors. By experimenting with various scenarios (not discussed in this chapter) we also noticed that the SVR method, compared to others, can provide good detection performance even with relatively small number of training samples.

Performance was also evaluated for the $N_r = 24$ case; however, as can be seen in the lower part of the Table 3.3, the results for the SVR and kernel PLSR methods did not vary significantly from the $N_r = 36$ case. This is in contrast to the LDA method which showed a significant drop in performance for the $N_r = 24$ case.

Fig. 3.7 shows regression results using the kernel PLSR method (with a Gaussian kernel). Here, the data are sorted so that the first 656 samples correspond to class 1 defects (i.e., injurious class), and the remaining 873 are non–injurious anomalies. In this case, the function f is a step function, with its discontinuity between values 656 and 657. The figure shows how the kernel PLSR method performs in fitting the data. A classification error results when a data sample with index in $[1, 656]$ is above the value 1.5, and is below 1.5 when its index is in the $[657, 1529]$ range. From Table 3.3, we see that this

Table 3.3: 8-inch pipeline: Comparison of performance among different injurious metal defect identification methods. A target value $y = 1$ corresponds to an injurious data sample, while a target value of $y = 2$ corresponds to a non–injurious class. Gaussian kernels are used with the SVR and PLSR methods.

method	sensitivity, $p(\hat{y} = 1 \mid y = 1)$	specificity, $p(\hat{y} = 2 \mid y = 2)$	% average performance
LDA, $N_r = 36$	0.9	0.94	92%
Multilayer Perceptron, $N_r = 36$	0.931	0.976	95.37%
RBFN, $N_r = 36$	0.942	0.951	94.64%
RLS, $N_r = 36$	**0.976**	**0.978**	**97.69%**
SVR, $N_r = 36$	**0.971**	**0.985**	**97.81%**
kernel PLSR, $N_r = 36$	0.951	0.99	97.05%
LDA, $N_r = 24$	0.878	0.915	89.64%
SVR, $N_r = 24$	0.97	0.979	97.44%
kernel PLSR, $N_r = 24$	0.959	0.977	96.79%

method correctly classifies the MFL samples 97% of the time.

We now show how dimensionality reduction of the feature space as discussed in Chapter 2 can aid in the visualization of the classifier. This can be readily accomplished using the kernelized PCA procedure. In this case, a Gaussian kernel was used. Fig. 3.8 shows the corresponding two–dimensional scatter plot for the same data set. Even though this two–dimensional representation is not sufficient for the most separable view of the two classes, it allows visualization on a sheet of paper. It is clear that this low-dimensional representation indeed shows reasonable geometric separation and clustering of subjects into injurious (shown by yellow rectangles with black edges) and non–injurious (shown by blue circles) classes. There are of course, a few overlapping points and these would lead to classification errors, but performance would be improved by an increase of dimensionality. Fig. 3.8 corresponds to the projection of the data samples onto the first and second major latent variables, which are selected through the maximum mutual information method.

We now show a second example of how dimensionality reduction can assist in visualization. Here we consider only the injurious group of data. We used kernel PCA (KPCA), with a Gaussian kernel, to show how different types of metal-losses cluster geometrically in a two–dimensional space.

The input into the KPCA algorithm were feature vectors of dimension $N_r = 36$. Fig. 3.9 shows the relative magnitude of the first several major principal components (eigenvalues) of the MFL data kernel matrix. The eigenvalues are sorted in descending order, and are scaled so that the first eigenvalue with biggest magnitude corresponds to 100 in the y-axis. It is clear that the first two

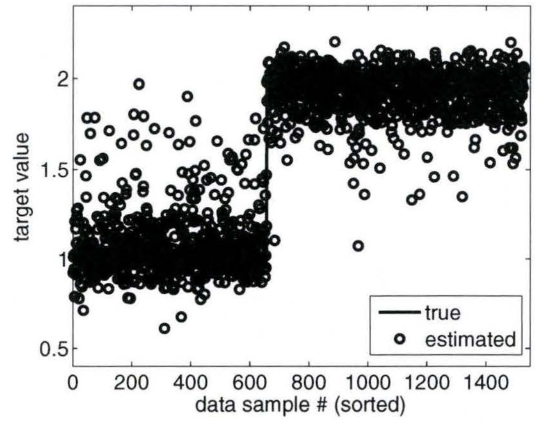

Figure 3.7: 8-inch pipeline: Estimated (circles) and true (solid-line) target values using the Kernel PLS regression method with a Gaussian kernel. The data samples are sorted. $y = 1$ denotes injurious class and $y = 2$ denotes non-injurious class. $N_r = 36$.

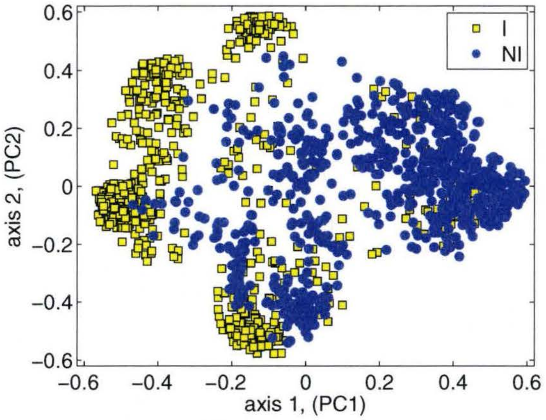

Figure 3.8: 8-inch pipeline: Scatter plot of the projection of MFL data samples into the first and second major latent variables using the KPCA method with a Gaussian kernel. 'I' denotes injurious (metal-loss) samples (including ML-GW and ML-PW classes), and 'NI' denotes non-injurious samples.

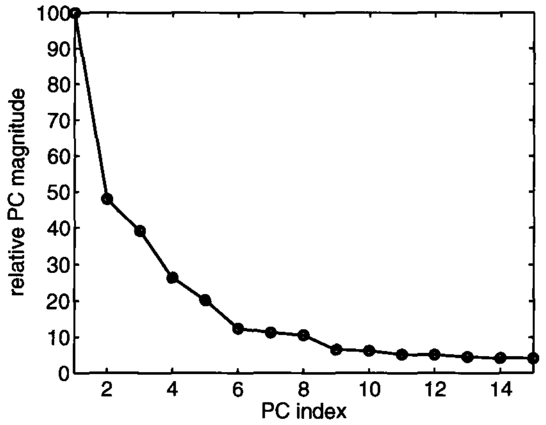

Figure 3.9: 8-inch pipeline: Plot of the relative magnitude of a few major principal components (PCs or eigenvalues) of the MFL data kernel matrix using the Kernel PCA method with a Gaussian kernel.

or three components dominate. This suggests that only a few latent variables with the highest PC magnitude could be enough to represent the data with small error. Fig. 3.10 shows the projection of the feature vectors onto the first and third major latent variables. These particular latent variables were again chosen using the maximum mutual information criterion. As can be seen, there are some overlapping samples in the figure, because the dimensionality of two is too low to adequately cluster the data. Nevertheless, only two dimensions were able to capture the data clustering and show the behaviour of the classifier.

As an example of the usefulness of this approach, a new test MFL data sample can be projected into this 2D space and the user/operator can then infer or visualize approximately the class of test sample based on how closely it is to the clusters for each class.

For clustering analysis through low-dimensional representation, in addition to using the KPCA method (as reported above), we further studied the results by the 'isometric embedding' (ISOMAP) [93], the 'locally linear embedding' (LLE) [94], and the Graph Laplacian [95] methods. The 2D and 3D graphs obtained by these methods had approximately similar clustering performance, however, the shape of the graphs were different in each method. These further results are not reported in this thesis due to space limitation. Nevertheless, these this satisfactory clustering performance confirms the automatic pipeline inspection methodology developed in this thesis.

In the second experiment, inspection data for a 10-inch pipeline is used for

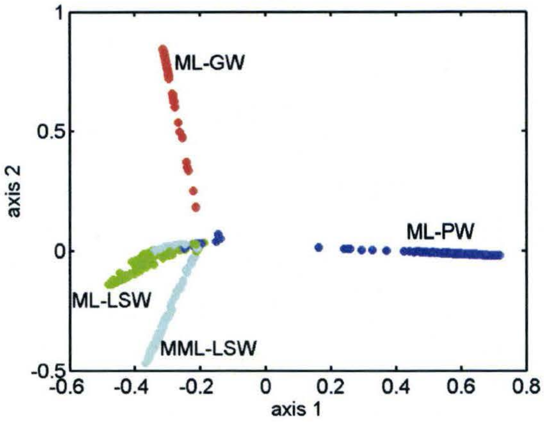

Figure 3.10: 8-inch pipeline: Scatter plot of the projection of MFL data samples into the first and third major latent variables using the KPCA method with a Gaussian kernel. For clarification, green dots correspond to ML-LSW, and cyan dots correspond to MML-LSW ('manufacturing metal-loss samples on long-seam weld').

performance evaluation. The performance numbers are shown in Table 3.4. Here, there are a total of $M_t = 1919$ samples, 909 of which correspond to injurious events and 1010 are non–injurious or benign anomalies. The data in the injurious class include 403 ML-PW, 120 ML-LSW, 109 ML-GW, 137 MML-LSW and 140 CL-LSW samples. The methodology used in this case is similar to that used in the first experiment. The injurious samples mis-classified with the SVR method comprise 11 ML-PW errors, 2 ML-LSW errors, 10 ML-GW errors, 1 MML-LSW, and 1 CL-LSW error. As can be seen, the kernel PLSR method outperforms the linear PLS regression method.

3.2.2 Metal-Loss/Crack Depth Estimation

In this case, we are given an image segment which corresponds to a injurious (including non–benign metal loss and crack-like defects), and we wish to assess its severity, or depth, relative to the pipe wall thickness. The MFL image data set used in the following experiments is distinct from those used in previous subsection, and corresponds to several 8, 10 and 12 in. pipes.

We use two approaches for this problem. In the first approach, we quantize the metal-loss/crack depth into two levels:
$y = 1$, 'less-severe loss': metal-loss/crack depth is less than a threshold δ.

Table 3.4: MFL data for a 10-inch pipeline: Comparison of performance among different metal-loss identification methods. $y = 1$ corresponds to an injurious metal defect, and $y = 2$ corresponds to a non–injurious or benign class. Gaussian kernels are used in kernelized PLSR and SVR.

method	sensitivity, $p(\hat{y} = 1\|y = 1)$	specificity, $p(\hat{y} = 2\|y = 2)$	% average performance
LDA, $N_r = 36$	0.879	0.863	87.12%
RLS, $N_r = 36$	0.968	0.966	96.72%
SVR, $N_r = 36$	**0.973**	**0.969**	**97.09%**
kernel PLSR, $N_r = 36$	0.946	0.977	96.17%
PLSR, $N_r = 36$	0.88	0.862	87.12%

$y = 2$, 'severe loss': metal-loss/crack depth is greater than or equal to the threshold δ.

In the second approach, we treat the depth as a continuous variable and use regression methods to estimate the depth. The defect depth refers to the maximum depth of the defect.

The dimensions of the metal defect anomalies used in this subsection are measured by a NDT technician at an actual dig-site, using various techniques.

In the following two experiments, a total of 58 samples of various metal loss and crack-like defects are used as our training/test data. Defect depths vary in the range of 10% to 61% of wall-thickness, where they reasonably uniformly cover a full range of values in the above depth range. Most of the crack-like defects are adjacent to or on the 'long-seam weld'. The data set contains both internal and external defects. The data set comprises 33 metal losses, 4 metal losses on a 'long seam weld' (LSW), 20 cracks and crack-like events on an LSW, and 1 crack-like event adjacent LSW.

In our third experiment, we show results from the quantized depth estimation approach. In this case, we use $\delta = 30\%$ of pipeline wall-thickness as our severity threshold. Performance results are shown in Table 3.5. There are 29 data samples in each class. It is seen that the SVR method with a polynomial kernel (of degree 7) performs better than the other methods. The nearest neighbor method shown in the table is a well-known classification technique used as a reference in our comparisons [4].

The fourth experiment shows results for the continuous-valued depth estimation approach. The data set is the same as that used in the third experiment. Fig. 3.11 shows the metal-defect depth estimation results using the SVR method with a polynomial kernel of degree 9. With reference to Eqs. (2.14) and ε-insensitive loss function of SVR, the parameters are $\epsilon = 0.05, \gamma =$

Table 3.5: Comparison of performance among different metal-loss severity identification methods. $y = 1$ corresponds to 'less-severe loss' or shallow loss-depth, and $y = 2$ corresponds to 'severe loss' or deep loss-depth. $N_r = 36$. The severity threshold is $\delta = 30\%$ of pipeline wall-thickness.

method	sensitivity, $p(\hat{y} = 1 \mid y = 1)$	specificity, $p(\hat{y} = 2 \mid y = 2)$	% average performance
Nearest Neighbor	0.621	0.828	72.41%
RLS	0.931	0.897	91.38%
SVR with linear kernel	0.931	0.931	93.1%
SVR with Gaussian kernel	0.966	0.931	94.83%
SVR with polynomial kernel	**1**	**0.966**	**98.28%**
PLSR with Gaussian kernel	0.966	0.897	93.1%

$0.001, d = 9$. The dashed line corresponds to the output of the regression function f using the SVR method, whereas the solid line (with sample points marked by circles) shows the true relative depth of the corresponding data samples. With this figure, the exact true distribution and range of defect depths can be seen. The mean-squared error for depth-estimation in this experiment is 7.73% (of pipeline wall-thickness). Given the simple set of features used, this figure is of sufficient accuracy, especially considering the complexity in the shape and structure of various actual metal defects used in this case, to determine repair schedules for the pipeline. In comparison, in a separate experiment (not shown), conducted using the radial basis function network (RFBN), the rms error was determined to be 9.94%.

An example, further to that of Fig. 3.4, of a defect that was included in the above experiment, is shown in Fig. 3.12. This shows a two-dimensional plot of the de-noised MFL signal amplitude of an actual metal-loss on an LSW for an 8-inch pipe. In contrast, Fig. 3.4 shows an MFL response of an actual metal loss on plain pipeline wall. In this case, the running speed of the ILI tool was 0.778 m/s. The rectangular area sampled by the Hall effect sensors before denoising was 58 × 115 sample points. The depth estimated using the SVR method discussed above is 45.7%, which corresponds to an error of 3.3% relative to the value obtained by field measurement. The signal amplitude and axis values in the figure have been scaled. It may be seen from this figure that the defect shape is very irregular. The fact that the proposed machine learning technique is capable of producing an accurate result shows the effectiveness of these methods over previous traditional methods which assume a very limited class of defect geometries, e.g. [27].

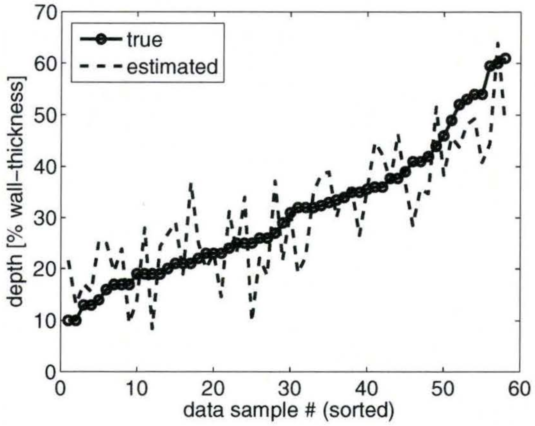

Figure 3.11: Metal-loss/crack-depth estimation: Sorted estimated (dotted-line) and true (solid-line) depth values using SVR with a polynomial kernel. $N_r = 24$.

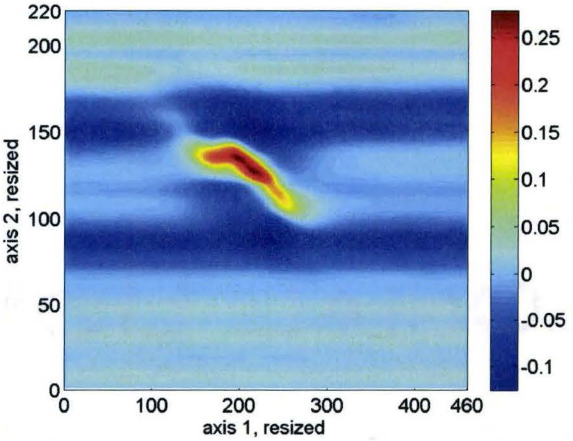

Figure 3.12: 2D view of an example MFL image (after de-noising) showing a metal-loss on LSW.

3.3 Discussion

Testing of oil and gas pipelines using MFL technique involves the detection of defects and anomalies in the pipe wall, and the evaluation of the severity of these defects. The difficulty with the MFL method is the extent and complexity of the analysis of the MFL images. In this thesis we showed how modern machine learning techniques can be used to considerable advantage in this respect.

In conclusion, in this part of thesis, a complete machine learning procedure for the inspection of MFL images from pipelines is presented and the test results using real data is studied. Mainly the detection of major metal-defects in a pipeline is studied, but two techniques for determining the depth and severity of the defect is also included. The average detection performance in recognizing major metal-defects versus benign or noise defects was over 95%. The root-mean square error in defect depth estimation was less than 8%. Furthermore, the kernel PCA method provided a low-dimensional representation of the MFL data and was presented as an effective visualization tool.

Several experiments were performed to verify these results are not due to over-fitting. Firstly, the performance figures indicated were obtained using a cross-validation procedure, and as such the reported performance results are an aggregate over many permutations of the data. Thus, relatively high performance must be obtained in each fold, which implies the feature selection and classification/regression processes are not over-fitting to the data.

Further, the clustering performance shown in Figs. 3.8 and 3.10 show that a simple nonlinear transform exists which is capable of relatively cleanly separating the clusters with a very simple boundary. The simplicity of the transformation in combination with the boundary, show that the classifier is not over-adapting to the data.

As a further indication of the integrity of the processor, we calculate the probability of these results happening by chance alone. With reference to results by SVR method with $N_r = 36$ reflected in Table 3.3, there are 656 injurious and 873 non-injurious samples, so the probability $p = p(\mathbf{x})$ of an injurious defect may be taken as $656/1529 = 0.429$. Assuming all subjects are independent, the probability of a prediction error is governed by a binomial distribution, which is parameterized by N, the number of samples, and p, in this case the probability of an injurious defect. Therefore, the probability of this level of performance (637 classifications as injurious and 19 as non-injurious out of N=656 actually injurious defects) occurring due to chance alone is evaluated from the binomial distribution as 3.9582×10^{-203}. Similarly, the value of p for the non-injurious case is 0.571, so the probability of estimating 860 non-injurious and 13 injurious out of 873 actually non-injurious samples due to chance alone is 2.0146×10^{-186}. Similar analysis can be done for results in Tables 3.4 and 3.5.

Chapter 4

Application to Neuroscience: Prediction of Treatment-Efficacy in Psychiatric Disorders

Here, several experiments are investigated in which pre-treatment clinical information is used to predict the treatment response for each individual patient suffering from major depressive disorder (MDD) or from schizophrenia, at the outset of therapy, thus improving therapeutic efficiency. Three treatments are studied: (i). Selective serotonin reuptake inhibitors (SSRI), an antidepressant for MDD, (ii). 'repetitive transcranial magnetic stimulation' (rTMS) therapy for MDD, and (iii). Clozapine therapy for schizophrenic patients.

When administering a particular therapy for a particular subject with MDD, a significant problem is to predict whether or not the therapy will be effective for the subject. In the larger view, a major problem in the treatment of MDD is that there are no objective procedures for selecting optimal treatments. In the following, we describe the application of machine learning methods, based primarily on the pre-treatment electroencephalography (EEG), for predicting the response of therapy to MDD. The SSRI and rTMS therapies are first studied. Then we will investigate the same problem using the same tools for prediction of response to clozapine therapy for schizophrenia.

All clinical studies, including the rTMS, SSRI, clozapine as well as al medical diagnosis studies discussed in this thesis (in the current and next chapters) were all approved by the Research Ethics Board of St. Joseph's Health Care, Hamilton, ON, Canada.

4.1 Clinical Data Analysis Procedure

The data analysis used in all clinical studies including the SSRI, rTMS, clozapine as well as medical diagnosis studies (to be discussed in next chapter), share the following procedure.

As described in Chapter 2, M_t epochs of pre-treatment resting or spontaneous EEG signals, are collected from M subjects who participated in the study. They were then prescribed the assigned therapy. The corresponding response outcome y of the patient to the treatment, after completion of a full treatment plan, is recorded. The possible values for the y are either "R" (responder), or "NR" (non–responder). The set of pre-treatment EEG recordings and the corresponding outcomes is referred to as a *training set.* These EEG signals from each subject are pre-processed to extract a large number N_c of candidate features that might be relevant for prediction of response. The regularized feature selection method explained in Section 2.2.3 is then used to find the indices of N_r discriminating features from the set of N_c candidate features using the training set to extract those features which are most indicative of the response outcome. These reduced-dimensionality features are then fed into a classifier which outputs the predicted response to the treatment.

We are interested in using regression methods (like SVR and PLSR) for classification. As explained in Section 2.4.3 and particularly in Table 2.1, in the operational phase, when predicting the response to a therapy, or making a diagnosis recommendation, the regression function is applied to each epoch in the test set. Then the regression outputs for all epochs belonging to each subject are averaged. This average value is then quantized to the nearest value (e.g., responder or non-responder class) and then this final decision value is reported as the estimated target value for the subject. When one wants to use a discrete classification method (like NN) instead, the final decision for each subject is the majority vote among the classification results for all associated epochs. For statistical methods (like MFA), for all epochs associated with a subject, the posterior probability values for each class is averaged. Then, for example, in the maximum likelihood decision method (as described in Section 2.3.5), the class with maximum probability will be selected as the final output.

We use the LnO nested cross-validation procedure explained in Section 2.4.3. Based on the recommendation in [87], we try to use a P-fold cross-validation procedure for performance analysis in most of our experiments, with $10 \leq P \leq 20$.

4.1.1 Details of EEG Data Recording

The EEG signals were collected under the auspices of Drs. Gary Hasey and Duncan MacCrimmon, Dept. of Psychiatry and Behavioral Neurosciences, Faculty of Health Sciences, McMaster University. Dr. Hasey is also with Mood Disorders Program, and Dr. MacCrimmon with the Schizophrenia Program at the St. Joseph Hospital, Hamilton, ON. All the clinical studies in this thesis are partly supported by the Canadian Psychiatric Research Foundation, the Ontario Mental Health Foundation and the Stanley Foundation.

The EEG recording procedure for the following experiments is described as follows. Twenty channels of EEG (standard 10–20 system referenced to linked ears) were recorded at a sampling frequency of 205 Hz, after approximately 10 days of medication withdrawal and before a 6-week trial of antidepressant treatment was administered. For the clozapine study however, the pre-treatment EEG data were collected without change to the patient's current medication regimen. A QSI-9500 EEG system is used, which filters the signals between [0.5Hz–80Hz] band and applies a notch filter at 60 Hz. During recording of the resting EEG data, the patient was in a semi-recumbent position in a sound attenuated, electrically shielded room. The process was administered by an experienced technician who prompted patients on signs of drowsiness. Sessions were arranged in the mornings and patients were requested to avoid coffee, drugs, alcohol and smoking immediately prior to the recording. For each patient, a maximum of 6 EEG data files each of 3.5 minutes duration were collected, 3 under eyes open (EO), and 3 under eyes closed (EC) conditions.

For de-artifacting, the data were partitioned into segments of 1 second duration. If the input signal on any electrode saturated the acquisition hardware at any time throughout the segment, the entire segment was rejected. The signals were then digitally bandpass filtered after recording between 3Hz and 38 Hz to partially mitigate the effects of eye movement and muscle artifacts.

For each EEG file, 60 seconds of the 3.5 minutes of data which are not contaminated by artifacts are selected. The selected data are divided into 2 epochs of 40 sec. duration with 50% overlap. Each epoch is further divided into 1 sec. windows with 50% overlap to calculate the statistical quantities which become the candidate features as described below. A 1 sec. window length enables analysis of frequencies above approximately 3 Hz. These settings result in a total of a nominal $M_p = 12$ epochs per subject[1].

The EEG data from 16 EEG electrodes of Fp1, Fp2, F7, F8, F3, F4, T3, T4,

[1] Some of the EEG recording sessions were missing and therefore some subjects only have 8 or 10 epochs available.

C3, C4, T5, T6, P3, P4, O1, O2 are used (excluding data in midline electrodes FZ, CZ, PZ and OZ). It was observed that most of the reduced (most relevant or most discriminating) features used for the prediction process are bilateral variables indicating e.g., power ratios, coherences, etc. from electrodes on opposite sides of the central parasagittal boundary. Also, we were interested in using an smaller number of EEG electrodes to reduce data acquisition time, as well as data processing time and memory usage. In another arrangement of our setup, we observed that using the 16 electrodes above exhibited similar performance compared to the case where all 20 electrodes are used. Therefore, the 16 electrodes mentioned above were used in our main experiments, but we also did some experiments with an even smaller number of EEG electrodes, as discussed later.

4.1.2 Candidate Numerical Features Extracted from EEG

The candidate features extracted from each data epoch include statistical quantities such as the spectral coherence between all electrode/channel pairs at various frequencies[2], the mutual information between electrode pairs, absolute and relative power spectral density (PSD) levels at various frequencies[3], the log ratio of left-to-right hemisphere powers, and anterior/posterior power ratios at various frequencies and between various electrode pairs. The use of 2nd–order statistics as candidate features can be justified by considering that a complete set of statistical moments fully describes a multi–variate random process. In our study, the statistical quantities that are selected as candidate features are a subset of the complete set of moments, and therefore offer a partial description of the process generating the EEG observations. Such quantities have been used in previous related studies; e.g., PSD values are used in [96–101]; left-to-right hemisphere powers are used in [98, 99]; and anterior/posterior power ratios are used in [99]. The work of [102] and [98, 99] used coherence between electrode pairs to assess the effect of the anti–psychotic drug *clozapine* and characterize depression, respectively. Also [100, 101, 103, 104] have used inter and intra-hemispheric power ratios as numerical indicators or predictors of treatment response. Mutual information was used in [105] to analyze EEG data abnormalities in 10 schizophrenic subjects as compared to 10 normals. Further, statistical quantities such as these have also proven useful in related

[2] The magnitude squared coherence estimate was calculated using the Welch averaged periodogram method by the *MathWorks MATLAB* software, ver. 7.1. See *www.mathworks.com.*

[3] Welchs' averaged modified periodogram method was used with *MATLAB* to calculate PSD estimates.

previous EEG classification problems [106, 107]. With the frequency resolution and the number of EEG electrodes considered in these experiments, the number of candidate features $N_c = 6988$.

4.1.3 Feature Normalization

Clinical feature values are normalized before feature selection and classification to improve performance. A convenient means for accomplishing this purpose is the *z-score*. Using EEG data from 91 normal subjects, the means μ_ℓ and standard deviations $\sigma_\ell, \ell = 1, \ldots, N_c$ of each candidate feature are calculated. The ℓth feature value $\tilde{x}_\ell$ from a feature vector $\tilde{\boldsymbol{x}}$ is then replaced by its z-score value $\frac{\tilde{x}_\ell - \mu_\ell}{\sigma_\ell}$. An alternative method studied in our experiments is to normalize each feature to have a maximum absolute magnitude of unity so that each feature is in the interval $[-1, 1]$. In this thesis, this method is referred to as the *Max1Mag* method. In our various experiments, we observed approximately similar average prediction performance with both of these normalization methods (i.e., with both *z-score* and *Max1Mag* methods). However the list of N_r discriminating features was slightly different for the two methods. The *z-score* normalization method is preferably used in this thesis.

4.2 SSRI Therapy for MDD

'Selective serotonin reuptake inhibitors' (SSRIs), are a class of antidepressants widely prescribed in the treatment of depression, as well as in the treatment of some other psychiatric disorders. SSRIs affect the level of a neurotransmitter *serotonin* that neurons in the brain use to send messages to one another. Messages are passed between two connecting nerve cells (neurons) across the synapse, a small gap between the cells using chemical mechanisms. For a signal to be propagated along a chain of neurons, neurotransmitters such as serotonin are released into the synaptic cleft by a proximal or presynaptic neuron then diffuse across the synaptic space to bind to receptors on the surface of the recipient or post-synaptic neuron. This binding initiates a series of events in the post-synaptic neuron which results in changes to the electrical activity (depolarization or hyperpolarization) of the post synaptic neuron. If the neuron is depolarized this signal is propagated across another synaptic cleft to the next neuron in that circuit. After binding to a synaptic receptor, neurotransmitters eventually uncouple from the receptor and are again released into the synaptic cleft where they remain active by again binding to a synaptic receptor or are deactivated by being taken up by the presynaptic neuron using an active transport system. This process is termed 'reuptake'.

As depression may be the result of insufficient binding of the neurotransmitter serotonin to post synaptic receptors, SSRIs may work by inhibiting the reuptake of serotonin, an action which allows more serotonin to be available for binding with post synaptic receptors [108].

Some most commonly prescribed SSRIs are paroxetine (with brandname or trade-name 'Paxil'), fluoxetine (with brandname 'Prozac') and sertraline (with brand name 'Zoloft').

As described previously in Chapter 1, a methodology that can employ pre-treatment measures to predict the response to an SSRI treatment, such as the one proposed in this thesis, would eliminate the inefficient trial-and-error process that often characterizes the management of MDD.

4.2.1 A Review of Previous Studies

Over the years, strategies have been developed to employ resting EEG, or quantitative EEG (QEEG) data, especially of the prefrontal area, as a method for understanding the biological heterogeneity of psychiatric syndromes and predicting treatment outcome in depressed subjects [96, 97, 107, 109–117]. In our earlier (not yet published) studies examining pre-treatment QEEG spectra we observed superior antidepressant response to a new physical treatment for depression, repetitive transcranial magnetic stimulation (rTMS), in subjects whose pre-treatment QEEG showed frontal alpha power asymmetry with relatively greater alpha power in the right compared to left frontal cortex. These findings together with those of others who reported that pre-treatment right-hemisphere delta and theta absolute powers were increased in drug-free depressed subjects, compared to controls, [100] suggest that the inter-hemispheric balance may play an important role in the pathophysiology of MDD. This is further supported by the observations of Bruder et al [101] who, employing EEG together with dichotic listening, found that greater alpha power in the right compared with the left hemisphere was associated with favorable response to the anti–depressant drug fluoxetine. The opposite hemispheric asymmetry predicted poor response.

With respect to the application of more sophisticated mathematical approaches to treatment response prediction in MDD, a quantitative approach using artificial neural networks (ANN) employing clinical and demographic, but not EEG, features has been developed [35]. However this approach offered no better results than using traditional techniques.

There have been other machine learning approaches for classification of EEG signals for prediction of response to treatment for MDD [118, 119]. However, these methods have resorted to *ad hoc* methods for feature reduction and classification, in that no explicit criterion of optimality is used for these processes. The method of Greenwald et al. [119] classifies a multi-variate set of features by performing a sequential one-dimensional decision rule on each individual variable. In contrast, the method developed in this study applies the machine learning principles that directly employ information theoretic, least-squares or other optimality criteria as discussed in Chapter 2.

Quantitative EEG measures such as 'cordance' [96] among others have shown potential value in predicting treatment outcomes. Cordance is a quantitative measure calculated from fast Fourier transformed EEG data, processed to obtain absolute and relative power in 4 frequency bands (delta: 0.5-4 Hz, theta-I: 4-8 Hz, theta-II: 8-12 Hz, and Beta: 12-20 Hz), localized to regions of the scalp. Cordance has been shown to be correlated with cortical perfusion in the region beneath each electrode. Medication responders reportedly show decreases (compared to pretreatment cordance) in prefrontal cordance measured after 48 hours and 1 week of treatment [96, 112]. In this study, pretreatment cordance by itself is not predictive of treatment response.

Loud sounds activate serotonergic mechanisms in the frontal cortex and hypothalamus in rats and it has been suggested that the 'loudness dependent auditory evoked potentials' (LDAEP) may reflect central serotoninergic activity. In humans enhancement of serotonin function with the SSRI citalopram decreases the LDAEP [114]. A 'strong' LDAEP in a patient with MDD may therefore reflect serotoninergic underactivity, this in turn indicating potential response to drugs which enhance serotoninergic activity such as the SSRIs. A 'weak' LDAEP may indicate the need for a medication that affects other neurotransmitters [115]. In this thesis, LDAEP data were not available therefore we used only the resting EEG.

Sleep polysomnography is another method of obtaining measurements and analysis of EEG data during various phases of sleep. Sleep polysomnography (delayed 'rapid eye movement'(REM) onset, increased delta sleep ratio) to predict antidepressant response were not particularly successful, [116]. While response to antidepressant may be correlated with changes in REM sleep [120], this involves measurements taken after the antidepressant has been started. Furthermore nocturnal EEG collection is lengthy and logistically complicated. During the acute state of depression, sleep continuity measures were associated with the number of previous depressive episodes, but did not correlate with the prospective course. However decreased slow wave sleep and increased rapid eye movement density were predictive of recurrences [121]. Patients with abnormal sleep EEG profiles are reported to have significantly poorer clinical response to short-term interpersonal psychotherapy than the patients with more normal sleep profiles. This suggests that underlying neurobiology as indicated by an abnormal sleep EEG may reflect a more marked disturbance of 'central nervous system' (CNS) arousal that warrants pharmacotherapy, [122]. However, as sleep EEG recording is logistically complex, expensive and very time consuming, methods of estimating clinical response to antidepressant treatments that employ standard waking EEG signal are preferred.

There are some other works related to neuroimaging (and particularly functional magnetic resonance imaging) in MDD. See [123] for a review. The most widely reported functional imaging abnormality in patients with MDD are prefrontal cortical hypoactivity [124], and hyperactivity of the subgenual cingulate [125] and the anterior cingulate cortex (ACC) [126]. Subjects with higher metabolic rates in these regions of the cingulate cortex reportedly respond better to sleep deprivation as well as to the antidepressant medications paroxetine, sertraline or venlafaxine [127–129]. These findings have been replicated with a low resolution EEG tomographic analysis (LORETA) of resting EEG-data, with increased current source density in the anterior cingulate cortex in responders to nortriptyline in the theta-frequency range [107].

These results, for the most part, have been observed during the resting state. It has been suggested that activation or challenge paradigms, and particularly those which use mood induction paradigms "may prove to be a more productive direction to take than resting state studies in depressive illness." [124]. One such paradigm involves pictures of faces showing sad or happy expressions to induce mood changes. When comparing activation patterns during happy and sad stimuli directly, greater activation has been demonstrated in the ventrolateral prefrontal cortex, the anterior cingulate cortex, the transverse temporal gyrus, and the superior temporal gyrus in healthy subjects [130] and in the prefrontal and anterior cingulate cortices in depressed volunteers [123, 131]. The intensity of self-rated sadness is reported to correlate with left sided amygdala activation [132]. Brain imaging methods are highly promising as potential response predictors, however, as with sleep EEG techniques, they are very complex, expensive and not readily available to the average clinician.

In this study, the above works have been used as a potential guide to neuroanatomical regions of interest with respect to predicting response to antidepressant treatment. However, we endeavored to develop a method that employs the inexpensive and readily available EEG. Our strategy was to fully exploit any neuropsychiatric salience that might be embedded in the EEG signal by using high performance machine learning methods of analyzing pretreatment EEG to predict response to SSRI treatment in subjects with MDD.

4.2.2 Experimental Details

Twenty two subjects (9 males, 13 females, age 20.6 to 62.6, mean 48.9 years) diagnosed with MDD using the internationally recognized Diagnostic and Statistical Manual - IV diagnostic criteria were treated with a 6 week course of an SSRI (particularly, the drug Sertraline hydrochloride, with the trade name Zoloft). Training data consisting of each subject's pre-treatment EEG and their ultimate response to the therapy were collected.

Often in studies using EEG, it is necessary to distinguish between the EO and EC cases. Most studies involving EEG use only EC data to avoid artifact. However, at least one study found greater pre-treatment alpha power asymmetry in responders to an SSRI vs non-responders for EO EEG [101]. In this study, it was also found that even though the selected features corresponding to the EO, EC, and EO– and EC–combined cases were different, the overall final performance did not vary significantly. Therefore in my study, combined EO and EC EEG measurements are used in the following experiments to make maximal use of the available data.

The definition of a responder to the SSRI medication in this case was taken to be at least a 25% improvement between the pre– and post–treatment Hamilton depression rating scales. This is a 17 item clinical procedure undertaken by psychiatric interview with the patient and yields a quantitative indication of the severity of depression. Although "response" is often defined as at least 50% improvement of depression rating scales, a recent review of the threshold for clinically significant improvement [133] concluded that this value is overly conservative as subjects considered to be "improved" by their own clinician showed only 23 to 42% improvement in scores using standardized symptom rating scales, such as the Hamilton Depression rating scale. Furthermore a 25% improvement in depression rating scale scores in the first few weeks of treatment has been shown to be indicative of more extensive improvement several weeks later [134]. Finally, only 50% of those patients who go on to reach full remission do so in the first 6 weeks [135]. For these reasons it is believed that an improvement of 25% in Hamilton depression rating scale scores is clinically significant.

In the SSRI experiments, the novel regularized feature selection method of 'maxKLD' which is based on maximum KL distance, as explained by Eq. (2.6) is used.

4.2.3 Results

Based on the EEG recording procedure described in Sect. 4.1.1 we nominally have 12 epochs for each subject. However, in the SSRI study, for one subject, only 10 epochs are available. Therefore, the total number of available epochs is $M_t = (12$ epochs/subject $\times 22$ subjects $-2) = 262$. See Section 4.1 for details of the data analysis process.

The performance of the proposed methodology for prediction of response to SSRI medication is documented in Table 4.6 for the case $N_r = 8$, where N_r is the number of discriminating features selected from N_c candidate features, as described in Section 2.2 and in Section 4.1. An L2O nested cross-validation procedure was used. As explained in Section 2.4.3, in the SSRI experiment, using L2O is equivalent to an 11-fold cross-validation procedure, as recommended by [87]. As may be seen, the performance is $\mu_H(\hat{\boldsymbol{\theta}}) = 86.6\%$ prediction rate (which corresponds to $\mu_B(\hat{\boldsymbol{\theta}}) = 0.165$, and $\mu_{\mathrm{KL}}(\hat{\boldsymbol{\theta}}) = 0.866$). The SVR and KPLSR methods yielded approximately equal performance. The number of major latent vectors which worked best for the KPLSR method is three. The parameter values used to calculate the performance measures in Sect. 2.4.2 are $C_{1|2} = 1.5, C_{2|1} = 1, P_1 = P_2 = 0.5$. It was found that the kernel-based prediction methods employed in this thesis are not very sensitive to the design parameter N_r, in that the performance results generated for $N_r = 5, 10, 12, 14, 16, 20$ were not significantly different compared to the $N_r = 8$ case, which means that they can tolerate some redundancy in the input data. Also, this means that there are several sets of discriminating features that have similar predictive power. The discriminating power of selected features is confirmed by observing that the "nearest-neighbor" (NN) classification method (see e.g. [4]) which is the traditional classifier that has no design parameter, results in an average performance reflected by $\mu_H(\hat{\boldsymbol{\theta}}) = 83.04\%$ (specificity=$P_{1|1}$=0.786, sensitivity=$P_{2|2}$=0.875), $\mu_B(\hat{\boldsymbol{\theta}}) = 0.2$ and $\mu_{\mathrm{KL}}(\hat{\boldsymbol{\theta}}) = 0.829$.

We also tested with L1O, L3O and L5O cross-validation procedures (instead of L2O) and found no significant change in the average prediction performance. This shows that the overall performance is not strongly dependent on the value of n in the LnO testing procedure, provided n is less than approximately 7, considering the fact that data for only 22 subjects are available in the SSRI study. Thus, we see that performance is not strongly dependent on the training set, provided the relative number of training samples is not too small. For $n \geq 7$ in LnO, it is observed that the average performance drops, as expected.

EEG recordings were also taken 7 days after onset of treatment for 21 out of the 22 subjects used in the study. As a third test of performance, the proposed prediction model was trained using all pre-treatment data from the 22 subjects using $N_r = 8$. This prediction model was then tested using all available post–7–days data. We observed that the average response prediction performance in this case is 88.5% (specificity=0.769, sensitivity = 1), as reflected in Table 4.7.

Table 4.6: The contingency table from the L2O nested cross-validation procedure, for predicting response to SSRI therapy for subjects with MDD. $N_r = 8$.

	predicted NR	predicted R	% correct
actual NR	12	2	85.7%
actual R	1	7	87.5%

Table 4.7: Contingency table obtained by testing the treatment-response predictor on the post–7–days EEG data, when the predictor is trained using all available pre-treatment data. $N_r = 8$.

	predicted NR	predicted R	% correct
actual NR	10	3	76.9%
actual R	0	8	100%

Even though background EEG signals are nonstationary random processes, these results suggest that the underlying statistical behaviour of the EEG that allows us to discriminate responders from non-responders is persistent for the significant duration of about 7 days. Thus, there is evidence that the proposed prediction method will give persistent results, i.e., results that will not vary over an extended period of time. This also indicates that 1 week of antidepressant medications do not appear to alter those aspects of the EEG signal most relevant to prediction of treatment response.

Finally for this post–7–days test data, it is interesting to observe that the 'linear discriminant analysis' (LDA) which is a standard classification technique, and has no design parameter to estimate, (see e.g., [4]), provides a prediction performance of 82.2% (specificity=0.769, sensitivity = 0.875). Compare this to Table 4.7. The LDA classifier incorporates a linear boundary, and no kernelization procedure was used in this experiment. The result is a very simple mathematical structure with no design parameter, for discrimination of the classes. The satisfactory performance level of 82.2% indicates that the clusters are well separated in the feature space. The experiment of Table 4.7 used a Gaussian kernel (with one design parameter) and a PLSR classifier (also with one design parameter), resulting in a mathematical structure with only slightly higher complexity than the LDA case. Thus, with well-separated clusters, a simple mathematical structure, and a relatively large number of training samples (in comparison to the number of parameters), the performance obtained in Table 4.7 is very unlikely to be a consequence of over-fitting.

We now show how low-dimensional representations based on nonlinear principal components, as discussed in Sect. 2.3.7, can be used to visualize the clustering behaviour associated with this problem. Fig. 4.1 shows a scatter plot of the $M_t = 262$ available training samples projected onto only the first two major nonlinear principal components. This figure was generated using the KPCA method with a Gaussian kernel using $N_r = 8$ common discriminating features among folds of a L1O procedure when applied to all 22 subjects. The patient index is written beside each data sample. Averaging the locations of the projected data samples belonging to each subject results in Fig. 4.2, in which each subject is shown with only a single point. These two figures show a noticeable clustering of the subjects into two classes, although the clustering is not perfect. This example further illustrates the fundamental integrity of the proposed prediction method, and shows that it is possible to select a set of reduced features from the background EEG which are indicative of response.

4.2.4 A List of Discriminating Features for SSRI Therapy

A list of the most relevant discriminating features is shown in Table 4.8. The features are sorted based on the frequency (in Hz) of the corresponding statistic. Columns 3 and 4 reflect the means and standard deviations of non-responder (NR) and responder (R) groups. These values however depend on the pre–processing and feature extraction procedure. All features in this table are intra-hemispheric coherence, mutual information and anterior to posterior log power ratios (or front-to-back PSD ratio denoted by 'F/B' in the table). A feature is listed in this table if it is used at least once throughout the LnO procedure.

The average value of all the intra-hemispheric coherence and mutual information features listed in Table 4.8 are lower for responder subjects than for non-responders. As well, the average values of the anterior to posterior log PSD ratios in this table (features 14 and 16) are close to zero for responders, indicating a more uniform distribution of 14 Hz EEG power levels. Non-responders on the other hand, had lower anterior to posterior EEG power ratios. Finally it should be noted that all identified frequencies are in the alpha or low beta bands.

From the statistics shown in Table 4.8, we see that the individual features are not significantly different between the R and NR groups, and thus single features on their own are incapable of discrimination. An important point is that the clusters corresponding to the prediction categories become distinct only when the features are *jointly* mapped into the N_r–dimension

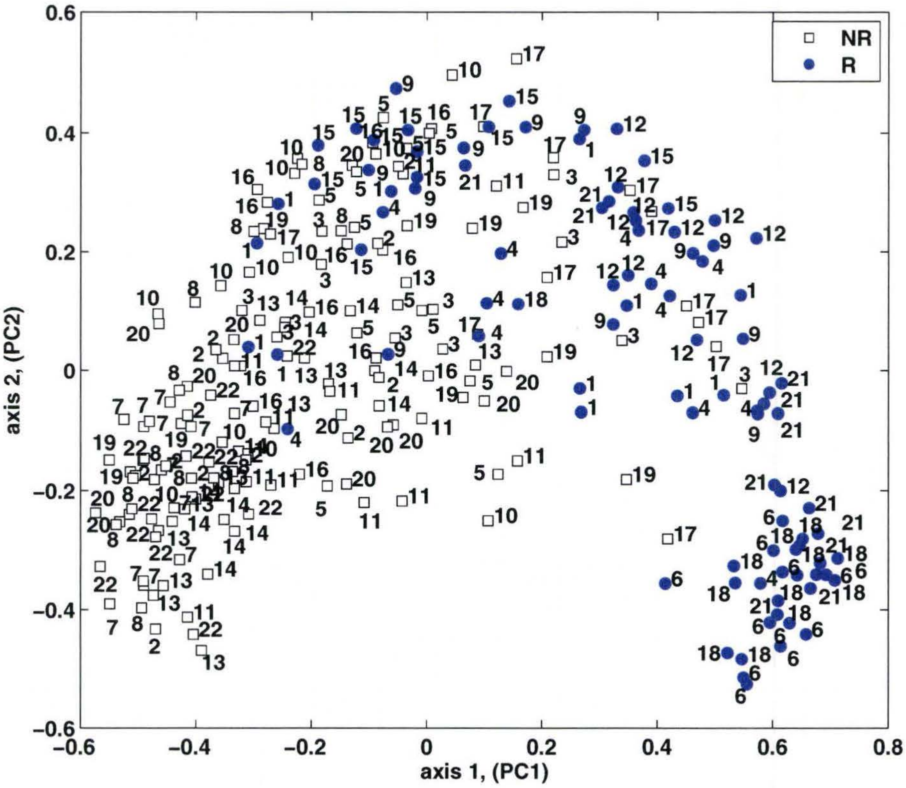

Figure 4.1: Scatter plot of the projection of the $N_r = 8$–dimensional feature vectors from all $M_t = 262$ available training epochs (approx. 12 epochs per subject) onto the first 2 major principal components, which are obtained using the KPCA method with a Gaussian kernel. The numbers identify epochs belonging to each subject.

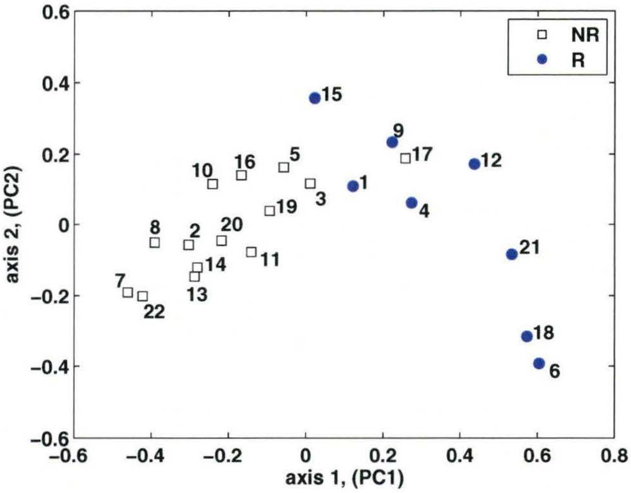

Figure 4.2: Same as Fig. 4.1, except that all points corresponding to each subject have been averaged. The clustering behaviour between the R and NR groups is clearly evident.

feature space.

An additional interpretation of the features is given in Fig. 4.3 which presents a graphical depiction of the most-relevant features listed in Table 4.8. A connection between two electrode sites in the figure corresponds to a selected feature which involves those two locations. Connections are shown by solid thick lines. This roughly indicates relations between EEG sensors that convey relevant information for our response-prediction task.

4.2.5 Using a Smaller Number of EEG Electrodes

Based on the EEG electrodes selected in the list of relevant features in Table 4.8, in another experiment, pre-treatment measurements from only 7 EEG electrodes consisting of Fp1, F3, F7, C3, T3, P3, T5, were used for the prediction study. The data from the remaining electrodes were discarded. These electrodes are all on the left side of the scalp. In this experiment, the set of $N_c = 950$ EEG-driven candidate features extracted from each epoch include the magnitude-squared spectral coherence between all electrode pairs as a function of frequency (from 4Hz to 20Hz with 1Hz resolution).

With the above design, the treatment-response prediction performance became approximately 82%. This indicates that this cheaper and more readily

Table 4.8: A list of the most discriminating features, showing the mean and standard deviation of each feature (after z-score normalization) over the non-responder (μ_N, σ_N) and responder groups (μ_R, σ_R) to SSRI therapy. The EEG channel designations shown are in accordance with the standard 10–20 system.

#	*Selected Feature*	μ_N, ($\pm$ σ_N)	μ_R, ($\pm$ σ_R)
1	Mutual Information between T3 & P3	0.769 (± 0.874)	-0.066 (± 0.637)
2	Mutual Information between T3 & T5	0.79 (± 0.901)	-0.306 (± 0.824)
3	Mutual Information between F4 & T4	0.93 (± 1.091)	0.004 (± 0.736)
4	Coherence at f=9Hz between T3 & T5	0.564 (± 0.867)	-0.566 (± 0.99)
5	Coherence at f=10Hz between T3 & T5	0.59 (± 0.973)	-0.688 (± 0.959)
6	Coherence at f=10Hz between T3 & P3	0.488 (± 0.888)	-0.456 (± 0.688)
7	Coherence at f=10Hz between C3 & T5	-0.241 (± 0.972)	-0.96 (± 0.796)
8	Coherence at f=10Hz between F7 & C3	0.041 (± 0.969)	-0.624 (± 0.932)
9	Coherence at f=11Hz between T3 & T5	0.773 (± 1.035)	-0.575 (± 0.99)
10	Coherence at f=12Hz between T3 & T5	0.846 (± 0.921)	-0.501 (± 1.019)
11	Coherence at f=12Hz between T3 & O1	0.82 (± 1.272)	-0.28 (± 0.739)
12	Coherence at f=13Hz between T3 & T5	0.826 (± 0.865)	-0.468 (± 1.011)
13	Coherence at f=14Hz between T3 & T5	0.775 (± 0.823)	-0.497 (± 0.98)
14	F/B PSD-ratio at f=14Hz, Fp1/F3	-0.24 (± 0.669)	0.575 (± 0.738)
15	Coherence at f=14Hz between C3 & T5	-0.0618(± 1.021)	-1.203 (± 0.86)
16	F/B PSD-ratio at f=14Hz, Fp1/C3	-0.709 (± 0.911)	0.369 (± 0.85)
17	Coherence at f=15Hz between T5 & P3	-0.472 (± 0.887)	-1.726 (± 1.072)
18	Coherence at f=15Hz between T3 & T5	0.746 (± 0.809)	-0.476 (± 0.921)
19	Coherence at f=15Hz between C3 & T5	-0.21 (± 0.875)	-1.243 (± 0.786)
20	Coherence at f=16Hz between T3 & T5	0.717 (± 0.831)	-0.373 (± 0.998)

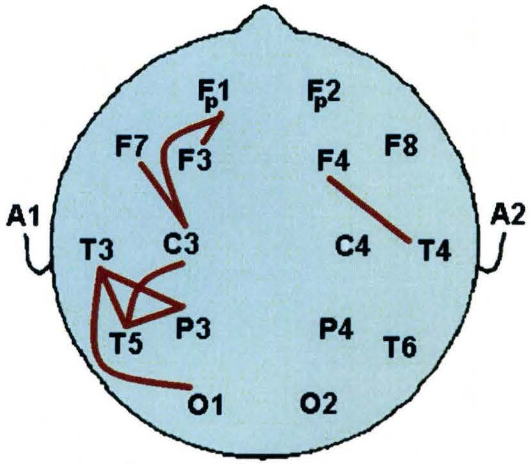

Figure 4.3: In SSRI case, a rough schematic drawing showing the most relevant features by connections between corresponding EEG electrodes, as reflected in Table 4.8. Electrodes A1 and A2 denote the linked ear reference.

applied configuration of electrodes may suffice when this proposed method is adapted for use in the clinic. It is to be noted that we experimented with several other configurations of EEG electrodes (e.g., some with 3 electrodes only), but the results are not discussed here to save space.

4.2.6 A Comparison of Feature Selection Methods

We now discuss a comparison of the following two feature selection methods described in Section 2.2:

- FS1. The novel 'maxKLD' feature selection method, as described by Eq. (2.6) using the weighted Gaussian *pdf* assumption of Eq. (2.9).
- FS2. The Feature selection method of Peng. et al. [68], which is based on mutual information, as described by Eq. (2.7).

The performance index used in the comparison was the average correct prediction performance, $\mu_H(\Theta)$, as defined by Eq. (2.63), measuring the final efficacy of the selected features in each method using a cross-validation procedure.

One advantage of the proposed 'maxKLD' criterion is that it admit a simplification based on a Gaussian approximation of the feature variables, as in (2.8) and (2.9). Note however, that the method of [68] is inherently not amenable to a closed-form Gaussian approximation for mutual information, due to the fact that the target variable y is discrete with only two values. Unfortunately however, the computation of the mutual information quantities is very fast when the y-variables is binary. This fact counter-acts the potential computational advantage of the Gaussian approximation of the maxKLD method. The net result is that the two methods require approximately the same execution time in this application.

Furthermore, the overall prediction performance of the two methods was evaluated for the SSRI case. The best performance index values were 86.6% for both methods. Thus, we observe there is no apparent advantage in performance of the proposed method. Nevertheless, an exposition of the method is included in this thesis, due to the scientific interest it may generate.

4.2.7 Discussion

Our findings are consistent with the results of Cook et al [96], who found that absolute and relative power, as well as cordance, in all four EEG bands recorded pre-treatment over the regions implicated in mood disorders; i.e., prefrontal Fpl-Fp2-Fpz FCl-FC2-Cz, left temporal T3-T5; and right temporal T4-T6 are not significant predictors of response.

Several research groups have only used alpha power, considering an increase indicative of less cognitive neural activity, and that depressed patients had greater inter-hemispheric alpha asymmetries. They only considered inter-hemispheric asymmetries and found significantly greater overall pre-treatment alpha asymmetry with the right hemisphere more active in non-responders to the SSRI fluoxetine than responders [101] and responders had significantly greater alpha in the occipital regions than non-responders and controls, and also greater right over left alpha asymmetry with non-responders having the opposite asymmetry [104].

Considering Table 4.8, the list of most discriminating features are in the alpha or low beta frequency band. However, unlike other studies, e.g., [104], inter-hemispheric power asymmetries are not among the most discriminating features found by our method. Instead, our results show that responders have more uniform alpha and low beta power anterior to posterior in the left hemisphere than non-responders who showed relatively greater posterior power.

Fourteen of the 20 features identified by our method are coherence between the EEG signals from two left hemispheric electrodes for alpha and very low beta frequencies. Coherence has not generally been a common feature selected for EEG analysis, especially in the field of psychiatric disorders. However, it has been used to discriminate subjects afflicted with MDD from normals

[98, 99], and in the study of the effect of clozapine therapy in schizophrenia [102].

As coherence and mutual information between several electrode pairs appear to be among the most highly predictive features (especially in the T3–T5 region), we might speculate that neural interaction or connectivity between the regions corresponding to the respective channels is highly relevant to SSRI response. In our study, responders had less coherence in the identified regions than non-responders, indicating less synchronism in the left prefrontal and temporal areas.

4.3 rTMS Therapy for MDD

Here, repetitive transcranial magnetic stimulation (rTMS, or TMS) therapy for MDD is investigated. rTMS therapy, approved in Canada and the USA for use in patients with MDD, employs strong pulsed magnetic fields administered through a magnetic coil placed near the head of the subject, to induce electrical currents in the brain to change the activity of neuron populations. Typically, a period of 2 to 6 weeks of application of rTMS is required before the patient experiences relief from depression. rTMS therapy can be considered as a non-invasive version of electroconvulsive therapy (ECT). As with ECT, rTMS is typically reserved for use when antidepressant medications prove ineffective. For an overview of rTMS therapy, its effectiveness and its various clinical applications, see, e.g. [136–142]. In general only 40% to 50% of depressed persons treated with rTMS respond to treatment. A means of determining, in advance, whether rTMS will be effective would be of great value.

The objective of this portion of the thesis is to evaluate the efficacy of an automatic data analysis procedure/model using pre-treatment EEG data, in which advanced machine learning methods are employed to predict whether or not an individual subject will be responsive or non-responsive to the rTMS therapy.

4.3.1 Clinical Details of the rTMS Therapy

In our study, a 2-week course of rTMS therapy was administered over the left[4] or right dorsolateral prefrontal cortex, or both. Two different TMS magnetic stimulator devices are used: (i) Dantec Magpro and (ii) Magstim

[4] The location is 5 cm anterior to the motor spot that elicited motor-evoked potentials in either the abductor pollicis brevis or the first interosseous dorsalis muscle of the right hand.

Superrapid. The magnetic power of the coils is up to 2 Tesla. The subjects are randomly assigned to receive any of the following:

1. Right low frequency TMS (plus left sham high frequency TMS)
2. Left high frequency TMS (plus right sham low frequency TMS)
3. Right low frequency AND left high frequency TMS

In all these, the high frequency value was 10Hz and the low frequency value was 1 Hz.

- Sham TMS: coil held with one wing touching the scalp and held at 90 degrees to a tangent at the scalp site. The device output setting for Sham TMS intensity was set at 30 units.
- High Frequency TMS: 20 trains at 10 Hz at 110% of motor threshold, train duration of 8 seconds, inter-train interval of 52 seconds using a figure 8 coil (total 16,000 pulses).
- Low frequency TMS: = 2 trains at 1 Hz, train duration of 60 seconds, inter-train interval of 3 minutes using a round coil (total 120 pulses). LF was administered immediately after HF TMS was completed.

On the day of the first rTMS treatment, subjects received either 50 mg of sertraline or 10 mg of citalopram (choice based on previous treatment history). The dose was doubled after 1 week. Further dose increments were made each week if the HamD improvement was less than 25% of the previous week's score (sertraline: 50 mg increments to a maximum of 200 mg/day; citalopram: 10 mg increments to a maximum of 60 mg/day).

After 2 weeks of active rTMS therapy, as outlined above, all subjects continued to receive the medication (sertraline or citalopram, as described above) for 4 more weeks. Then at the end of 6 weeks of treatment, the clinical response (to be described later) is measured and is compared to the pre-treatment condition.

4.3.2 A Review of Previous Studies

There are several studies which investigated using various clinical rating attributes and biographical parameters as predictors of treatment-response in rTMS therapy. Padberg et al. [143] found that improvement of depression after partial sleep deprivation was inversely correlated with improvement after rTMS therapy. However this data analysis is done within 2 weeks with 10 sessions of rTMS therapy administered 5 days after sleep deprivation. Fitzgerald et al. [137] employed linear regression models and used the improvement in Montgomery-Asberg depression rating scale (MADRS) as the treatment response variable. They analyzed the baseline clinical and demographic data of 40 subjects receiving rTMS and found that a greater degree of agitation in a baseline rating of psychomotor disturbance [144] was associated with better response. Fregni et al. [145] studied the relevance of demographic, depression and treatment characteristics, psychiatric and drug history to predict the response to rTMS applied on left dorsolateral prefrontal cortex. They concluded that rTMS results in better outcome in younger and less treatment-resistant patients. Brakemeier et al. [146] found that a high level of sleep disturbances was a significant response predictor. Also, a low score of treatment resistance and a short duration of episode were positive predictors. As in the above two studies (i.e., [145, 146]), most patients also received concomitant antidepressant medication, the response patterns may rather refer to combined treatment than rTMS alone. Brakemeier et al. [147] stressed this point and studied the clinical predictors in 79 drug-free subjects and found that the two formerly mentioned studies above (i.e., [145, 146]) could not be validated. However they found a confirmative result that a high level of therapy resistance (prior to starting the therapy) is associated with poor outcome. Lisanby et al. [148] studied the clinical predictors of acute outcome of active or sham TMS, in which changes in MADRS after 4 weeks was used as the response variable. They found that the number of prior treatment failures was the strongest predictor for positive response to acute treatment with TMS. They also claimed that shorter duration of current illness and lack of anxiety comorbidity may also correspond to an increased likelihood of response.

Langguth et al. [149] studied baseline alterations of regional cerebral blood flow measured by single photon emission computed tomography (SPECT) imaging as well as baseline clinical characteristics as predictors of response to rTMS. In a data population of 24 subjects and using a multivariate regression model, they found that high pretreatment anterior cingulate activity and low treatment-resistance to pharmacologic therapy were positive predictors. Schiffer et al. [150] studied an alternative technique in which baseline lateral visual field stimulation is used to predict the clinical outcome of 10-day course of rTMS, and found that there was a significant correlation between the percent improvement in response to rTMS and their lateralized affective responses to lateral visual field stimulation.

The short-term (less than a few minutes, or immediate) effect of rTMS on EEG signals have been studied in a few papers including [151–155]. rTMS appears to alter intrahemispheric directed coherence, and increase alpha and delta frequency power. In this thesis we examined pretreatment EEG as a predictor of the long-term response to a standard course (6-weeks) of rTMS therapy to treat MDD in treatment resistant subjects.

As discussed in Sect. 4.2, there are significant research activities to employ resting EEG as a method for predicting treatment outcome in depressed subjects, however, there are only a few studies for prediction of response to rTMS therapy using EEG data, to be discussed as follows.

In our earlier work examining pre-treatment EEG spectra, a superior response to rTMS in subjects who showed frontal alpha power asymmetry with relatively greater alpha power in the right compared to left frontal cortex was observed. Price et al. [156] studied correlations between EEG features including individual alpha power, alpha frequency as well as asymmetry indexes (using 9 EEG electrodes including F3, F4, F7, F8, T3, T4, Fz, Cz, Pz) and clinical response in 39 subjects with treatment-resistant depression, and found that there is weak evidence of predicted correlation between these features and clinical rating change. They concluded that the use of these features for clinical assessment is not supported by their results. Funk and George [157] describe a recent multisite study in which EEG data from 2 electrodes (Fpl and F3) will be measured in 240 subjects with MDD, with the aim of investigating whether or not the changes in prefrontal EEG power or asymmetry are associated with response to rTMS. However, in this publication, data on only 4 subjects were presented as a case series and no definitive conclusions are possible. Spronk et al. [158] studied the rTMS effect on 8 subjects and found that pre-treatment QEEG did not show treatment specific effects.

4.3.3 Participants

Subjects were recruited into an rTMS study approved by the Research Ethics Board of St. Joseph's Health Care, Hamilton, ON, Canada. In this study, 41 subjects diagnosed with unipolar MDD were treated with active/true rTMS. The exclusion criterion included failure to at least two courses of antidepressant. The other exclusion criteria for the rTMS study (as well as for the SSRI study discussed in Section 4.2) were: active suicidality, substance abuse, uncontrolled medical illness, personal or 1st degree family history of epilepsy, cardiac pacemaker, or any intracranial metal object. All subjects also received concurrent treatment with SSRI antidepressant medications (mostly 'Sertraline hydrochloride', with the trade name Zoloft). All subjects gave informed consent. Available socio-demographic and clinical information for participants

Table 4.9: Demographic information of 41 subjects with major depressive disorder who participated in the study. See text for definition of items marked by *.

Information	Range
Age at start of treatment [years]	Average=45.9, std=10.3, min=20.3, max=65.8
Gender	Female: 28 (68.3%), Male: 13 (31.7%)
Handedness*	Average= 0.768, std=0.55, min= -1, max=1
Pre-treatment HamD score*	Average=21.3, std=3.7, min=15, max=29
rTMS type administered*:	
True left, Sham right	20 subjects
True left, True right	9 subjects
Sham left, True right	12 subjects

are shown in Table 4.9. HamD denotes the 17-item Hamilton depression rating score. Left rTMS (true or sham) include high frequency pulses,but right rTMS (true or sham) include low frequency pulses administered to dorsolateral prefrontal cortex. Handedness ranking was as follows. 1:Right-handed (30 subjects), -1: Left-handed (3 subjects), 0.75: Mostly right-handed, but some left (6 subjects), 0.25: some right, some left, but right is used more than left (1 subjects), -0.25: some left, some right, but left-hand is used more than right (1 subjects).

4.3.4 Definition of Response

Two response criteria are considered. In the first criterion, subjects were classified as "responders" if at least a $\beta = 25\%$ improvement in their 'Hamilton depression rating' (HamD) score after a period of six weeks of treatment is observed. This is a clinical procedure undertaken by psychiatric interview with the patient and yields a quantitative indication of the severity of depression. See Section 4.2.2 about relevance of a 25% improvement in six weeks. In the second criterion, a $\beta = 50\%$ or greater improvement in HamD is used as the response criterion. Therefore, for our purposes, the HamD percentage change value is discretized into two values (or classes), corresponding to responder (R) when it is larger than or equal to β, and non-responder (NR) otherwise. A non-responder is denoted by response indicator $y = 1$, and responder by $y = 2$.

4.3.5 Analysis of Prediction Performance

In the rTMS experiment, the training set consists of $M_p = 12$ EEG epochs from each of $N_p = 41$ subjects, for a total of $M_t = 492$ epochs altogether. See

Table 4.10: Performance results for predicting the response to rTMS therapy for subjects with MDD, with an L4O test procedure (which is equivalent to 11-fold cross-validation) and when $\beta = 25\%$ response threshold is used. Both EO & EC data are used.

	predicted NR	predicted R	% correct
actual NR	17	5	77.27% = specificity
actual R	3	16	84.21% = sensitivity
			average = 80.74%

Table 4.11: rTMS therapy: Prediction performance results when at least $\beta = 50\%$ improvement in pre-post Ham-D score is used as the response criterion. Both EO & EC data are used.

	predicted NR	predicted R	% correct
actual NR	25	5	83.33% = specificity
actual R	2	9	81.82% = sensitivity
			average = 82.58%

Section 4.1 for data analysis process. The regularized feature selection method in Eq. (2.7) is used.

Table 4.10 shows the prediction performance for rTMS therapy for the 41 subjects available for this study (consisting of 22 NR and 19 R subjects) when using all available EO and EC EEG data, and $\beta = 25\%$ (as the response threshold) is used. The average correct prediction rate is 80.74%. The L4O cross-validation procedure with $N_r = 8$ discriminating features are used. Note that with 41 subjects, this is equivalent to an 11-fold cross-validation. This is the result obtained by the best SVR model; however, the results obtained by the MFA, RLS and kernel PLSR methods are not significantly different.

If we use at least a $\beta = 50\%$ improvement instead of a $\beta = 25\%$ improvement in the HamD rating score as the response criterion, we will have 11 R and 30 NR subjects in our data set. For this scenario, with an L4O cross-validation test, the average performance rate as shown in Table 4.11 is 82.6% using the MFA prediction model with two mixtures and 1 factor (i.e., the best model parameters are found to be $K = 2$, $m = 1$ as defined in the MFA model in Section 2.3.6). $N_r = 8$ discriminating features are used in each step of the L4O test.

It is noticed that the regularized 'maximum relevance and minimum redundancy' feature selection criterion by Peng et al. [68] based on mutual information, gave approximately similar performance as compared to our 'maxKL' feature selection method where KL distance is employed in a regularization framework (as described in Section 2.2.3).

Further, the use of 'only-EO' data[5] or 'only-EC' data is studied in alternative experiments. It is found that the average prediction performance is approximately similar to the previous case where all EO and EC data were used together. Note that in theory, it is expected to have a lower level of variations in statistical properties of the measured data in each of these alternative experiment when only one type of EEG data is used (i.e., either EO or EC, not both). However, on the other hand, the high performance of the combined EO & EC case shows that both EC and EO data convey some common discriminating statistical information. It is also observed that the list of N_r discriminating features are different in each experiment, showing that each of the EO and EC conditions conveys some possibly different kind of discriminating information. Note that for example, for the EO case, the number of available data samples (M_t) is smaller, which is a disadvantage from machine learning point of view; however the data has lower variance which compensates the lower data size. For example, using $N_r = 8$ and employing the 'nearest–neighbor (NN)' classification method (which has no design parameter), the average prediction performance was 78.47% when only EO data (corresponding to $M_t = 246$ in our study) is used in both the training and test phases. Compare this figure with the one shown in Table 4.10 for the EO&EC case.

With regard to data clustering performance, Fig. 4.4 shows a subject-wise scatter plot of pre-treatment data of the 41 subjects who received rTMS therapy, projected onto only the first and second major nonlinear principal components. The objective in this figure is to check the clustering behaviour of the two classes (R versus NR group of subjects) for this 2-dimensional representation, which is useful only for the purpose of visualization. Again, the Max1Mag normalization and the KPCA method with a Gaussian kernel are used to generate the figure. Each subject is shown with only one point which is the average location of projected pre-treatment data samples/epochs belonging to the subject. The subjects are given an arbitrary ID number before starting the experiment (blind to treatment response), and this number corresponding to each individual is written beside each point in the figure. R subjects are shown with blue circles and NR subjects with red squares. The clustering result in this figure further justifies the discriminating power of simple quantitative features extracted from pre-treatment EEG data.

When a $\beta = 50\%$ response threshold is used, employing all EO & EC EEG

[5] Using 'only-EO' data means using only EO EEG files for training the predictor model, and then the model is tested on only the EO data available for each subject. This means that no EC data is involved in training or test.

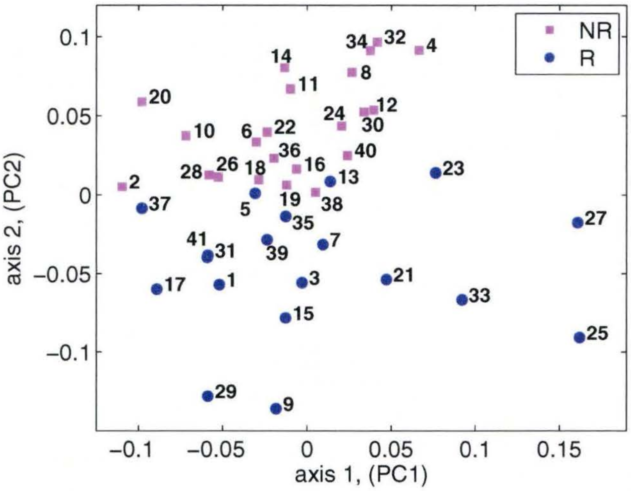

Figure 4.4: For rTMS therapy, subject-wise scatter plot of projected pre-treatment EEG data, with $\beta = 25\%$ improvement as the response threshold. This shows a clustering of responsive (R) and non-responsive (NR) group of subjects.

data, the clustering performance by the KPCA method is shown in Fig. 4.5 with $N_r = 14$ features. Compare this figure to Fig. 4.4. The two clusters associated with the R and NR group of subjects in Fig. 4.4 are more closely–spaced (compared to the clusters in Fig. 4.5), which confirms the slightly lower prediction performance obtained in the $\beta = 25\%$ case versus the $\beta = 50\%$ case. Note that the list of discriminating features is different compared to the case when a $\beta = 25\%$ response threshold is used. There are a few overlapping points which correspond to a mis-classification (or mis-prediction), as reflected in Table 4.11.

Comparing the prediction performance between rTMS and SSRI therapies, it can be seen that the treatment-efficacy prediction performance is lower for rTMS therapy compared to SSRI therapy. The reason might be the fact that there are several factors that could affect the efficacy of rTMS therapy. One of them is the exact scalp area where the TMS magnetic stimulations are applied in each subject. Variations in head size (relative to normal) could result in an inefficient TMS location, which might reduce rTMS efficacy and complicate the treatment process. The other factor to make the problem more complex is that some of the subjects got TMS treatments over the left dorsolateral prefrontal cortex, and some were treated over right side, or both sides. Subjects may

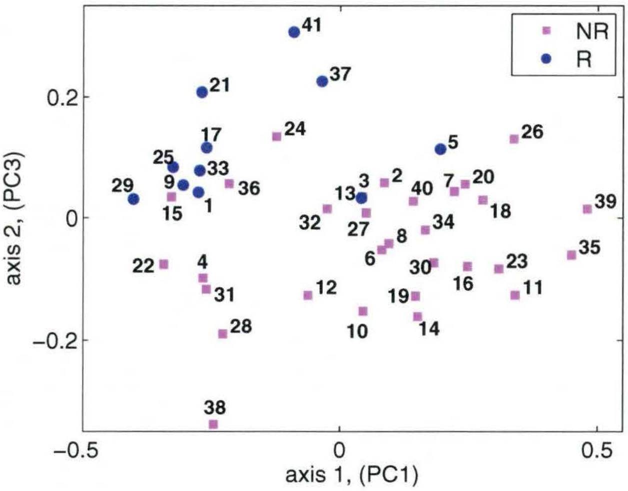

Figure 4.5: For rTMS therapy, subject-wise scatter plot of projected pre-treatment data, when $\beta = 50\%$ improvement threshold is used.

respond differentially to these different types of rTMS or these types of rTMS may not have equivalent efficacy. Therefore, the rTMS study needs to be investigated more carefully.

4.3.6 A List of Discriminating Features

A list of discriminating features for prediction of response to rTMS therapy when the response is considered to be at least a $\beta = 50\%$ improvement based on pre-versus-post HamD change is shown in Table 4.12. The features are sorted based on the frequency in Hz associated with the corresponding statistic. The frequency values shown have approximately 1Hz resolution, and therefore one might consider the neighboring frequencies as well, expecting only a slight change in performance. Similarly, since our spatial resolution is limited, one might consider similar features derived from neighboring EEG electrodes, (particularly if one is experimenting with an EEG system that has a higher number of electrodes than the standard 10–20 system that was used here). The list of relevant features derived using the regularized feature selection method depends on various factors including the EEG feature extraction and pre-processing procedure. The words 'F/B' and 'R/L' stand for the 'front-to-back' and 'right-to-left' power spectral density (PSD) ratios (calculated as log difference of PSD values), respectively. Note that some of the features

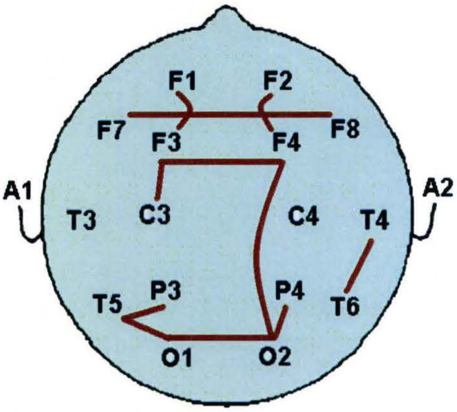

Figure 4.6: In rTMS case and for $\beta = 50\%$ response threshold: A rough schematic drawing which shows a list of some relevant features by connections, as reflected in Table 4.12.

listed in Table 4.12 are related and therefore the list has some redundancy.

It is found that no individual feature listed in Table 4.12 has significant discriminating ability on its own. It is the joint information in the collection of all N_r features (i.e., in the form of multivariate or multi-dimensional feature vector of size N_r) that results in the performance numbers listed in Table 4.11. Note that as an extension to this work, one might use a combination of the simple features we used to obtain improved discrimination performance.

As an alternative explanation of the features, Fig. 4.6 presents a graphical depiction of the most-relevant features listed in Table 4.12. A connection between two electrode sites in the figure corresponds to a selected feature which involves those two locations. Connections are shown by solid thick lines. This roughly indicates relations between EEG sensors that convey relevant information for our response-prediction task. Like in the previous SSRI case, the features listed in this figure and Table 4.12 may give some clues about the locality and interconnection of neurological mechanisms associated with a positive response to rTMS therapy.

4.4 Clozapine Therapy for Schizophrenia

Compared with other antipsychotic medications, the atypical antipsychotic medication clozapine is recognized to have superior therapeutic effectiveness in

Table 4.12: A list of discriminating features/attributes for prediction of response to rTMS antidepressant therapy with $\beta = 50\%$ improvement response threshold and using all EO & EC EEG data.

#	*EEG-driven Numerical Feature*
1	Mutual Information between T5 & O1
2	Correlation between T5 & O1
3	Coherence at f=4Hz between F2 & F4
4	Coherence at f=5Hz between F2 & F4
5	Coherence at f=5Hz between F3 & C3
6	Coherence at f=6Hz between F3 & C3
7	Coherence at f=7Hz between F3 & C3
8	Coherence at f=7Hz between P4 & O2
9	Coherence at f=7Hz between F3 & F4
10	Coherence at f=8Hz between F3 & C3
11	Coherence at f=8Hz between P4 & O2
12	F/B PSD-ratio at f=8Hz, F4/O2
13	R/L PSD-ratio at f=10Hz, O2/O1
14	R/L PSD-ratio at f=11Hz, O2/O1
15	R/L PSD-ratio at f=12Hz, O2/O1
16	F/B PSD-ratio at f=12Hz, T4/T6
17	Coherence at f=16Hz between T5 & O1
18	R/L PSD-ratio at f=17Hz, O2/O1
19	R/L PSD-ratio at f=18Hz, O2/O1
20	Coherence at f=18Hz between T5 & P3
21	Coherence at f=18Hz between T5 & O1
22	Coherence at f=20Hz between T5 & O1
23	F/B PSD-ratio at f=23Hz, F1/F3
24	R/L PSD-ratio at f=28Hz, F8/F7
25	R/L PSD-ratio at f=36Hz, F8/F7

the treatment of chronic medication-resistant schizophrenia, e.g., [159]. However, clozapine may produce serious side effect such as seizures, cardiac arrhythmias or bone marrow suppression with neutropenia [39]. According to a recent Cochrane review, about 34% of treatment-resistant patients respond to clozapine while 3.2% develop blood problems [159]. As the hematological side effects can be life threatening, blood samples to monitor the white blood cell count must be collected as frequently as weekly. The logistic difficulties for the patient and the treatment team are substantial. A method that could reliably determine, before the onset of therapy, whether a given patient will or will not respond to clozapine would greatly assist the clinician in determining whether the risks and logistic complexity of clozapine are outweighed by the potential benefits.

QEEG or EEG may offer some promise in this regard. EEG abnormalities in schizophrenic subjects and EEG changes due to clozapine therapy have been the focus of a number of clinical studies, [43,44,102,160–169].

Based on findings in 17 schizophrenic subjects, Knott et al. [170] found that the clozapine-induced improvement of psychopathology symptom ratings using the Positive and Negative Syndrome Scale (PANSS) was correlated with pretreatment QEEG inter and intra-hemispheric spectral power asymmetry. Greater pretreatment anterior to posterior asymmetry in the delta frequency range was associated with greater improvement in negative symptoms while greater pretreatment anterior to posterior theta asymmetry predicted improvement of positive symptoms and global improvement. Larger inter-hemispheric asymmetry in the theta and beta frequencies in the central and anterior temporal regions were, respectively, predictive of greater improvement in positive and negative symptoms. Gross et al. [171] also found that changes in the theta frequency in QEEG with clozapine treatment, particularly in the middle electrodes over the fronto-central scalp area, were a more sensitive indicator for the evaluation of clozapine treatment efficacy than the serum clozapine level. Though these methods reveal important relationships between QEEG variables and clinical outcome, a series of simple correlational analyses do not readily yield a "responder" or "non-responder" dichotomous categorization for an individual patient.

The above analyses employed standard simple statistical methods. Machine learning techniques are finding increasing application in psychiatry, particularly when multi-dimensional, noisy, highly complex data or multi-modal data sets are analyzed together, see e.g., [172]. For example, the support vector machine (SVM) techniques to select spectro-temporal patterns from multichannel magnetoencephalogram (MEG) data collected during a verbal working memory task have been used to distinguish schizophrenic from control subjects [173]. Machine learning algorithms using structural brain magnetic resonance (MRI) images [174], functional MRI (fMRI) data [175, 176] and combined genomic and clinical data [177] have been employed to separate schizophrenic, bipolar and healthy control subjects.

Machine learning approaches to prediction of clozapine treatment-efficacy have also been employed. Lin et al. [178] describes a study in which a feed-forward multilayer perceptron network (with a back-propagation error training technique) is employed using clinical and pharmacogenetic data to predict clozapine response in schizophrenic subjects. Five pharmacogenetic variables and five clinical variables (including gender, age, height, baseline body weight, and baseline body mass index) were collated from 93 schizophrenic subjects taking clozapine, including 26 responders. Using this method, they obtained an overall prediction accuracy rate of 83.3%.

Guo et al. [175] describes a Bayesian hierarchical model using pretreatment fMRI and positron emission tomography (PET) information coupled with patient characteristics (e.g. medical or family history and genotype) as training data to predict changes in brain activity in 16 schizophrenic subjects following treatment with two atypical antipsychotics (risperidone or olanzapine). The authors postulated that predicting drug-induced changes in brain activity would assist the clinician in determining optimal drug choice.

However, the clinical utility of these mathematical approaches is negatively impacted by the expense and unavailability of complex techniques such as fMRI, PET, genetic screening and MEG. In contrast, electroencephalography (EEG) is an inexpensive, non-invasive technique widely available in smaller hospitals and in community laboratories. Therefore, predictive algorithms dependent on EEG measurements are more practical. Furthermore, since the required EEG data is obtained during the resting state, only minimal cooperation is required from the patient. Thus, an EEG based method of predicting treatment response would have many advantages over imaging methods such as MRI, PET or MEG.

The goal of the present study is to examine the utility of machine learning methods for processing EEG signals to predict the response of schizophrenic subjects to clozapine.

4.4.1 Description of Subjects and the Clinical Assessment Procedures

Subjects, comprising both in-patients and out-patients, were recruited from the Schizophrenia Program at St Joseph's Hospital, Centre for Mountain

Health Services, Hamilton, Ontario. All subjects met both *Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association, 4th Edition* (DSM-IV) [45] for schizophrenia and the Kane et al. [179] criteria for treatment resistance. Patients meeting these criteria may be considered to be "severely symptomatic", i.e., as suffering acutely from schizophrenia. All subjects gave informed consent.

Data from two groups of schizophrenic subjects were used in this retrospective study. The first group (Group A) consists of 23 subjects. Group B is an independent sample of 14 subjects. Available socio-demographic and clinical information for Groups A and B are shown in Tables 4.13 and 4.14. Symptom severity after clozapine treatment is measured in Group A using the positive and negative syndrome scale (PANSS) score [180]. As PANSS scores were not available for Group A subjects prior to clozapine treatment, pre-treatment symptom severity was assessed through a quantitative clinical assessment (QCA) conducted by review of the clinical record guided by the structure of the PANSS. The QCA procedure is outlined at the end of this subsection. All QCA raters were blind to the machine learning outcome predictions. QCA was used to assess psychopathology both pre and post clozapine treatment in Group B. PANSS evaluations are not available for Group B subjects. In Tables 4.13 and 4.14, the education level rating is done as follows: 1: grade 6 or less, 2: grade 7 to 12 without graduating, 3: graduated high school, 4: part college, 5: graduate 2 years college, 6: graduate 4 years college, 7: part graduated/professional school, 8: completed graduated professional school.

We now discuss how we determine whether a patient is a responder (R) or non-responder (NR). In this retrospective pilot study quantifying clinical response is complicated by the absence of pre-treatment PANSS scores. We were therefore obliged to define response on the basis of a single post treatment PANSS score. To do this we created post-treatment PANSS score[6] thresholds δ_1 to assess response: first we rank-ordered all subjects by post treatment PANSS score then chose a value of δ_1 (88.5) such that our 23 subjects were divided into responder (R) and non-responder (NR) classes with roughly equal number of subjects (R=12, NR=11).

Having R and NR groups of similar size has advantages with respect to the machine learning process; however, this assumes that clinically significant improvement is seen in about 50% of those treated with clozapine. Others have reported that, on average, only 34% of treatment-resistant schizophrenic

[6] Using the PANSS data, the 'total rank' (TR) score is used as the clinical assessment in our experiments. TR is the sum of three scales in PANSS: 1. general rank, (GR), 2. positive (or productive) symptoms scale, (PSS), 3. negative (or deficit) symptoms scale, (NSS). This means that TR=GR+PSS+NSS.

Table 4.13: Demographic information of the 23 subjects (denoted Group A) who participated in the clozapine study. The lower 4 items in the table are scales related to the PANSS clinical rating score. The item with * is described in the text. Avg denotes 'average', and std denotes 'standard deviation'.

Information	**Range**
Age at start of treatment [years]	Avg=41.2, std=8.4, min= 28.8, max=57
Gender	Female: 11 (48%), Male: 12 (52%)
Educational Level*	Avg=3.1, std=1.4, min=2, max=7
Age at symptom onset [years]	Avg=21.2, std=5, min=14, max=32
Total # of Hospitalizations (Pre-clozapine)	Avg=9.7, std=13, min=0, max=63
Duration total of Hospitalization (Pre-clozapine) [days]	Avg=615.7, std=928, min=0, max=3789
Chlorpromazine Equivalents (Pre-clozapine) [mg]	Avg=726.6, std=636, min=40, max=2485
Clozapine dose [mg/day]	Avg=344.6, std=157, min=50, max=600
Post-treatment Positive Symptoms Scale	Avg=17.8, std=3.4, min=11, max=24
Post-treatment Negative Symptoms Scale	Avg=23, std=3.9, min=12, max=32
Post-treatment General Symptoms Rank (GR)	Avg=46.3, std=5.7, min=32, max=56
Post-treatment Total Rank (PSS+NSS+GR)	Avg=87.2, std=10.9, min=58, max=101

Table 4.14: Available demographic information of the 14 schizophrenic subjects denoted by Group B.

Information	**Range**
Age at start of treatment [years]	Avg=35.7, std=10, min= 22, max=55.5
Gender	Female: 6 (43%), Male: 8 (57%)
Educational Level*	Avg=3.3, std=1.64, min=2, max=7
Age at symptom onset [years]	Avg=21.3, std=5.28, min=15, max=31
Total # of Hospitalizations (Pre-clozapine)	Avg=6.43, std=6.9, min=0, max=18
Duration total of Hospitalization (Pre-clozapine) [days]	Avg=470.8, std=627, min=0, max=1879
Chlorpromazine Equivalents (Pre-clozapine) [mg]	Avg=628, std=404, min=40, max=1169
Clozapine dose [mg/day]	Avg=396.4, std=101, min=200, max=500

patients will respond to clozapine. For this reason we also reanalyzed our data using a value $\delta_1 = 83.5$ which yields a 30% response rate (i.e. with 7 R and 16 NR subjects in group A).

We must confirm that the pre-treatment QCA means of the R and NR subgroups of group A subjects are not significantly different, so that the post-treatment PANSS rating alone accurately indicates the effect of the treatment on the subject. To this end, we conducted a hypothesis test on the means, assuming the QCA data points are independent and normally distributed, and that the variances of the R and NR groups are identical. It is straightforward to show that the respective likelihood ratio is F-distributed. In this case, df=10, 11 for the numerator and denominator, respectively, with F=1.1056 and p=0.43. Thus, there is no evidence to suggest the pre-treatment QCA means of the two groups are significantly different.

Group B subjects are defined as responders to clozapine therapy if there is an improvement of at least 25% between the pre- and post-QCA scores. This level of relative change represents a clinically significant improvement in symptom severity considering the fact that all the subjects in our study were in the treatment-resistant population [181]. See e.g., Kane et al. [179] who used a 20% relative change as response indicator.

4.4.1.1 The QCA Clinical Rating Procedure

The QCA clinical rating procedure was devised in the context of an unrelated earlier naturalistic retrospective un-published clinical study of treatment resistant schizophrenic patients being considered for clozapine treatment. The subjects in the present study were included in this previous study. An experienced clinician reviewed all the available clinical descriptive information of the patient's symptomatology prior to beginning a course of clozapine. Reported symptoms, corresponding to those described in the PANSS, were rated as: present, moderate or severe on a one to six point scale. Only explicitly described symptoms were scored and the clinical rater was instructed not to infer the presence of potential symptoms. The same rating was repeated, based on case records describing current symptoms at the time (usually after approximately six months) when the decision was made to either discontinue or continue with on-going maintenance clozapine therapy.

4.4.2 Results

The training set consists of a nominal M_p EEG epochs from each of N_p subjects, for a total of M_t epochs altogether. In the clozapine experiment, $N_p = 23$ and $M_p = 12$. However for two subjects in this study a total of 8 and

10 epochs (instead of 12) are available. Therefore the total number of epochs is $M_t = 270$.

See Chapter 2 as well as Section 4.1 for details of the feature extraction, regression/classification and evaluation methods used, and see Sect. 4.1.1 for details of the EEG recording process for this study. The regularized feature selection as illustrated by Eq. (2.7) is used. For responders, the values $y = 1$ and for non-responders $y = 2$ are arbitrarily assigned. As discussed in Section 4.1, the regression/classsification function is applied to each epoch. In the test phase, all the regression outputs are averaged for all epochs associated with each subject. The quantization is done by comparing the average of regression function for all epochs corresponding to each subject to the threshold 1.5, which means that an average of less than 1.5 is determined as being an R subject, otherwise it is reported as an NR subject.

4.4.2.1 Treatment-Efficacy Prediction Performance

The first set of results uses data from group A which consists of 23 subjects. The set of candidate features were extracted from the pre-treatment EEG data and then reduced into a set of $N_r = 8$ most relevant features. SVR and kernel PLSR models/classifiers were then trained. The prediction performance was then evaluated using the leave–two–out (L2O) testing procedure (i.e., a 12-fold cross-validation) as discussed in Section 2.4.3. However, similar performance is obtained with the L1O procedure. The number of latent variables in kernel PLSR model as well as other design parameters in both PLSR and SVR models are found using the nested cross-validation parameter optimization method described in Sect. 2.4.3.

The performance evaluation results using the combined EO and EC EEG data sets together for the 23 subjects, for a value $\delta_1 = 88.5$ and $N_r = 8$ are summarized in Table 4.15(i), where it is seen that the overall prediction performance is 87.12%. When δ_1 is reduced to 83.5 corresponding to a 30% responder rate, the overall performance becomes 89.7% as reflected in Table 4.15(ii). For this experiment, the performances of both the SVR and kernel PLSR regression methods were identical. In the kernel PLSR method, two major latent vectors are used. These results indicate that it is indeed possible to predict the response to clozapine therapy using the proposed methods. Further experiments were performed using a range of δ_1 from 83.5 to 92.5; prediction performance was above 85% in all cases.

The results using data from both subject groups A and B are now presented. For this second experiment, we train the classifiers using group A as training data, and then test the prediction performance over group B (with 14

Table 4.15: Performance results predicting the response to clozapine therapy in Group A subjects using combined EC and EO EEG data, and $N_r = 8$. Subjects with a post-treatment PANSS score (total-rank) of less than δ_1 are considered responsive (R).

(i) $\delta_1 = 88.5$	predicted R	predicted NR	% correct
actual R	10	2	83.33% = sensitivity
actual NR	1	10	90.91% = specificity
			average = 87.12%

(ii) $\delta_1 = 83.5$	predicted R	predicted NR	% correct
actual R	6	1	85.7% = sensitivity
actual NR	1	15	93.75% = specificity
			average = 89.7%

Table 4.16: Independent test performance using subjects in group A as training data (with $\delta_1 = 88.5$), and group B as test subjects. Response to clozapine therapy is defined as a more than a 25% improvement in the QCA score.

(i) $\delta_1 = 88.5$	predicted R	predicted NR	% correct
actual R	6	1	85.7% = sensitivity
actual NR	1	6	85.7% = specificity
			average = 85.7%

subjects). A group B responder in this case is defined as a subject having an improvement of at least 25% between the pre- and post-QCA scores. The average treatment efficacy prediction performance for this experiment was 85.7% as reflected in Table 4.16. This shows a satisfactory prediction performance under different conditions when the classifier is trained on one set, and then tested on another independent set.

We now show an example where a view of N_r-dimensional feature space is represented on a surface. In this case, the N_r-dimensional feature space is compressed into 2 dimensions using the low-dimensional representation technique described in Sect. 2.4. Fig. 4.7 shows a scatter plot of 270 points corresponding to the 270 available epochs of EEG data from the group A subjects. This figure was generated using the kernel PCA method with a Gaussian kernel. Filled circles correspond to responders and squares to non-responders. The number written beside each data sample is the corresponding subject index, assigned arbitrarily. Averaging the location of all data samples belonging to each subject results in Fig. 4.8, in which each subject is shown with one point.

Table 4.17: A list of discriminating features for clozapine treatment-efficacy prediction using pre-treatment EEG information. $\delta_1 = 88.5$. The mean and standard deviation of each feature over the responder (μ_R, σ_R) and non-responder groups (μ_N, σ_N) are shown.

#	*Selected EEG-driven Feature*	μ_R, ($\pm$ σ_R)	μ_N, ($\pm$ σ_N)
1	Mutual Information between T3 & P3	1.056 (± 0.868)	0.109 (± 0.571)
2	Mutual Information between T3 & O1	1.498 (± 2.382)	-0.169 (± 0.734)
3	Mutual Information between C3 & P3	-0.169 (± 0.997)	-1.154 (± 0.608)
4	Correlation between F8 & T4	1.072 (± 0.596)	0.56 (± 1.105)
5	Coherence at f=4Hz between T3 & P3	0.725 (± 0.865)	-0.356 (± 0.885)
6	Coherence at f=5Hz between T3 & P3	0.838 (± 0.853)	-0.178 (± 0.753)
7	Coherence at f= 6Hz between T3 & O1	0.713 (± 1.411)	-0.375 (± 0.864)
8	Coherence at f=6Hz between T3 & P3	0.718 (± 1.007)	-0.241 (± 0.812)
9	Coherence at f=6Hz between C3 & O1	-0.518 (± 1.007)	-1.169 (± 0.871)
10	Coherence at f= 9Hz between T3 & O1	0.816 (± 1.194)	-0.199 (± 1.011)
11	Coherence at f=10Hz between T3 & T5	0.532 (± 0.844)	-0.254 (± 1.088)
12	Coherence at f=10Hz between T3 & P3	0.774 (± 0.765)	-0.044 (± 0.913)
13	Coherence at f=10Hz between C3 & P3	0.007 (± 0.736)	-1.072 (± 1.3)
14	Coherence at f=11Hz between C3 & P3	0.008 (± 0.78618)	-0.804 (± 1.039)
15	Coherence at f=11Hz between T3 & P3	0.869 (± 0.728)	0.065 (± 0.815)
16	Coherence at f=12Hz between T3 & P3	1.06 (± 0.771)	0.092 (± 0.881)
17	Left to right PSD-ratio at f=12Hz, T5/T6	0.11 (± 0.828)	0.894 (± 1.348)
18	Coherence at f=12Hz between T3 & T5	0.688 (± 0.789)	-0.183 (± 0.928)
19	Coherence at f=13Hz between F7 & F3	0.418 (± 0.812)	-0.165 (± 1.221)
20	Left to right PSD-ratio at f=16Hz, T5/T6	-0.069 (± 0.793)	1.01 (± 1.274)
21	Left to right PSD-ratio at f=29Hz, T5/T6	-0.26 (± 0.983)	0.537 (± 0.866)

The clustering between the R and NR groups is clearly evident in this figure. Thus, this two-dimensional representation demonstrates that 1) a set of discriminating features exists that is sufficient for distinguishing responders and non-responders, and 2) the proposed machine learning techniques are indeed capable of predicting the long-term outcome of schizophrenic subjects being treated with clozapine. The clustering performance shown in this figure is indicative that the nonlinear transformation imposed using our Gaussian kernel in conjunction with a linear classifier or regression technique will perform well.

4.4.2.2 A List of Discriminating Features

Even though we have selected the value of N_r to be eight in our experiments, we show a list of the 21 most relevant EEG features of interest in Table 4.17. The features are sorted based on frequency (in Hz) of the corresponding statistic. Each of the features listed in the table is selected at least once over all L2O iterations in the cross-validation procedure.

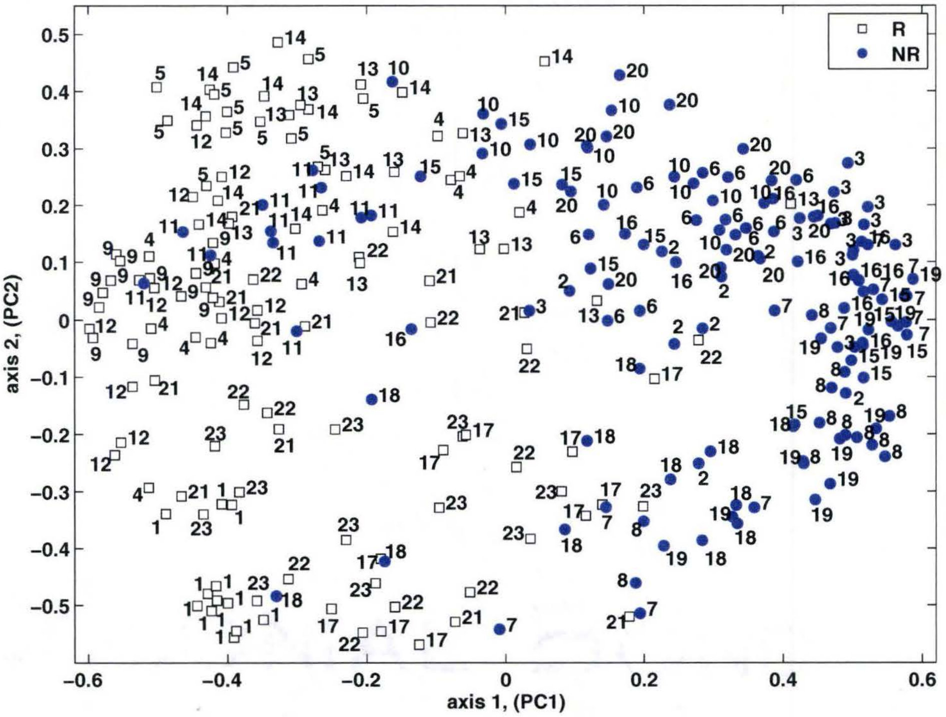

Figure 4.7: The clozapine study: a demonstration of the clustering behaviour of the selected discriminating features. The $N_r = 8$ dimensional feature space compressed into 2 dimensions using the KPCA method. There are nominally 12 data points corresponding to multiple EEG epochs from each subject. The subject index corresponding to each point is indicated on the plot. The clustering behaviour between the R and NR groups is clearly evident.

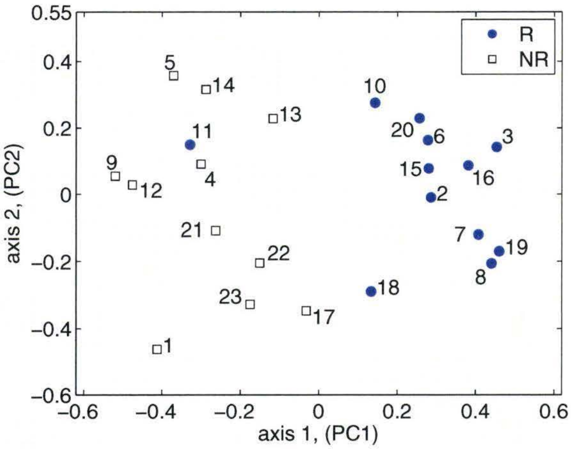

Figure 4.8: Same as Fig. 4.7, except that all data points belonging to each subject in Fig. 4.7 are averaged to provide one point per subject.

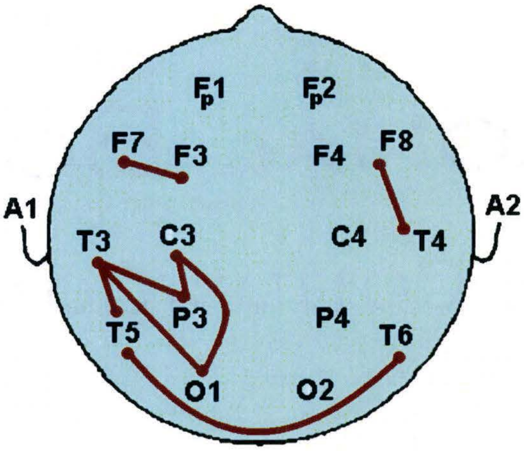

Figure 4.9: In clozapine case, a rough schematic drawing showing a list of some relevant features by connections, as reflected in Table 4.17. Connections are shown by solid thick lines. Electrodes A1 and A2 represent the linked ears reference.

As an example graphic explanation, Figure 4.9 is a depiction of the most-relevant features selected in Table 4.17. A connection between two electrode sites in the figure corresponds to a selected feature which involves those two locations. It roughly indicates any relations between EEG sensors that convey relevant information for our prediction problem.

4.4.3 Discussion

We can provide some further evidence of the validity of the proposed response prediction method for the clozapine therapy, as follows. Note that similar argument can be made based on the specific results obtained for each of the SSRI and rTMS therapies. First, the clustering behaviour shown in Figure 4.8 shows clean separation of the clusters, which is a strong indication that the reduced features can indeed discriminate long-term response. Also, with the LnO cross-validation procedure, different test and training samples are used in each iteration, and yet overall, a reasonable performance level is attained. This suggests the proposed machine learning procedure is consistent across variations of the input data. Particularly in the clozapine case, a final argument to suggest validity of the proposed method is with regard to the results of Table 4.16. Here, the prediction procedure is trained on Group A data and tested on a completely independent set of Group B data. Even though performance degrades somewhat, the resulting performance of 85.7% is still quite satisfactory. We can further examine the integrity of the proposed prediction procedure by evaluating the probability that our demonstrated prediction performance would have been due to chance alone. With reference to Table 4.15(i), there are 12 responders and 11 non-responders, so the probability p of a responder may be taken as $12/23 = 0.5212$. Assuming all subjects are independent, the probability of a prediction error is governed by a binomial distribution, which is parameterized by N, the number of samples, and p, in this case the probability of a responder. Therefore, the probability of this level of performance (10 classifications as R and 2 as NR out of $N = 12$ true responders) occurring due to chance alone is evaluated from the binomial distribution as 0.0226. Similarly, the value of p for the non-responder case is 0.4783, so the probability of estimating 10 NR and 1 R out of 11 non-responders due to chance alone is 0.0036. Similarly, for the case of Table 4.15(ii), the corresponding figures are 0.0039 and 0.0211 for the R and NR groups, respectively. Thus we see that these figures are negligibly small and we can conclude the prediction results are almost certainly a consequence of the distinguishing characteristics of the EEG measurements obtained from the two groups.

By employing more advanced analytical models, the present study was designed to extend and improve upon the utility of EEG in predicting the responsiveness to clozapine as investigated in other studies. Although Gross et al. [171] found that changes in EEG features correlated with outcome, post treatment EEG data was required. Our methodology is more potentially useful to the clinician as prediction is possible using EEG data collected before this potentially toxic treatment is initiated. Knott et al. [170] found that pre-treatment QEEG asymmetry was predictive of response. We observed that QEEG inter and intra hemispheric asymmetry were a useful features, but we were able to extend the findings by [170] to develop a method that automatically categorizes subjects into responder and non-responder groups. Both [171] and [170] looked at a limited set of quantitative EEG features as treatment-response indicators.

The goal of this study was to propose a new clinical data analysis method and derive an empirical set of EEG features predictive of response to clozapine, not to derive neurological information regarding the pathophysiology of schizophrenia. Nevertheless the clustering of relevant EEG features in the temporo-parietal area of the dominant hemisphere, as seen in Table 4.17 and in Fig. 4.9, may be of some interest to those studying regional brain activity patterns in patients with schizophrenia. Others have described bilateral reduced grey matter volume in the temporal lobes (e.g., Okugawa et al., [182]) and electrophysiological abnormalities in the left temporo-parietal region on EEG (e.g., Faux et al., [183]) in schizophrenic patients.

In this study, as in the study of prediction of response to SSRI medication, coherence variables were prominent among the features selected as predictive of treatment response. In contrast to the results associated with good response in the SSRI treated sample, coherences are in every case higher in the responder than in the non-responder group. Coherence in the temporo-parietal area seemed particularly important in that 7 of the 21 listed features involve either coherence or mutual information between the T3 and P3 electrodes. The neurophysiological meaning of this is obscure at this time. However, GABAergic inhibitory networks have been implicated in the pathophysiology of schizophrenia and in the response to clozapine [184], and these appear to be involved in regulating EEG coherence between the temporal and parietal areas [185]. One might speculate that patients with higher coherence in this area are afflicted with a subtype of illness that is more likely to respond to clozapine.

This retrospective study suffers from some weaknesses. Most notably our QCA clinical rating is based on chart review and therefore likely to be less accurate than a standardized PANSS. However, our raters were clinicians expert in the treatment of schizophrenia and familiar with the subjects being evaluated. The QCA would therefore have reasonable clinical validity. The high predictive accuracy of our algorithm in both group A and group B subjects even in the face of this source of outcome variance may speak to the robustness of this methodology. As QCA and PANSS ratings were completed years before this project they could not have been influenced by the machine learning assignment into responder and non-responder groups.

4.5 Concluding Discussion on Response Prediction

The performance of the proposed method suggests that suitably–selected features extracted from the EEG cluster according to how the patient responds to the treatment under consideration. Thus, the pretreatment EEG appears to contain information regarding brain functioning that is relevant to, and predictive of, the therapeutic effect of SSRI or rTMS antidepressant medications, or of the antipsychotic clozapine.

Most studies, have used a small preselected set of features and sometimes even a limited set of electrodes to determine whether it is possible to predict response to one or several antidepressant drugs. Our proposed feature selection process is novel in this respect, in that we have considered a large number of features including those, or similar ones, already cited in the literature, and reduced them into a much smaller set that is most statistically related to the response variable. In this way, rather than hypothesizing beforehand whether a particular feature is indicative of response and then verifying the hypothesis as required by previous approaches, the proposed method automatically identifies relevant features without the need for costly experimental verification. Thus our method can identify salient features that could be missed with previous methods.

Our experiments with SSRI and clozapine response prediction led to two important practical observations that may be relevant to deployment of this methodology in the clinical setting. First the predictive algorithms appear to work very well whether the test subject is taking medications at the time of the EEG sampling or not. This is relevant as taking a patient off medications not only delays initiation of new treatment but also can be associated with clinical worsening due to withdrawal symptoms. Secondly, effective response prediction using a reduced number of EEG electrodes is possible. This translates as less complex, less time consuming and less expensive EEG methodology.

What is particularly encouraging to us is that our feature selection process which reduces numerous EEG candidate features into an optimized reduced feature set has resulted in a feature set that includes frequencies at sites in the brain previously identified by other researchers as relevant to depression and its treatment. This confirms the validity of the proposed feature selection method and further suggests that the selected features have physiological/anatomical and clinical relevance.

Because some of the features have strong statistical dependencies, the set of selected features (e.g., in Table 4.8) is not unique. Some of the features may be replaced with others, with small penalty in performance. However, because of the inter-dependence of these features, a replaced feature set could be indicative of the same neurological information as the original and therefore likely correspond to closely related spatial locations and frequencies.

In our experiments, only non-significant differences between nonlinear SVR and kernel PLSR classifier models were observed. However, it was noticed that the SVR method provides good prediction performance even with a relatively small number of training samples. An additional benefit of the kernel PLSR model on the other hand is that it inherently provides a low-dimensional representation of the data. A larger dataset is needed for more robust comparisons. From various experiments done in this research, it is observed that the advantage of using Gaussian and polynomial kernels is apparent when the value of N_r is small. When N_r is larger, a linear kernel may give better results. This could be due to the fact that it is harder to find optimum design parameters for a more complex kernel. Despite this, the best overall test performance is obtained using a small number of relevant features and nonlinear kernels.

Note that most of the selected features listed in Tables 4.8, 4.12 and 4.17 are coherence or mutual information measures between a given pair of electrodes. These quantities are indications of synchrony between the respective regions. The fact that our proposed machine learning procedure has identified these synchronies as being predictive of response may provide valuable clues to the psycho-neuro-science community towards the understanding of the pharmacological mechanisms of SSRI, rTMS and clozapine therapies. Further exploration of this idea remains an exciting topic for further investigation, yet its development is beyond the scope of this thesis.

Looking further into Tables 4.8, 4.12 and 4.17 and Figures 4.3, 4.6 and 4.9, the difference between the relevant features in the SSRI, rTMS and clozpaine cases are in agreement with our goal of finding exclusive treatment-efficacy prediction models for each therapy, thus allowing the user to determine which therapy suits the subject under test. It can also be seen that there are several common features that have predictive value in both SSRI and clozapine therapies, however, these two cases share little common features with the rTMS therapy. The other observation is that several features in the frequency band greater than 17Hz (i.e., in the beta range) are found to be among the most relevant features for the rTMS therapy, but this is not the case for the SSRI case. A few of the relevant features in rTMS case are right–to–left PSD ratios, but in contrast, these kind of features were not among most relevant ones in the SSRI case.

It must be noted that the results for the SSRI, rTMS and clozapine studies are derived using a relatively small quantity of data. Our findings must be replicated with a much larger sample of training and test subjects before they can be accepted with confidence.

Chapter 5

Application to Neuroscience: Medical Diagnosis

5.1 Background

Psychiatric disorders are among major global diseases affecting a significant portion of the population [33], and in industrialized countries mental illnesses may account for about 16% of total health care costs [32]. In this chapter we propose an automatic diagnosis methodology which can further be used in treatment-planning, and towards better understanding of such disorders.

Currently, most psychiatric clinicians make a diagnosis based upon a standard set of diagnostic criteria such as the 'Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association' (DSM) [45] or the 'International Statistical Classification of Diseases and Related Health Problems' (ICD) [46]. The symptoms and signs of a neuro-psychiatric disease or condition are ordinarily reviewed and the critical information is discovered as follows: the clinician hears the presenting complaint, elicits subjective symptoms and, in some cases, conducts a physical examination of the patient. Based upon the information available at the time, a range of diagnostic possibilities is considered. The most likely diagnosis is designated the "preferred diagnosis". The other diagnostic possibilities are then listed in decreasing order of probability to form a "differential diagnosis". The preferred and differential diagnoses then suggest further analysis, including using laboratory and other clinical tests that will help to rule in or rule out the various entities in the diagnostic list.

The first step in an efficient treatment for a mental illness or disorder is a correct diagnosis. This can be a more difficult task than it might seem.

Though the diagnostic criteria of different conditions are designed to differentiate subjects with this condition from those with other conditions requiring other forms of treatment, often specific symptoms can appear in more than one diagnostic category and diagnostic criteria can overlap to the point where confident differentiation is impossible. Even the psychiatric expert can have difficulty distinguishing certain psychiatric conditions, e.g. psychotic depression from schizophrenia or, most notably, differentiating major depressive disorder (MDD) from bipolar depression (BD) (or the depressed phase of bipolar affective disorder– BAD). This distinction is highly relevant as the antidepressant medications that would be quite appropriate for MDD may, in the patient with BD, induce mania, or rapid cycling between depression and mania thus making the condition considerably worse [186]. Furthermore, current clinical diagnostic procedures are imperfect, i.e. all subjects meeting diagnostic criteria for an illness such as major depressive disorder (MDD) do not all respond to the same treatment e.g. antidepressant medication. This observation provides compelling clinical evidence for very substantial biological heterogeneity within a single diagnostic category. Automatic diagnosis tools such as the one based in this chapter, may help to improve the decision confidence and can further provide a finer classification of psychiatric illnesses or mood disorders.

There are various studies based on classic statistical analysis[1] which investigated employing data obtained from QEEG or EEG, biological measurements, other laboratory instruments, and various clinical markers in the diagnosis of psychiatric illnesses and mental conditions. Various biochemical markers, genetic markers, and brain imaging studies have been used to differentiate subjects with particular mental conditions from healthy volunteers and/or from subjects with other mental conditions.

The study in [187] (based on a sample of 11 subjects with a history of depression and 11 normal control subjects) reported that frontal brain asymmetry (FBA) is a potential marker for depression. It was found that subjects with current or previous incidence of depressive disorders tend to have an FBA ratio that lies towards the extremities of the distribution. In [188], the EEG data is analyzed to compare normal subjects with subjects suffering various mental disorders, and found that a decrease in low frequency band power (including delta band: 0.5–3.5Hz, and/or theta band: 4–7.5Hz) of the QEEG can be regarded as a specific sign of brain dysfunction, and is correlated with cortical atrophy as seen in brain magnetic resonance imaging (MRI). EEG abnormality during sleep is also investigated for diagnosis. For example, [189] used the EEG to study whether sleep rhythms differ between schizophrenic

[1] By classic statistical analysis, we mean using methods based on standard statistical tests.

subjects (N=18), healthy individuals (N=17), and a psychiatric control group with a history of depression (N=15). They reported that the schizophrenic group had a significant reduction in centroparietal EEG power, from 13.75 to 15 Hz, in relation to both the healthy and depression groups. Also they reported a decrease in sleep spindle number, amplitude, duration, and integrated spindle activity in schizophrenic subjects.

Use of the 'mutual entropy' between sensors of EEG (as a more advanced statistical feature) is discussed in [190] to diagnose Alzheimer's disease (AD) patients from normal subjects. Further, the work of [191] used spectral entropy and sample entropy measures (of each of 19 scalp electrodes) to study the difference between AD subjects (N=11) and healthy subjects (N=11).

As a more automatic numerical analysis approach, *machine leaning* methods have also been employed in many studies related to diagnosis. Machine learning or pattern recognition methods are used to learn the identification, classification or regression models based on the information in the training data set. The trained model is then used during the test phase. For example, the SVM technique to select spectro-temporal patterns from multichannel magnetoencephalogram (MEG) data collected during a verbal working memory task have been used to distinguish schizophrenic (N=15) from control subjects (N=23) [173], and they obtained 91.8% average diagnosis performance. Machine learning algorithms using structural brain MRI images [174], functional MRI (fMRI) data [176] and combined genomic and clinical data [177] have been employed to separate schizophrenic, bipolar disorder and healthy control subjects. Imaging techniques such as positron emission tomography (PET), magnetic resonance imaging (MRI) or MEG provide brain images with high neuroanatomical spatial resolution; however, it will be difficult to develop these technologies as clinically–proven diagnostic aids in psychiatric disorders [192]. In addition they are very expensive, and not readily available outside major medical centers. In contrast, the EEG is non-invasive, inexpensive and readily available in most community laboratories and hospitals. These attributes would potentially make the EEG a very practical tool as a diagnostic aid, particularly when combined with modern signal processing techniques.

As for EEG signal processing, the study [193] used EEG data and employed an 'artificial neural network' (ANN) method to differentiate subjects suffering from schizophrenia, depression and normal healthy persons. In a population with 10 subjects in each of the three classes mentioned above, the correct prediction performance rates they obtained with a multi-layer perceptron neural network (trained with the back-propagation technique) was 60%, 60%, and 80%, for each respective class.

A very basic form of classifier with a restricted structure was proposed in [119]. The method of [119] classifies a multi-variate set of features by performing a sequential one-dimensional decision on each individual variable. In contrast, the present work incorporates a joint multi-variate classification procedure which is based on a single multi-variate decision function. The present work uses a much more flexible and efficient classifier structure. Furthermore, the described method allows the classifier decision boundaries to be defined in an arbitrary/complex manner.

The study in [194] used the 'multiple discriminant analysis' method applied on EEG data to separate primary degenerative dementia from major depressive disorder, and obtained 91.1% average diagnosis performance. [195] employed principal component analysis (PCA) for feature enhancement and then the cosine 'radial basis function neural network' (RBFNN) model and the wavelet-chaos-neural network method are used for detection of epilepsy and seizure. The study in [74] used the SVM method for signal classification in epileptic subjects.

Brain computer interface (BCI) using EEG data is another closely-related problem. For example, [196] employed the 'Gaussian mixture model' (GMM) in the BCI. See e.g. [117,197–199] for other related studies.

The objective of this study is to help make the medical/clinical diagnosis procedure 'automatic' through the use of a cognitive/intelligent procedure that extracts critical quantitative information from EEG data and uses a powerful statistical machine learning methodology. These efficient mathematical methods permit the extraction of more useful information from the EEG than was ever possible using simple spectrum analysis, simple classical statistical analysis or visual inspection. The diagnosis result provided by the proposed method can assist the physician to make a better final decision and improve the overall diagnosis performance in clinical practice.

5.2 Participants

Our study sample consists of a total of 207 adult subjects, including 64 subjects with MDD or unipolar depression, 40 subjects with chronic schizophrenia, 12 subjects with bipolar depression (BD), also known as the *bipolar affective disorder* (BAD) and 91 healthy subjects. All psychiatric subjects were recruited from the case load of the St Josephs Hospital, Center for Mountain Health Services, Hamilton, Ontario. They were carefully diagnosed using the appropriate DSM criteria by experienced psychiatrists specializing in the management of either mood disorders or schizophrenia. In most subjects with MDD the diagnosis was confirmed using the Structured Clinical Interview for DSM (SCID). Also all subjects with schizophrenia (comprising both in-patients

Table 5.18: Some demographic and clinical information for 23 schizophrenic subjects.

Information	Range
Age at start of treatment [years]	Average=41.2, min=28.8, max=57
Gender:	Female:11 (47.8%), Male:12 (52.2%)
Educational Level	Average=3.1, min=2, max=7
Age at symptom onset [years]	Average=21.2, min=14, max=32
Clozapine dose [mg/day]	Average=344.6, min=50, max=600
Total Hospitalizations (Pre-Clozapine)	Average=9.7, min=0, max=63
Duration total of Hospitalization (Pre-Clozapine) [days]	Average=615.7, min=0, max=3789
Chlorpromazine Equivalent dose [mg]	Average=726.6, min=40, max=2485

Table 5.19: Some demographic and clinical information of 64 patients with major depressive disorder (MDD, or major depression).

Information	Range
Age at start of treatment [years]	Average=46.7, min=20.3, max=65.8
Gender:	Female:42 (65.6%), Male:22 (34.4%)
No. of patient in SSRI therapy	22 (including 13 female, 9 male)
Pre-treatment Ham-D17 in SSRI	Average=23.3, min=18, max=40
No. of patient in rTMS therapy	42 (including 29 female, 13 male)
Pre-treatment Ham-D17 in rTMS	Average=21.3, min=15, max=29

and out-patients) met both the DSM-IV criterion for schizophrenia [45] and the Kane et.al [179] criterion for treatment resistance. Patients meeting these criteria may be considered to be suffering acutely from schizophrenia. In summary, the clinical diagnosis (used as our reference value) is done with high accuracy. All subjects gave informed consent to participate in the study.

Table 5.18 shows some sociodemographic and clinical information of the schizophrenic patients who participated in the study, and Table 5.19 shows some information for patients with major depressive disorder.

5.3 Data Analysis

In subjects with major depression (unipolar or bipolar), resting EEG signals are measured after 10 days of medication withdrawal and before commencement of treatment. In subjects with schizophrenia, again the EEG data correspond to the pre-treatment stage; however, for clinical reasons, they do not have a complete 'drug washout' before recording the EEG.

The EEG measurements and feature extraction procedures were the same as those in Chapter 4. For details of the EEG data recording, see Section

4.1.1. For details of the candidate numerical features extracted from EEG, see Section 4.1.2. For details of data analysis procedure, see Section 4.1. For multi-class feature selection, we use the *feature index collection* or the *feature combination* method described in Section 2.2.5.

We used a statistical decision/identification approach which provides a high-performance diagnosis tool with three relevant benefits: (i). The class-conditioned diagnosis likelihoods for various diagnosis possibilities are estimated directly, (ii). It allows the incorporation of prior information (if available), (iii). Various decision costs in the multi-class diagnosis problem can be optimally incorporated. To estimate the probability density functions, we specifically used a mixture of factor analysis (MFA) model and the expectation-maximization method [82] for estimation of model parameters, as discussed in Section 2.3.6.

In learning the statistical diagnosis model, for simplicity, the maximum likelihood classification rule is used. This uses the most uncertain but fair prior information (i.e., using equiprobable assumption) and using default decision cost values (i.e., taking $C_{i,j} = 1$ for $i \neq j$ and $C_{i,i} = 0$, for all i and j).

In the following experiments, all recorded eyes-open and eyes-closed EEG data are used collectively. Our final diagnosis result for each patient is based on averaging the likelihood values for all corresponding data epochs before a final decision is made, in a manner similar to the way the results from multiple epochs were combined in Chapter 4.

In measuring the diagnosis performance, "actual" or "reference" diagnosis, is what the expert physician diagnosed in the clinic, and "estimated" diagnosis is what our cognitive diagnosis model generated using only EEG data with machine learning methodologies.

5.4 Diagnosis Performance Results

We studied several diagnosis experiments to evaluate the performance of the proposed method from various perspectives using the cross-validation procedure discussed in Section 2.4.3.

In experiment 1, we studied a two-class diagnosis scenario using the pre-treatment EEG to differentiate normal (healthy) subjects from subjects who suffer from either schizophrenia or MDD. Table 5.20 reflects the diagnosis performance using $N_r = 14$ selected EEG features. The cognitive diagnosis system made 9 errors out of a total of 195 subjects, and the average performance is 95.2%. Also we noticed that from the group of subjects with a mental disorder (either MDD or schizophrenia), all 3 misdiagnosed by the cognitive system are among those clinically diagnosed by physician as schizophrenic.

Table 5.20: Experiment 1: Diagnosis performance results for differentiating healthy subjects from subjects with mental disorders (who have either schizophrenia or MDD), using pre-treatment EEG information.

	Estimated as MDD or Schizophrenic	Estimated as normal	Total No.
Clinically diagnosed as MDD or or Schizophrenic	101 (97.1%)	3	104
Normal (or healthy)	6	85 (93.4%)	91
		$\mu_H(\Theta)$= 95.2%	195

Table 5.21: Experiment 2: Diagnosis performance results for recognizing subjects with MDD from subjects with schizophrenia.

	Estimated to have MDD	Estimated to be Schizophrenic	Total No.
Clinically diagnosed as MDD	57 (89%)	7	64
Clinically diagnosed as Schizophrenic	5	35 (87.5%)	40
		$\mu_H(\Theta)$= 88.3%	104

In experiment 2, we investigated a two-class diagnosis problem to differentiate subjects with MDD from schizophrenic subjects. Table 5.21 reflects the diagnosis performance using 14 selected discriminating features selected among all candidate quantitative features described previously. The number of misdiagnosed cases is 12. This may be justifiable due to the fact that there are many common symptoms between these two types of mental disorders (e.g. the "negative symptoms" of schizophrenia [180] are very similar to depressive symptoms) and it is reported that this kind of confusion happens in routine clinical practice in a small but noticeable percentage of cases.

Experiment 3 is an additional binary classification example, this time between MDD and BAD subjects. These two conditions are very difficult to distinguish, and even in the absence of a past history of an episode of mania or hypomania, often impossible to differentiate in a clinical setting. Due to the imbalance in the number of training samples in the BAD and MDD groups (12 BAD and 64 MDD subjects respectively), the classification procedure becomes biased toward the majority population, reducing the level of performance. To avoid this difficulty we used the following procedure. We divided the 64 MDD subjects into 4 subsets each of size 16 subjects. Then 4 separate diagnosis

Table 5.22: Experiment 3: Diagnosis performance results for recognizing subjects with BAD from subjects with MDD, using pre-treatment EEG information. $N_r = 8$ and L2O procedure is used.

	Estimated to have MDD	Estimated to have BAD	Total No.
Clinically diagnosed as MDD	60 (93.8%)	4	64
Clinically diagnosed as BAD	4	44 (91.7%)	12×4
		$\mu_H(\Theta)$= 92.7%	76

experiments are performed where in each experiment each of the MDD subsets are sequentially tested against the 12 subjects with BAD. The overall contingency table is then constructed by adding the respective entries from the individual tables obtained from each experiment, as shown in Table 5.22. Again, the combined performance level of approximately 92% offers promising potential for the proposed method. In experiments 2 and 3, the values for K and m of the MFA model were determined in each fold of the cross-validation procedure from the candidate sets $[1, \ldots, 5]$ and $[1, \ldots, 4]$, respectively. A list of discriminating features found in this experiment are reflected in Table 5.23. The features are sorted based on frequency of the corresponding statistic.

Furthermore, in experiment 3, a comparison among different diagnosis or classification models is shown in Table 5.24. Using a grid search and the cross-validation procedure (using the training data), the best design parameters for each method is found, and Table 5.24 reflects the best performance attainable with each technique. It shows that the MFA model outperforms other methods. RBFN stands for 'radial basis function neural network'. The 'linear discriminant analysis' (LDA) is a standard classification technique used as a reference in our comparison [4]. See [5, 69] for details of the support vector machine (SVM) method.

In experiment 4, a three-class diagnosis problem is studied. Table 5.25 shows the result when $N_r = 42$ relevant features are used to construct the diagnosis model. The diagnosis performance is above 85% in all cases. For feature selection, the 'feature combination' method as described in Subsection 2.2.5 is used to generate the results as reflected in Table 5.25.

5.4.1 Clustering Performance

Here we investigate the (unsupervised) clustering performance, using a two-dimensional (2D) representation of the feature space, to gain insight into how

Table 5.23: A list of the most discriminating features in Experiment 3, diagnosis of MDD versus BAD.

#	*Selected EEG-driven Feature*
1	F/B PSD-ratio at f=5 Hz between T3/T5
2	Coherence at f=5 Hz between T6 & P4
3	Coherence at f=12 Hz between T6 & O2
4	Coherence at f=12 Hz between T6 & P4
5	F/B PSD-ratio at f=15 Hz, F1F7F3/T3C3
6	Coherence at f=17 Hz between O1 & O2
7	Coherence at f=17 Hz between T5 & O2
8	Coherence at f=18 Hz between T5 & O2
9	Coherence at f=21 Hz between O1 & O2
10	Coherence at f=22Hz between O1 & O2
11	F/B PSD-ratio at f=22 Hz, C4/O2
12	F/B PSD-ratio at f=34 Hz, F1/F3

Table 5.24: In Ex. 3, Comparison of diagnosis performance among different methods in recognizing subjects with BAD from subjects with MDD. $N_r = 8$. $y = 1$ corresponds to the decision that that the patient is suffering from MDD, and $y = 2$ corresponds to BAD.

method	specificity, $p(\hat{y} = 1 \mid y = 1)$	sensitivity, $p(\hat{y} = 2 \mid y = 2)$	% average performance
NN	0.859	0.771	81.5%
RBFN	0.875	0.813	84.4%
LDA	0.859	0.833	84.6%
SVM	0.906	0.833	87%
kernel PLSR	0.906	0.854	88%
RLS	0.922	0.875	89.8%
MFA	0.938	0.917	92.7%

Table 5.25: Experiment 4: Three-class diagnosis performance results for detecting the type of psychiatric disorder: MDD vs. schizophrenia vs. healthy.

	Estimated MDD	Estimated Schizophrenic	Estimated Normal	Total No.
Diagnosed as MDD	55 (85.9%)	6	3	64
Diagnosed as Schizophrenic	3	35 (87.5%)	2	40
Normal (or healthy)	4	7	80 (87.9%)	91
			$\mu_H(\Theta)$= 87.1%	195

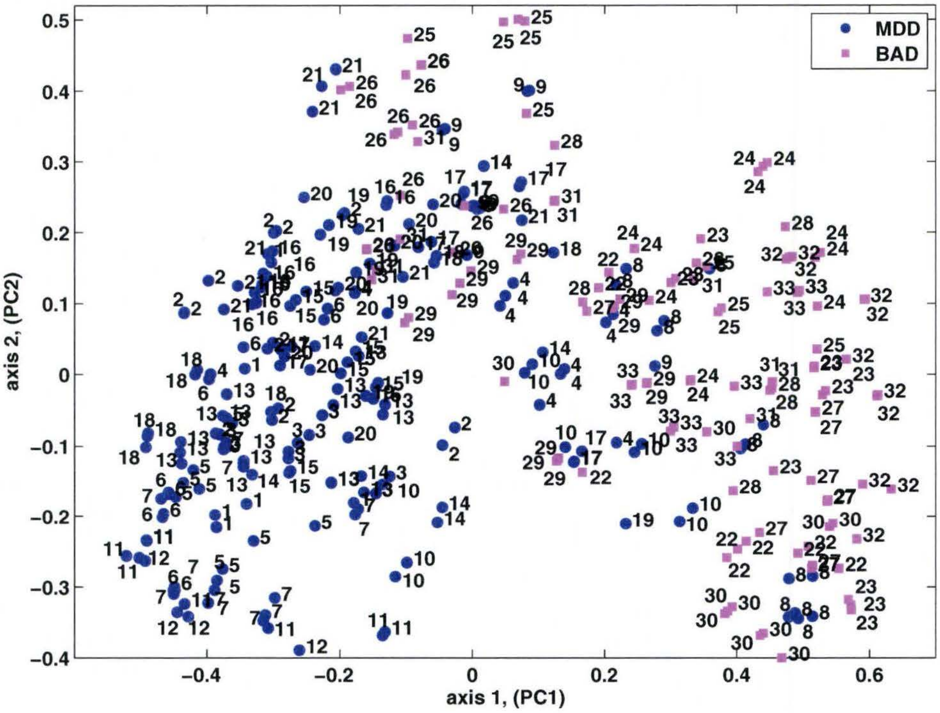

Figure 5.1: For Ex. 3: Scatter plot of the projection of pre-treatment data samples onto the first 2 major principal components using the KPCA method. Clustering of the BAD versus MDD epochs is described. $N_r = 14$.

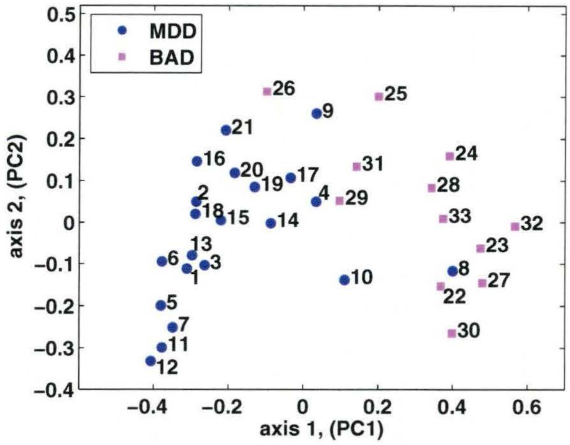

Figure 5.2: For Ex. 3: Subject-wise scatter plot of projected pre-treatment data for diagnosis between BAD versus MDD subjects, using the KPCA method. See Fig. 5.1.

well the small set of selected features (of size N_r) discriminate the diagnostic classes.

The two-dimensional scatter plots for clustering analysis, were generated using two principal vectors of the KPCA method with a Gaussian kernel, as discussed in Section 2.3.7.

We performed a clustering analysis for diagnosis experiments 1 to 3 discussed previously. In experiment 3, which considers diagnosis between BAD versus MDD subjects, Fig. 5.1 shows a scatter plot of 352 points corresponding to 352 distinct epochs of data. The points represent a collection of EO and EC sessions of pre-treatment EEG recording of 33 subjects (33 = 21 MDD + 12 BAD), projected onto only the first two major nonlinear principal components. Fig. 5.1 shows 223 MDD plus 129 BAD distinct epochs. The number of distinct epochs for each patient varies. The subject index is written beside each point. It shows a noticeable clustering of the subjects into the MDD and BAD groups. However, there are some overlaps as expected, due to the fact that this 2D representation is a limited one, and that better separation can be obtained in the higher N_r–dimensional feature space. Nevertheless this figure shows the diagnostic power of just two KPCA latent variables.

Averaging the location of projected epochs belonging to each subject results in Fig. 5.2, in which each subject is shown with only one point.

Fig. 5.3 shows a subject-wise scatter plot for Experiment 2 which considers diagnosis between MDD versus schizophrenia using $N_r = 14$ selected features

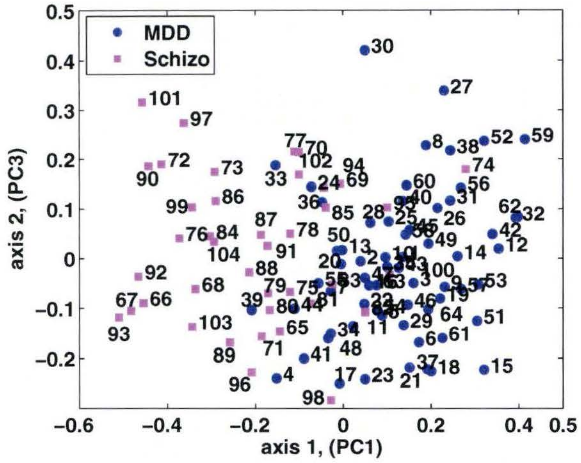

Figure 5.3: For Ex. 2: diagnosis between MDD versus schizophrenia: Subject-wise scatter plot of projected pre-treatment data, using the KPCA method.

projected onto two nonlinear principal eigenvectors. Before averaging the 2D location of data epochs belonging to each subject, we had 1723 epochs (which included 1066 MDD samples and 657 schizophrenic samples). The subject index is shown beside each point in the graph. The first and third principal components of KPCA are chosen for this 2D plot.

Fig. 5.4 shows a subject-wise scatter plot for Experiment 1, i.e., diagnosis between normal (healthy) subjects versus either MDD or schizophrenia. Before averaging the 2D location of epochs belonging to each subject, we had 3111 distinct data samples/epochs (which included 1640 samples with MDD or schizophrenia, and 1471 normal samples). Subject indices 1 to 64 are MDD, subjects 65 to 104 are schizophrenic, and indices 105 up to 195 are normal subjects.

It is to be noted that for clustering analysis, in addition to using KPCA method (as reported here), we further studied the results by 'isometric embedding' (ISOMAP) [93], the 'locally linear embedding' (LLE) [94], and the Graph Laplacian [95] methods and we obtained approximately similar clustering performance; however, the shape of the graphs (i.e., the spatial distribution of clusters) were different in each method. These further results are not reported in this thesis due to space limitations. Nevertheless, this satisfactory clustering performance confirms the proposed diagnosis method developed in this thesis.

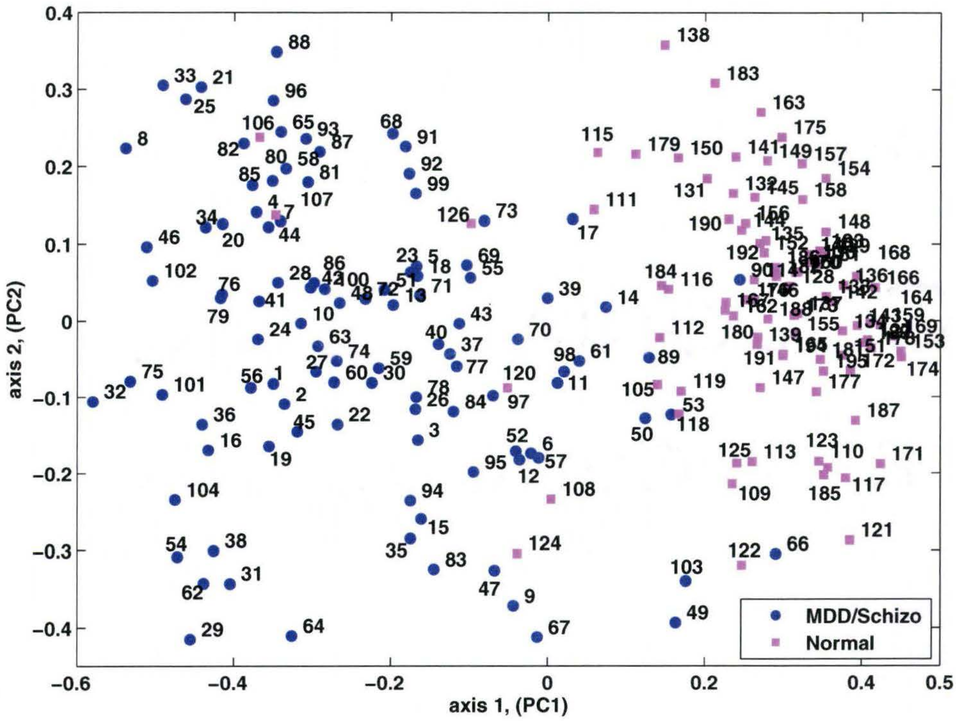

Figure 5.4: For Ex. 1: diagnosis between normal (healthy) subjects versus psychiatric subjects with either MDD or schizophrenia: Subject-wise scatter plot of projected pre-treatment data, using the KPCA method. $N_r = 14$.

By the data analysis methods used in this thesis, the 4-class diagnosis performance is found to be low. The following two reasons could explain this limitation: (i) Each of the feature selection methods illustrated in Sections 2.2.2, 2.2.3 and 2.2.5 does not give good performance for the multi-class diagnosis case when the number of classes is more than 3. (ii). The number of subjects for the BAD class is only 12 which is very small compared to MDD, schizophrenia and normal classes. Therefore, the data available for a 4-class diagnosis study is imbalanced which lowers the performance of both feature selection and classification processes.

5.5 Discussion

In our experiments, we first tested the diagnosis performance using only EO data, or using only EC data[2], and obtained similar diagnosis performance numbers as compared to the results in Section 5.4. In either of these experiments, the number of data epochs is less than the case when all available EO and EC are used together. Despite the fact that the EEG data has different characteristics in the EO versus EC case, our diagnosis models were able to find common discriminative features from the two EEG recording conditions.

In conclusion, the proposed EEG–based methodology, consisting of the feature selection method and the MFA classification procedure [82] is found to be very efficient for diagnosis of psychiatric disorders. The superior performance of the MFA method for this application, in comparison to other forms of classifier, is very likely due to its ability to model a low-dimensional nonlinear manifold using a combination of linear components. Furthermore, the proposed method outputs a soft decision in the form of a likelihood statistic for each of the classes, as opposed to a hard decision as in the case of other common forms of classifier. This provides the clinician with the likelihood of occurrence of each of the illnesses, for a given patient. This can be of value e.g., in prescribing treatment in the case of a co-morbid illness.

The clustering results in Figures 5.1– 5.4 show that the respective classes are separable using a simply-described kernel function (a Gaussian kernel with only a single parameter) and a linear boundary. The fact that satisfactory results can be obtained with such a simple model is an indication that overfitting is not a dominant issue.

Findings such as these suggest that machine learning may find an important place in the tool chest of the medical practitioner, particularly when

[2] Using only EC data means training classifiers using only EC data and then testing on only EC samples of EEG data.

experienced psychiatric personnel are not readily available. Confirmation of diagnosis may permit the clinician to initiate appropriate treatment while awaiting expert psychiatric assessment, which even in urban areas may not be available for weeks or months.

Chapter 6

Conclusions and Future work

6.1 Conclusions

Automated machine learning processes for (i) pipeline inspection, (ii) treatment-response prediction and medical diagnosis applications are developed. The promising performance indicated by our experiments shows the potential for these methods to be further exploited for commercial use.

For pipeline inspection, in this thesis we showed that modern machine learning methodology can be efficiently employed for detection and sizing of metal defects using MFL images. All the data used in our experiments were real data. Two metal-loss detection (i.e., recognizing major metal-defects versus benign defects or noise) experiments were studied: The first one include 1529 MFL image segments collected from an 8-inch pipeline. The second experiment include 1919 MFL image samples collected from a 10-inch pipeline. Several cross-validation tests are investigated. In addition, we studied low-dimensional representation or clustering of the input data using the KPCA method. The average detection performance in all metal-loss detection experiments was found to be over 95%. The relatively distinct clustering of various metal-defects in 2-dimensional space further confirms that it is possible to differentiate metal-defects by a machine and obtain high performance. After detecting serious or *injurious* metal-losses, the second step devised in this thesis was depth estimation. Estimation of defect depth using 58 real data samples was found to be efficient and the root-mean square estimation error was less than 8%. In practice, the developed automated detection and estimation system shows the potential to assist the technician who examines long records of MFL images of a gas/oil pipeline for defects. The automated machine can replace or, if used in combination, facilitate the slow and error–prone inspection process currently done by human operators.

Though the findings with respect to detection of pipeline defect are important, the results from the neuroscience application was the main accomplishment of this PhD work. For prediction of treatment-efficacy, three separate clinical studies are analyzed: (i). SSRI antidepressant therapy, $M = 22$ subjects, (ii). repetitive transcranial magnetic stimulation (rTMS) therapy, $M = 41$ subjects, (iii). the antipsychotic drug clozapine, $M = 23$ subjects. The performance of the proposed machine learning methodology suggests that suitably selected features extracted from the EEG cluster according to how the patient responds to the treatment under consideration. Thus, the pretreatment EEG appears to contain information regarding brain functioning that is relevant to, and predictive of, the therapeutic effect of the SSRI, rTMS, or clozapine treatments. In the pilot studies considered in this thesis, machine learning methods were found to be capable of employing pretreatment EEG data to accurately predict (with over 85% performance) whether a given patient will or will not respond to treatment with SSRI, clozapine or rTMS. We studied the results by cross-validation tests, several experiments with independent training and test sets, as well as clustering analysis. The list of discriminating features found through the feature selection procedure, also showed potential neuropathological relevance based on previous literature. By analyzing the average and standard deviation values for responder and non-responder groups, we found that each single relevant feature has some predictive value. This can be considered as another confirmatory result. While each single feature on its own is not sufficient for obtaining high prediction performance, the joint combination of the features does offer adequate performance.

Most related clinical studies for treatment-response prediction, including those cited in this study, have used a small preselected set of features to determine whether it is possible to predict response to one or several antidepressant drugs. Our proposed feature selection process is novel in this respect, in that we have considered a large number of features including those, or similar ones, already cited in the literature, and reduced them into a much smaller set that is most statistically related to the response variable. In this way, rather than hypothesizing beforehand whether a particular feature is indicative of response and then verifying the hypothesis as required by previous approaches, the proposed method automatically identifies relevant features without the need for costly experimental verification. The proposed method may identify features that could be missed by previous methods.

For diagnosis, in a training set of 207 subjects, including 64 subjects with major depressive disorder (MDD), 40 subjects with chronic schizophrenia, 12 subjects with bipolar depression and 91 normal or healthy subjects, the average correct diagnosis rate attained using the proposed method is over 85%, as determined by various cross-validation experiments. The promise is that, with further development, the proposed methodology could serve as a valuable adjunctive tool for the medical practitioner

As serious mood disorders like major depression (or MDD) are common conditions, the economic and clinical benefits of this work are substantial, if validated in a larger data sample. The proposed methodology if validated, could benefit patients, physicians, as well as public and private health-care and disability insurers. For expert clinicians, for example, the treatment-response prediction as well as diagnosis reports generated by the proposed methods could be considered an adjunctive tool permitting more confident diagnosis by the attending clinician. Further, the methodology if implemented as a remotely accessible web-based system, could help in providing healthcare service to remote areas which do not have access to expert physicians.

The machine learning process in all of the above applications was the same: Feature extraction, dimensionality reduction and then classification/regression. For feature extraction, from the measured data, we calculated a large set of candidate quantitative features which were already used in the related literature. Since we don't know which features are actually discriminative, the feature selection procedure is used to find those items that are statistically discriminative. The regularized feature selection based on maximizing the mutual information or maximizing the Kullback-Leibler distance was found to be efficient when the final classification performance based on the selected features are considered. After this stage, the reduced-dimensionality features are fed to the classification or regression model which generated the final target variable. For classification/regression we used RLS, SVR and PLS methods as well as a statistical decision method based on MFA model. These methods are all found to be efficient with small differences in average prediction performance.

The most concerning issue with respect to this work is that of overfitting i.e. developing a predictive algorithm that performs very well in the training sample, but much less so when tested in a new set of subjects. To some extent this concern was mitigated by use of the nested cross-validation procedure. Using cross-validation with independent training and test datasets diminishes the likelihood of the over-fitting issue. Nonetheless the small data sample size, and the reanalysis of the same subject's data multiple times during our iterative cross-validation procedure could still result in some degree of overfitting. The most convincing test of the general accuracy of any machine learning algorithm is testing in an entirely new and independent test sample, as was done during our clozapine experiment. Our group will focus on this validation step in future studies.

There are other machine learning and clinical data analysis issues which remain to be solved. In the next section, the suggested further work is described.

6.2 Future Works and Recommendations

The following are various topics for future research:

6.2.1 Further Study in the Neuroscience Application

Some further study topics for treatment-response prediction as well as diagnosis applications are:

1). The data used in the neuroscience aspect of this thesis are the result of pilot studies. A much larger quantity of data must be acquired for the development of an expanded training set before the proposed methods can be employed in clinical context. Therefore first validating the results in a larger independent test data sample and then combining these new data with our existing database to create a larger training set are future research topics.

2). In the neuroscience application, the data analysis procedure and methodology described in this thesis could be extended to construct models that predict the response to various other treatments available for patients with MDD, BAD or schizophrenia. Furthermore, it may be possible also to incorporate information from other sources (such as symptom rating scales, scores from personality inventories and other psychiatric evaluations, the levels of various hormones etc. in the blood, demographic and socioeconomic information, etc.) to improve the performance and to further reduce the decision ambiguities. The same machine learning methodology for treatment-response prediction and diagnosis can be used for various other therapies and psychiatric disorders that are not experimented in this thesis.

The author is currently involved in a study partially funded by Magstim Co. Ltd., Carmarthenshire, Wales, UK. The objective of this Magstim-funded (REB approved) program includes collecting pre-treatment clinical data for the purpose of prediction of response to the following treatments for major depressive disorder (MDD): cognitive behaviour therapy (CBT), Escitalopram, Venlafaxine and Bupropion. The plan is to recruit 120 subjects with major depressive disorder (MDD) over the next 18-24 months. For each subject, we collect a pre-treatment EEG and various clinical assessments such as personality inventories, quality of life indicators, etc. Non-responders are switched to another form of treatment after a six-week period. There is also an additional

REB-approved related study in progress in our research group, on response-prediction to electro-convulsive therapy (ECT) for subjects with MDD. Data is also currently being collected for this study.

3). In the neuroscience application, a large number of EEG sensors (16 electrodes in the standard 10-20 system) is used, and a relatively good prediction performance is obtained. Therefore a topic worthy of further investigation is using a minimum configuration of EEG electrodes that will be sufficient to obtain adequate prediction and diagnosis performance. Using a smaller number of EEG electrodes is useful from many aspects: cost, EEG electrode assembly time and therefore total data acquisition time, and convenience for the patient.

4). In the experiments for the neuroscience application, a collection of previously described candidate numerical features are used. These features are an indication of an associated neurological function, e.g. coherence indicates synchrony between respective brain regions. A topic of future investigation is the use of more mathematically complex features or the use of combinations of these simple features as input to the classifiers/regressors. For example, a candidate conjunction feature to use is KPCA nonlinear principal components extracted from current features (i.e., features described in Section 4.1.2).

5). Based on the results in the neuroscience application (such as Tables 4.8 and 4.17 and Figures 4.3 and 4.9), the list of discriminating features that were found to be predictive of treatment-response in conjunction with the location of the respective EEG electrodes, require further investigation. Such tables and figures may give some clues about the locality and interconnection of neurological mechanisms associated with a positive response to the corresponding treatments and to the possible understanding of the psycho-pathology of various psychiatric disorders. Further investigation of this matter with the help of psychiatric experts and clinicians remains a promising topic for future work.

6.2.2 Further Study in the Pipeline Inspection Application

A critical part of the automatic pipeline inspection is the collection of large samples of real data from various pipelines and various operating conditions: various pipeline usages (Gas, Oil at various pressure levels), various pipeline diameters, various materials, etc. Therefore, an area of future work is extending the result to a larger dataset and validating the findings.

Another area of further work is incorporating machine learning techniques as a coarse but fast model for 'space mapping' methods (see e.g. [16]), which employ finite-element modeling of magnetic fields in the pipeline. Also, another use of machine learning methods is preliminary detection of metal defects, and then using inversion methods or space mapping methods to further analyze the sector of MFL data where a defect is found.

The following sections present further ideas that may lead to improved performance in both pipeline inspection and neuroscience applications.

6.2.3 Improving Efficiency of Feature Selection Process

The computational cost and efficiency of the regularized feature selection method based on maximum KL distance (i.e., the maxKLD method), as explained in Section 2.2.3 and Eq. (2.6), needs to be further improved. An area for future work is an efficient design of the optimization problem to incorporate minimization of the redundancy among selected features. One way to do this could be to use a revised second term that more efficiently quantifies the statistical dependence, or normalizing the second term so that the difference between pdf of two separate features is properly addressed.

Another topic of future work is using a multi-dimensional feature selection or a *feature subset selection* method instead of selecting the discriminating features one by one. A further work could also be measuring the feature selection performance based on the classification performance rather than doing feature selection and classification processes separately.

6.2.4 Multi-class Feature Selection

For the multi-class medical diagnosis problem, as an extension to the *feature index collection* method (as previously discussed in Section 2.2.5), a multi-class feature selection procedure based on the one-versus-one multi-class classification procedure might improve the feature selection performance. The proposed procedure is as follows. Multiple two-class classifiers (similar to [1, 73, 74] but incorporating a novel multi-class feature selection strategy) are combined to perform the final multiclass classification task. Assume that we have N_p classes of illness/disorder. All possible binary classifiers are built separately and the feature selection is done independently for the corresponding binary diagnosis problem. The total number of binary classifiers are $N_p(N_p - 1)/2$, and only binary feature selection is used in building each binary classifier. Finally the overall diagnosis result can be the majority vote among the results of all binary classifiers, or a similar structure. We denote this as the *multi-binary* feature selection and diagnosis method.

6.2.5 Concatenated Dimensionality Reduction

Here, based on the methods in Sect. 2.2 and 2.3.7 a concatenated method for dimensionality reduction will be described. This is a two-stage process as described below.

1. Select a relatively large list of N_b most-relevant features using the KL distance or 'mutual information' criterion, or using a regularized feature selection method as described in Sect. 2.2. Here, $N_b \gtrsim 10N_r$.

2. Then use manifold learning[1], or a low-dimensional representation method, as discussed in Section 2.3.7, project the data into a much lower-dimensional manifold (with dimensionality N_r). After this process, we will have N_r transformed or projected features. For example, in our neuroscience application, where initial number of candidate features (N_c) are in the few thousands range, we could extract $N_b \cong 110$ most relevant features out of the N_c available, and then determine the final features with dimensionality $N_r = 5$ (using manifold learning methods), that are then fed to the classification and regression models.

By performing the above two-step dimensionality reduction procedure, we will get a more compact representation of the measured information. Note that each of the N_r features calculated as above is a linear or non-linear combination of the N_b features, while in feature selection methods like the one described in Section 2.2, N_r features are selected and the remaining information in $N_c - N_r$ features are discarded, which might result in loss of some significant information.

6.2.6 Using Bayesian Network Models

A topic of interest for future work is building a classification/decision model based on a Bayesian network structure. A Bayesian network (BN) is a statistical graphical model that represents a set of variables and their probabilistic independencies. Bayesian networks are directed acyclic graphs whose nodes represent variables and whose arcs encode conditional independencies between the variables [78, 200, 201]. BNs are a useful representation for hierarchical Bayesian models. In such a model, parameters are treated like any other random variable, and become nodes in the graph. "Fundamental to the idea of a

[1] By manifold learning methods, we mean methods like KPCA [71], Graph Laplacian [95] that explore nonlinear data structures.

graphical model is the notion of modularity — a complex system is built by combining simpler parts", [202].

The flexibility of a BN structure allows combining various data types. The other benefit is providing likelihood values that will help make a better decision.

6.2.7 Semi-supervised Learning and Inference

The existence of similar structures across a set, or family, of patterns is often recognized as a low-dimensional manifold embedded in high dimensional space. One of the general assumptions in order to have an efficient semi-supervised learning is 'smoothness': If two points $\mathbf{x}_1$ and $\mathbf{x}_2$ are close in a high-density region in the measurement space (i.e., they belong to the same cluster), then the corresponding outputs (or labels) y_1 and y_2 are also close.

In semi-supervised learning, in addition to labeled data where the target value, class/category is associated with the measured data, we have many unlabeled data points available. Using the information inherent in the available unidentified (or unlabeled) data samples, the goal is to improve the classification/regression/clustering performance. See [203] for a review of popular methods.

1. Using geometric properties of data in high-dimensional observation space (i.e., using graph-based and manifold learning methods), [204–206].

2. Building statistically generative data model

In problems where the geometric distribution of sample points is such that graph-based methods will not help, then using 'statistical generative models' is an alternative. Generative models assume a model $p(\mathbf{x}, y) = p(y)p(\mathbf{x}|y)$ where $p(\mathbf{x}|y)$ is an identifiable mixture probability distribution, for example Gaussian mixture models. With a large amount of unlabeled data, the mixture components can be estimated; then the labeled examples (at least one labeled example per mixture component) will be used to fully determine the mixture distribution. The expectation-maximization (EM) method [207], for example, can be used to train the generative mixture model. However, the EM procedure is prone to local maxima. If a local maximum is far from the global maximum, unlabeled data may hurt learning. If the mixture model assumption is correct, unlabeled data could improve accuracy.

The idea of semi-supervised learning can be used in dimensionality reduction process as well. Similar to the proposed 'concatenated dimensionality reduction' procedure described in Section 6.2.5, a semi-supervised variation of this method is also proposed that uses unlabeled test data in re-projecting the entire available data (including the training plus test data) and re-training the classifiers in the re-projected space. However, in the first stage, the feature selection is done using only the training data. The proposed process is as follows:

1. Using the training set (which has known class labels for all samples) select a relatively large list of most-relevant features using the KL distance or 'mutual information' criterion, or a regularized feature selection method described in Sect. 2.2. The result is N_b discriminating or relevant features.

2. Using the list of N_b feature indices found in the previous step (i.e., using the labeled training set), determine the N_b-dimensional test (or unlabeled) set.

3. Using the collection of all N_b-dimensional labeled and unlabeled data (i.e., the combination of training and test sets) use manifold learning, or low-dimensional representation methods, such as KPCA discussed in Section 2.3.7, to project the data into a much lower-dimensional space (with dimensionality N_r). After this process, we will have N_r transformed or projected features. Then use the projected features of only the training set to design or learn the classification function. Then test the classifier on the projected features of the test set.

By the above method, since now a larger collection of data (as compared with the training and labeled set alone) is used to build the manifold and then the low-dimensional projection process (from N_b-dimensional space into N_r-dimensional space), this semi-supervised concatenated dimensionality reduction method might result in a better performance. Note that a major problem in manifold learning is the *sampling problem*, meaning that there are not enough data samples in the high-dimensional space to sufficiently and smoothly represent the shape of the manifold. Using the test data might help in this regard by providing a somewhat more smoothed manifold with a more dense distribution of data samples.

6.2.8 Data Fusion

Data fusion (see e.g. [208–210]) occurs at two main levels. The first level is 'feature-level data fusion' or 'signal fusion' in which the data and information from a variety of sources of information and features are combined in an efficient way. In the neuroscience application, for example, feature fusion means combining information from a variety of clinical and laboratory assessments and tests. The second level is 'processor data fusion', which is alternatively called 'algorithm fusion', 'decision fusion', or 'classifier fusion'. The parameters and structure of data fusion and processor fusion are determined using training data.

In the neuroscience application, for example, the clinical and laboratory data that are measured for a patient may come from different sources of information (such as clinical ratings, EEG, MRI, radiology, laboratory assessment, pharmacogenetic data, medical history, etc.). Data from each source has different properties and require proper information preprocessing, the output of which can be referred to as "raw feature data". The raw feature data goes through a feature extraction, feature ranking and discriminative feature selection procedure.

Also, a collection of preliminary or low-level classification/prediction models (or processors) are employed. Each method uses discriminative features to calculate preliminary estimation, decision and prediction results. In our research, each preliminary processor uses the feature data from all available data sources. Each preliminary processing model has its own numerical properties, and special processors/models are superior in particular cases. The idea of data fusion is to combine these results in an optimal way for more accurate and more robust performance in treatment planning and optionally in medical diagnosis. Data fusion can be regarded as a data-reduction mapping from multiple inputs of information into a smaller number of outputs. Another less optimal option to perform data fusion is to run the complete course of processing/analysis for each source of information separately, and then combine these high-level results afterwards.

There is cross-correlation, statistical dependency and shared information among estimates/decisions/outcomes reported by preliminary processors/models. An adaptive learning procedure can be used to determine the best implementation of a data fusion model, using the training set based on a determined optimality criterion. Ultimately the goal is to correct the error of each processor/model by the other processors/models. The basic assumption is that errors among processors/models are not the same for the same input data. Another simple 'processor fusion' method is to use a voting/ranking procedure, where each processor/model generates a decision variable instead of a continuous score. There are several voting and weighted averaging techniques, including a majority vote. See [208,209,211] for example, for a survey of such methods.

Statistical (or probabilistic) data fusion is another option to implement processor fusion. This fusion method is based on estimating statistical properties of processors and then combining them to satisfy a statistical optimality criterion (such as obtaining maximum likelihood, obtaining minimum probability of error, or minimizing the decision cost based on Bayesian decision risk). See [212], for example.

Bibliography

[1] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems*, vol. 12. Cambridge, MA: MIT Press, 2000, pp. 547–553.

[2] F. Rosenblatt, *Principles of Neurodynamics: Perceptron and Theory of Brain Mechanisms.* Spartan Books, 1962.

[3] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Prentice Hall, 2008.

[4] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. USA: Elsevier Academic Press, 2008.

[5] V. N. Vapnik, *Statistical Learning Theory.* John Wiley and Sons, 1998.

[6] P. Langley, *Elements of Machine Learning.* Morgan Kaufmann Publishers, 1996.

[7] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2008.

[9] V. E. Loskutov, A. F. Matvienko, B. V. Patramanskii, and V. E. Shcherbinin, "The magnetic method for in-tube nondestructive testing of gas and oil pipelines: The past and the present," *Russian Journal of Nondestructive Testing*, vol. 42, no. 8, pp. 493–504, Aug. 2006.

[10] R. C. Ireland and C. R. Torres, "Finite element modelling of a circumferential magnetiser," *Sensors and Actuators A (Physical)*, vol. 129, no. 1-2, pp. 197–202, May 2006.

[11] S. O'Connor, L. Clapham, and P. Wild, "Magnetic flux leakage inspection of tailor-welded blanks," *Measurement Science and Technology*, vol. 13, no. 2, pp. 157–162, Feb. 2002.

[12] W. Lord, *Nondestructive Testing Monographs and Tracts, Volume 3 – Electromagnetic Methods of Nondestructive Testing.* New York: Gordon and Breach Science Publishers, 1985.

[13] A. A. Carvalho, J. M. A. Rebello, L. V. S. Sagrilo, C. S. Camerini, and I. V. J. Miranda, "MFL signals and artificial neural networks applied to detection and classification of pipe weld defects," *NDT&E International*, vol. 39, no. 8, pp. 661–667, Dec. 2006.

[14] K. C. Hari, M. Nabi, and S. V. Kulkarni, "Improved FEM model for defect-shape construction from MFL signal by using genetic algorithm," *IET Science, Measurement and Technology*, vol. 1, no. 4, pp. 196–200, July 2007.

[15] A. Joshi, L. Udpa, S. Udpa, and A. Tamburrino, "Adaptive wavelets for characterizing magnetic flux leakage signals from pipeline inspection," *IEEE Transactions on Magnetics*, vol. 42, no. 10, pp. 3168–3170, Oct. 2006.

[16] R. K. Amineh, S. Koziel, N. K. Nikolova, J. W. Bandler, and J. P. Reilly, "A space mapping methodology for defect characterization from magnetic flux leakage measurements," *IEEE Trans. Magnetics*, vol. 44, no. 8, pp. 2058–2065, Aug. 2008.

[17] W. Han and P. Que, "A modified wavelet transform domain adaptive FIR filtering algorithm for removing the SPN in the MFL data," *Measurement*, vol. 39, no. 7, pp. 621–627, Aug. 2006.

[18] M. Afzal and S. Udpa, "Advanced signal processing of magnetic flux leakage data obtained from seamless gas pipeline," *NDT&E International*, vol. 35, no. 7, pp. 449–457, Oct. 2002.

[19] J. Tao, Q. Peiwen, C. Liang, and L. Liang, "Research on a recognition algorithm for offshore-pipeline defects during magnetic-flux inspection," *Russian Journal of Nondestructive Testing*, vol. 41, no. 4, pp. 231–238, Apr. 2005.

[20] S. Mukhopadhyay and G. P. Srivastava, "Characterisation of metal loss defects from magnetic flux leakage signals with discrete wavelet transform," *NDT&E International*, vol. 33, pp. 57–65, 2000.

[21] Y. Zhang, Z. Ye, and X. Xu, "An adaptive method for channel equalization in MFL inspection," *NDT&E International*, vol. 40, no. 2, pp. 127–139, Mar. 2007.

[22] Z. Zeng, L. Xuan, Y. Sun, L. Udpa, and S. Udpa, "Probability of detection model for gas transmission pipeline inspection," *Research in Nondestructive Evaluation*, vol. 15, no. 3, pp. 99–110, July/Sep. 2004.

[23] H. Goedecke, "Ultrasonic or MFL inspection: Which technology is better for you?" *Pipeline and Gas Journal*, vol. 230, no. 10, pp. 34–41, Oct. 2003.

[24] C. Mandache and L. Clapham, "A model for magnetic flux leakage signal predictions," *Journal of Physics D: Applied Physics*, vol. 36, no. 20, pp. 2427–2431, Oct. 2003.

[25] J. B. Nestleroth and R. J. Davis, "Application of eddy currents induced by permanent magnets for pipeline inspection," *NDT&E International*, vol. 40, no. 1, pp. 77–84, Jan. 2007.

[26] R. W. Tucker, S. W. Kercel, and V. K. Varma, "Characterization of gas pipeline flaws using wavelet analysis," in *Proceedings SPIE, Sixth International Conference on Quality Control by Artificial Vision*, 2003, pp. 485–493.

[27] R. K. Amineh, N. K. Nikolova, J. P. Reilly, and J. R. Hare, "Characterization of surface breaking cracks using one tangential component of magnetic leakage field," *IEEE Trans. Magnetics*, vol. 44, no. 4, pp. 516–524, Apr. 2008.

[28] S. Mandayam, L. Udpa, S. S. Udpa, and W. Lord, "Invariance transformations for magnetic flux leakage signals," *IEEE Trans. Magnetics*, vol. 32, no. 3, pp. 1577–1580, May 1996.

[29] J. Philip, C. B. Rao, T. Jayakumar, and B. Raj, "A new optical technique for detection of defects in ferromagnetic materials and components," *NDT&E International*, vol. 33, no. 5, pp. 289–295, July 2000.

[30] C. A. Beck and S. B. Patten, "Adjustment to antidepressant utilization rates to account for depression in remission," *Comprehensive Psychiatry*, vol. 45, no. 4, pp. 268–274, July 2004.

[31] C. A. Beck, S. B. Patten, J. V. A. Williams, J. L. Wang, S. R. Currie, C. J. Maxwell, and N. El-Guebaly, "Antidepressant utilization in Canada," *Social Psychiatry and Psychiatric Epidemiology*, vol. 40, no. 10, pp. 799–807, Oct. 2005.

[32] M. Jäger, P. Sobocki, and W. Rössler, "Cost of disorders of the brain in Switzerland– with a focus on mental disorders," *Swiss Medical Weekly*, vol. 138, no. 1–2, pp. 4–11, 2008.

[33] C. S. Dewa, A. Lesage, P. Goering, and M. Craveen, "Nature and prevalence of mental illness in the workplace," *Healthcare Papers*, vol. 5, no. 2, pp. 12–25, 2004.

[34] R. A. Kowatch, T. J. Carmody, G. J. Emslie, J. W. Rintelmann, C. W. Hughes, and A. J. Rush, "Prediction of response to fluoxetine and placebo in children and adolescents with major depression: a hypothesis generating study," *Journal of Affective Disorders*, vol. 54, no. 3, pp. 269–276, Aug. 1999.

[35] A. Serretti, P. Olgiati, M. N. Liebman, H. Hu, Y. Zhang, R. Zanardi, C. Colombo, and E. Smeraldi, "Clinical prediction of antidepressant response in mood disorders: Linear multivariate vs. neural network models," *Psychiatry Research*, vol. 152, no. 2–3, pp. 223–231, Aug. 2007.

[36] M. H. Trivedi, A. J. Rush, S. R. Wisniewski, A. A. Nierenberg, D. Warden, and e. a. L. Ritz, "Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice," *American Journal of Psychiatry*, vol. 163, no. 1, pp. 28–40, Jan. 2006.

[37] A. J. Rush, M. H. Trivedi, S. R. Wisniewski, A. A. Nierenberg, J. W. Stewart, and e. a. D. Warden, "Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report," *American Journal of Psychiatry*, vol. 163, no. 11, pp. 1905–1917, Nov. 2006.

[38] D. C. Malone, "A budget-impact and cost-effectiveness model for second-line treatment of major depression," *Journal of Managed Care Pharmacy*, vol. 13, no. 6 (Suppl A), pp. S8–S18, 2007.

[39] C. R. Young, M. B. B. Jr., and C. M. Mazure, "Management of the adverse effects of clozapine," *Schizophrenia Bulletin*, vol. 24, no. 3, pp. 381–390, 1998.

[40] V. Menon and S. Crottaz-Herbette, "Combined EEG and fMRI studies of human brain function," *International Review of Neurobiology*, vol. 66, pp. 291–321, 2005.

[41] P. L. Nunez and R. Srinivasan, *Electric fields of the brain: The neurophysics of EEG*, 2nd ed. Oxford University Press, 2006.

[42] E. Niedermeyer and F. L. da Silva, Eds., *Electroencephalography: Basic principles, clinical applications, and related fields*, 5th ed. USA: Lippincott Williams & Wilkins, 2004.

[43] J. R. Hughes and E. R. John, "Conventional and quantitative electroencephalography in psychiatry," *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 11, pp. 190–208, May 1999.

[44] K. L. Coburn, E. C. Lauterbach, N. N. Boutros, K. J. Black, D. B. Arciniegas, and C. E. Coffey, "The value of quantitative electroencephalography in clinical psychiatry: A report by the committee on research of the american neuropsychiatric association," *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 18, pp. 460–500, Nov. 2006.

[45] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR*, 4th ed. American Psychiatric Publishing Inc., June 2000.

[46] World Health Organization, "International statistical classification of diseases and related health problems (ICD), 10th revision," Available online by WHO: http://apps.who.int/classifications/apps/icd/icd10online/, 2007.

[47] G. M. Hasey, A. Khodayari-Rostamabad, J. P. Reilly, D. J. MacCrimmon, and H. de Bruin, "Expert system for determining patient treatment response," International Patent Application in WIPO, PCT/CA2009/000195, Feb. 2009.

[48] A. Khodayari-Rostamabad and J. P. Reilly, "SVM classifier approach to enumerate directional signals impinging on an array of sensors," in *Proceedings Canadian Conference on Electrical and Computer Engineering, CCECE'06.* Ottawa, ON: IEEE, May 2006, pp. 2190–2193.

[49] A. Khodayari-Rostamabad, J. P. Reilly, N. K. Nikolova, J. R. Hare, and S. Pasha, "Machine learning techniques for the analysis of magnetic flux leakage images in pipeline inspection," *IEEE Transactions on Magnetics*, vol. 45, no. 8, pp. 3073–3084, Aug. 2009.

[50] A. Khodayari-Rostamabad, G. M. Hasey, J. P. Reilly, D. J. MacCrimmon, and H. de Bruin, "A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy," accepted and to appear in: *Clinical Neurophysiology*, 2010, doi: 10.1016/j.clinph.2010.05.009.

[51] A. Khodayari-Rostamabad, J. P. Reilly, G. M. Hasey, H. DeBruin, and D. J. MacCrimmon, "Using pre-treatment EEG data to predict response to SSRI treatment for MDD," in *Proceedings 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*: accepted and to appear, 2010.

[52] ——, "Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model," in *Proceedings 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*: accepted and to appear, 2010.

[53] A. Khodayari-Rostamabad, J. P. Reilly, G. M. Hasey, H. de Bruin, and D. J. MacCrimmon, "A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder," to be submitted to: *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2010.

[54] A. Khodayari-Rostamabad, G. M. Hasey, J. P. Reilly, D. J. MacCrimmon, and H. de Bruin, "A Bayesian decision model for the diagnosis of psychiatric disorders," to be submitted to: *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2010.

[55] ——, "Analyzing pre-treatment EEG to predict the response to repetitive transcranial magnetic stimulation therapy in patients with treatment-resistant MDD," to be submitted to: *Journal of Psychiatric Research*, 2010.

[56] M. Ravan, J. P. Reilly, L. J. Trainor, and A. Khodayari-Rostamabad, "A machine learning approach for distinguishing age of infants using audio evoked potentials," submitted to: *Clinical Neurophysiology*, 2010.

[57] G. M. Hasey, A. Khodayari-Rostamabad, J. P. Reilly, D. J. MacCrimmon, and H. de Bruin, "Digital expert and neuro-psycho-biological signal processing as a predictor of response to treatment," US Provincial Patent Application 61/064177, Feb. 2008.

[58] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003.

[59] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994.

[60] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997.

[61] S. B. I. Guyon, J. Weston and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002.

[62] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 261–274, June 2008.

[63] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.

[64] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *The Journal of Machine Learning Research*, vol. 5, pp. 845–889, 2004.

[65] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[66] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.

[67] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. USA: John Wiley & Sons, 2006.

[68] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[69] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[70] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *'Subspace, Latent Structure and Feature Selection Techniques', Lecture Notes in Computer Science*, C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds. Springer, 2006, pp. 34–51.

[71] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[72] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, no. 1, pp. 1–50, Apr. 2000.

[73] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[74] I. Güler and E. D. Übeyli, "Multiclass support vector machines for EEG-signals classification," *IEEE Trans. Information Technology in Biomedicine*, vol. 11, no. 2, pp. 117–126, Mar. 2007.

[75] H. Abdi, *Encyclopedia of Measurement and Statistics.* Thousand Oaks, 2007, ch. Partial least square regression (PLS regression), pp. 740–744.

[76] R. Rosipal and L. J. Trejo, "Kernel partial least squares regression in reproducing kernel Hilbert space," *Journal of Machine Learning Research*, vol. 2, no. 2, pp. 97–123, 2002.

[77] N. Krämer and M. L. Braun, "Kernelizing PLS, degrees of freedom, and efficient model selection," in *Proceedings Int. Conf. Machine Learning*, 2007, pp. 441–448.

[78] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural Computation*, vol. 11, no. 2, pp. 305–345, 1999.

[79] H. Zhang, "The optimality of naive Bayes," in *Proceedings of the 17th International FLAIRS conf.*, 2004.

[80] C. Fraley and A. E. Raftery, "Model based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, June 2002.

[81] F. Dellaert, "The expectaion maximization algorithm," College of Computing, Georgia Institute of Technology, USA, Tech. Rep. GIT-GVU-02-20, Feb. 2002.

[82] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," University of Toronto, Toronto, Canada, Department of Computer Science Technical Report, CRG-TR-96-1, 1996.

[83] Z. Ghahramani and M. J. Beal, *Advances in Neural Information Processing Systems 12.* USA: MIT Press, 2000, ch. Variational inference for Bayesian mixtures of factor analysers, pp. 449–455.

[84] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

[85] N. Ueda, R. Nakano, Z. Ghahramani, and G. Hinton, "Pattern classification using a mixture of factor analyzers," in *Proceedings IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing*, Aug. 1999, pp. 525–534.

[86] J.-H. Zhao and P. L. H. Yu, "Fast ML estimation for the mixture of factor analyzers via an ECM algorithm," *IEEE Trans. Neural Networks*, vol. 19, no. 11, pp. 1956–1961, Nov. 2008.

[87] R. Kohavi, "A study of cross validation and bootstrap for accuracy estimation and model selection," in *Proceedings 14th Int. Joint Conf. on Artificial Intelligence*, 1995, pp. 1137–1143.

[88] R. Nishii and S. Tanaka, "Accuracy and inaccuracy assessments in land-cover classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 37, no. 1, pp. 491–498, Jan. 1999.

[89] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[90] S. Varma and R. Simon, "Bias in error estmation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 91, Feb. 2006.

[91] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. USA: Prentice Hall, 2007.

[92] J. C. Pesquet, H. Krim, and H. Carfantan, "Time-invariant orthonormal wavelet representations," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 1964–1970, Aug. 1996.

[93] J. Tenenbaum, V. d. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[94] L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[95] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2003.

[96] I. A. Cook, A. F. Leuchter, M. Morgan, E. Witte, W. F. Stubbeman, M. Abrams, S. Rosenberg, and S. H. Uijtdehaage, "Early changes in prefrontal activity characterize clinical responders to antidepressants," *Neuropsychopharmacology*, vol. 27, no. 1, pp. 120–131, July 2002.

[97] A. M. Hunter, I. A. Cook, and A. F. Leuchter, "The promise of the quantitative electroencephalogram as a predictor of antidepressant treatment outcomes in major depressive disorder," *Psychiatric Clinics of North America*, vol. 30, no. 1, pp. 105–124, Mar. 2007.

[98] H. Hinrikus, A. Suhhova, M. Bachmann, K. Aadamsoo, U. V. ohma, J. Lass, and V. Tuulik, "Electroencephalographic specral asymmetry index for detection of depression," *Medical and Biological Engineering and Computing*, vol. 47, pp. 1291–1299, 2009.

[99] V. Knott, C. Mahoney, S. Kennedy, and K. Evans, "EEG power, frequency, asymmetry and coherence in male depression," *Psyhiatry Research: Neuroimaging Section*, vol. 106, no. 2, pp. 123–140, Apr. 2001.

[100] J. S. Kwon, T. Youn, and H. Y. Jung, "Right hemisphere abnormalities in major depression: quantitative electroencephalographic findings before and after treatment," *Journal of Affective Disorders*, vol. 40, no. 3, pp. 169–173, Oct. 1996.

[101] G. E. Bruder, J. W. Stewart, C. E. Tenke, P. J. McGrath, P. Leite, N. Bhattacharya, and F. M. Quitkin, "Electroencephalographic and perceptual asymmetry differences between responders and nonresponders to

an SSRI antidepressant," *Biological Psychiatry*, vol. 49, no. 5, pp. 416–425, Mar. 2001.

[102] V. J. Knott, A. LaBelle, B. Jones, and C. Mahoney, "EEG coherence following acute and chronic clozapine in treatment-resistant schizophrenics," *Experimental & Clinical Psychopharmacology*, vol. 10, no. 4, pp. 435–444, Nov. 2002.

[103] G. Ulrich, H.-J. Haug, and E. Fähndrich, "Acute vs. chronic EEG effects in maprotiline- and clomipramine-treated depressive patients in the prediction of therapeutic outcome," *Journal of Affective Disorders*, vol. 32, pp. 213–217, 1994.

[104] G. E. Bruder, J. P. Sedoruk, J. W. Stewart, P. J. McGrath, F. M. Quitkin, and C. E. Tenke, "Electroencephalographic alpha measures predict therapeutic response to selective serotonin reuptake inhibitor antidepressant: pre- and post-treatment findings," *Biological Psychiatry*, vol. 63, pp. 1171–1177, 2008.

[105] S. H. Na, S.-H. Jin, S. Y. Kim, and B.-J. Ham, "EEG in schizophrenic patients: mutual information analysis," *Clinical Neurophysiology*, vol. 113, pp. 1954–1960, 2002.

[106] R. J. Davidson, "Anterior cerebral asymmetry and the nature of emotion," *Brain and Cognition*, vol. 20, no. 1, pp. 125–151, 1992.

[107] D. Pizzagalli, R. D. Pascual-Marqui, J. B. Nitschke, T. R. Oakes, C. L. Larson, H. C. Abercrombie, S. M. Schaefer, J. V. Koger, R. M. Benca, and R. J. Davidson, "Anterior cingulate activity as a predictor of degree of treatment response in major depression: evidence from brain electrical tomography analysis," *American Journal of Psychiatry*, vol. 158, no. 3, pp. 405–415, Mar. 2001.

[108] G. Racagni and M. Popoli, "The pharmacological properties of antidepressants," *International Clinical Psychopharmacology*, vol. 25, no. 3, pp. 117–131, May 2010.

[109] C. Mulert, G. Juckel, M. Brunnmeier, S. Karch, G. Leicht, R. Mergl, H. J. Moller, U. Hegerl, and O. Pogarell, "Rostral anterior cingulate cortex activity in the theta band predicts response to antidepressive medication," *Journal of Clinical EEG & Neuroscience*, vol. 38, no. 2, pp. 78–81, Apr. 2007.

[110] G. Ulrich, H. J. Haug, R. D. Stieglitz, and E. Fähndrich, "Are there distinct biochemical subtypes of depression? EEG characteristics of clinically defined on-drug responders and non-responders," *Journal of Affective Disorders*, vol. 15, no. 2, pp. 181–185, Sept.-Oct. 1988.

[111] V. Knott, C. Mahoney, S. Kennedy, and K. Evans, "Pre-treatment EEG and it's relationship to depression severity and paroxetine treatment outcome," *Pharmacopsychiatry*, vol. 33, no. 6, pp. 201–205, Nov. 2000.

[112] M. Bares, M. Brunovsky, M. Kopecek, P. Stopkova, T. Novak, J. Kozeny, and C. Hoschl, "Changes in QEEG prefrontal cordance as a predictor of response to antidepressants in patients with treatment resistant depressive disorder: a pilot study," *Journal of Psychiatric Research*, vol. 41, no. 3–4, pp. 319–325, Apr.-June 2007.

[113] e. M. Bares, "Early reduction in prefrontal theta QEEG cordance value predicts response to venlafaxine treatment in patients with resistant depressive disorder," *European Psychiatry*, vol. 23, no. 5, pp. 350–355, Aug. 2008.

[114] P. J. Nathan, R. Segrave, K. L. Phan, B. O'Neill, and R. J. Croft, "Direct evidence that acutely enhancing serotonin with the selective serotonin reuptake inhibitor citalopram modulates the loudness dependence of the auditory evoked potential (LDAEP) marker of central serotonin function," *Human Psychopharmacology: Clinical and Experimental*, vol. 21, no. 1, pp. 47–52, 2006.

[115] C. Mulert, G. Juckel, M. Brunnmeier, S. Karch, G. Leicht, R. Mergl, H.-J. Moller, U. Hegerl, and O. Pogarell, "Prediction of treatment response in major depression: integration of concepts," *Journal of Affective Disorders*, vol. 98, no. 3, pp. 215–225, Mar. 2007.

[116] S. V. Argyropoulos and S. J. Wilson, "Sleep disturbances in depression and the effects of antidepressants," *International Review of Psychiatry*, vol. 17, no. 4, pp. 237–245, 2005.

[117] A. S. Korb, I. A. Cook, A. M. Hunter, and A. F. Leuchter, "Brain electrical source differences between depressed subjects and healthy controls," *Brain Topography*, vol. 21, no. 2, pp. 138–146, Dec. 2008.

[118] S. C. Suffin, W. H. Emory, and L. J. Brandt, "Electroencephalography based systems and methods for selecting therapies and predicting outcomes," US Patent, US 7-177-675, 2007.

[119] S. D. Greenwald, C. P. Smith, J. C. Sigl, and P. H. Devlin, "System and method of assessment of neurological conditions using EEG," US Patent, US 7-231-245, 2007.

[120] G. E. Ott, U. Rao, K.-M. Lin, L. Gertsik, and R. E. Poland, "Effect of treatment with bupropion on EEG sleep: relationship to antidepressant response," *International Journal of Neuropsychopharmacology*, vol. 7, no. 3, pp. 275–281, Aug. 2004.

[121] M. Hatzinger, U. M. Hemmeter, S. Brand, M. Ising, and E. Holsboer-Trachsler, "Electroencephalographic sleep profiles in treatment course and long-term outcome of major depression: association with DEX/CRH-test response," *Journal of Psychiatric Research*, vol. 38, no. 5, pp. 453–465, Sept.-Oct. 2004.

[122] M. E. Thase, D. J. Buysse, E. Frank, C. R. Cherry, C. L. Cornes, A. G. Mallinger, and D. J. Kupfer, "Which depressed patients will respond to interpersonal psychotherapy? the role of abnormal EEG sleep profiles," *American Journal of Psychiatry*, vol. 154, no. 4, pp. 502–509, Apr. 1997.

[123] G. S. Malhi and J. Lagopoulos, "Making sense of neuroimaging in psychiatry," *Acta Psychiatrica Scandinavica*, vol. 117, no. 2, pp. 100–117, Nov. 2007.

[124] P. M. Grasby, "Imaging strategies in depression," *Journal of Psychopharmacology*, vol. 13, no. 4, pp. 346–351, 1999.

[125] W. C. Drevets, D. Öngür, and J. L. Price, "Neuroimaging abnormalities in the subgenual prefrontal cortex: implications for the pathophysiology of familial mood disorders," *Molecular Psychiatry*, vol. 3, no. 3, pp. 220–226, May 1998.

[126] H. S. Mayberg, S. K. Brannan, R. K. Mahurin, P. A. Jerabek, J. S. Brickman, J. L. Tekell, J. A. Silva, S. McGinnis, T. G. Glass, C. C. Martin, and P. T. Fox, "Cingulate function in depression: a potential predictor of treatment response," *Neuroreport: Clinical Neuroscience and Neuropathology*, vol. 8, no. 4, pp. 1057–1061, Mar. 1997.

[127] S. Saxena, A. L. Brody, M. L. Ho, N. Zohrabi, K. M. Maidment, and L. R. Baxter, "Differential brain metabolic predictors of response to paroxetine in obsessivecompulsion disorder versus major depression," *American Journal of Psychiatry*, vol. 160, pp. 522–532, Mar. 2003.

[128] R. J. Davidson, W. Irwin, M. J. Anderle, and N. H. Kalin, "The neural substrates of affective processing in depressed patients treated with venlafaxine," *American Journal of Psychiatry*, vol. 160, no. 1, pp. 64–75, Jan. 2003.

[129] G. E. Bruder, J. W. Stewart, J. D. Schaller, and P. J. McGrath, "Predicting therapeutic response to secondary treatment with bupropion: dichotic listening tests of functional brain asymmetry," *Psychiatry Research*, vol. 153, no. 2, pp. 137–143, Oct. 2007.

[130] U. Habel, M. Klein, T. Kellermann, N. J. Shah, and F. Schneider, "Same or different? neural correlates of happy and sad mood in healthy males," *NeuroImage*, vol. 26, no. 1, pp. 206–214, May 2005.

[131] G. S. Malhi, J. Lagopoulos, P. B. Ward, V. Kumari, P. B. Mitchell, G. B. Parker, B. Ivanovski, and P. Sachdev, "Cognitive generation of affect in bipolar depression: an fMRI study," *European Journal of Neuroscience*, vol. 19, no. 3, pp. 741–754, Jan. 2004.

[132] S. Posse, D. Fitzgerald, K. Gao, U. Habel, D. Rosenberg, G. J. Moore, and F. Schneider, "Real-time fMRI of temporolimbic regions detects amygdala activation during single-trial self-induced sadness," *NeuroImage*, vol. 18, no. 3, pp. 760–768, Mar. 2003.

[133] B. Bandelow, D. S. Baldwin, O. T. Dolberg, H. F. Andersen, and D. J. Stein, "What is the threshold for symptomatic response and remission for major depressive disorder, panic disorder, social anxiety disorder, and generalized anxiety disorder?" *Journal of Clinical Psychiatry*, vol. 67, no. 9, pp. 1428–1434, 2006.

[134] V. Henkel, F. Seemüller, M. Obermeier, M. Adli, and e. a. M. Bauer, "Does early improvement triggered by antidepressants predict response/remission? — Analysis of data from a naturalistic study on a large sample of inpatients with major depression," *Journal of Affective Disorders*, vol. 115, no. 3, pp. 439–449, June 2009.

[135] A. J. Rush, D. Warden, S. R. Wisniewski, M. Fava, M. H. Trivedi, B. N. Gaynes, and A. A. Nierenberg, "STAR*D: revising conventional wisdom," *CNS Drugs*, vol. 23, no. 8, pp. 627–647, Aug. 2009.

[136] G. Hasey, "Transcranial magnetic stimulation in the treatment of mood disorder: A review and comparison with electroconvulsive therapy," *Canadian Journal of Psychiatry*, vol. 46, pp. 720–727, 2001.

[137] P. B. Fitzgerald, T. L. Brown, N. A. Marston, Z. J. Daskalakis, A. D. Castella, and J. Kulkarni, "Transcranial magnetic stimulation in the treatment of depression: a double-blind, placebo-controlled trial," *Archives of General Psychiatry*, vol. 60, pp. 1002–1008, 2003.

[138] T. Wagner, A. Valero-Cabre, and A. Pascual-Leone, "Noninvasive human brain stimulation," *Annual Review of Biomedical Engineering*, vol. 9, pp. 527–565, 2007.

[139] D. J. L. G. Schutter, "Antidepressant efficacy of high-frequency transcranial magnetic stimulation over the left dorsolateral prefrontal cortex in double-blind sham-controlled designs: a meta-analysis," *Psychological Medicine*, vol. 39, pp. 65–75, 2009.

[140] F. Padberg and M. S. George, "Repetitive transcranial magnetic stimulation of the prefrontal cortex in depression," *Experimental Neurology*, vol. 219, pp. 2–13, 2009.

[141] P. B. Fitzgerald, K. Hoy, Z. J. Daskalakis, and J. Kulkarni, "A randomized trial of the anti-depressant effects of low- and high-frequency transcranial magnetic stimulation in treatment-resistant depression," *Depression & Anxiety*, vol. 26, pp. 229–234, 2009.

[142] J. P. Lefaucheur, "Methods of therapeutic cortical stimulation," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 39, pp. 1–14, 2009.

[143] F. Padberg, C. Schule, P. Zwanzger, T. Baghai, R. Ella, P. Mikhaiel, H. Hampel, H. J. Moller, and R. Rupprecht, "Relation between responses to repetitive transcranial magnetic stimulation and partial sleep deprivation in major depression," *Journal of Psychiatric Research*, vol. 36, pp. 131–135, 2002.

[144] G. Parker, D. Hadzi-Pavlovic, P. Boyce, K. Wilhelm, H. Brodaty, P. Mitchell, I. Hickie, and K. Eyers, "Classifying depression by mental state signs," *British Journal of Psychiatry*, vol. 157, pp. 55–65, 1990.

[145] F. Fregni, M. A. Marcolin, M. Myczkowski, R. Amiaz, G. Hasey, D. O. Rumi, M. Rosa, S. P. Rigonatti, J. Camprodon, M. Walpoth, J. Heaslip, L. Grunhaus, A. Hausmann, and A. Pascual-Leone, "Predictors of antidepressant response in clinical trials of transcranial magnetic stimulation," *International Journal of Neuropsychopharmacology*, vol. 9, pp. 641–654, 2006.

[146] E. L. Brakemeier, A. Luborzewski, H. Danker-Hopfe, N. Kathmann, and M. Bajbouj, "Positive predictors for antidepressive response to prefrontal repetitive transcranial magnetic stimulation (rTMS)," *Journal of Psychiatric Research*, vol. 41, pp. 395–403, 2007.

[147] E. L. Brakemeier, G. Wilbertz, S. Rodax, H. Danker-Hopfe, B. Zinka, P. Zwanzger, N. Grossheinrich, B. Varkuti, R. Rupprecht, M. Bajbouj, and F. Padberg, "Patterns of response to repetitive transcranial magnetic stimulation (rTMS) in major depression: Replication study in drug-free patients," *Journal of Affective Disorders*, vol. 108, pp. 59–70, 2008.

[148] S. H. Lisanby, M. M. Husain, P. B. Rosenquist, D. Maixner, R. Gutierrez, A. Krystal, W. Gilmer, L. B. Marangell, S. Aaronson, Z. J. Daskalakis, R. Canterbury, E. Richelson, H. A. Sackeim, and M. S. George, "Daily left prefrontal repetitive transcranial magnetic stimulation in the acute treatment of major depression: clinical predictors of outcome in a multisite, randomized controlled clinical trial," *Neuropsychopharmacology*, vol. 34, pp. 522–534, 2009.

[149] B. Langguth, R. Wiegand, A. Kharraz, M. Landgrebe, J. Marienhagen, U. Frick, G. Hajak, and P. Eichhammer, "Pre-treatment anterior cingulate activity as a predictor of antidepressant response to repetitive transcranial magnetic stimulation (rTMS)," *Neuroendocrinology Letters*, vol. 28, pp. 633–638, 2007.

[150] F. Schiffer, I. Glass, J. Lord, and M. H. Teicher, "Prediction of clinical outcomes from rTMS in depressed patients with lateral visual field stimulation: a replication," *Journal of Neuropsychiatry & Clinical Neurosciences*, vol. 20, pp. 194–200, 2008.

[151] J. Hongkui and M. Takigawa, "Observation of EEG coherence after repetitive transcranial magnetic stimulation," *Clinical Neurophysiology*, vol. 111, pp. 1620–1631, 2000.

[152] N. N. Boutros, A. P. Miano, R. E. Hoffman, and R. M. Berman, "EEG monitoring in depressed patients undergoing repetitive transcranial magnetic stimulation," *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 13, pp. 197–205, 2001.

[153] K. Iramina, T. Maeno, Y. Kowatari, and S. Ueno, "Effects of transcranial magnetic stimulation on EEG activity," *IEEE Trans. on Magnetics*, vol. 38, pp. 3347–3349, 2002.

[154] I. Griskova, O. Ruksenas, K. Dapsys, S. Herpertz, and J. Hoppner, "The effects of 10 hz repetitive transcranial magnetic stimulation on resting EEG power spectrum in healthy subjects," *Neuroscience Letters*, vol. 419, pp. 162–167, 2007.

[155] G. Fuggetta, E. F. Pavone, A. Fiaschi, and P. Manganotti, "Acute modulation of cortical oscillatory activities during short trains of high-frequency repetitive transcranial magnetic stimulation of the human motor cortex: a combined EEG and TMS study," *Human Brain Mapping*, vol. 29, pp. 1–13, 2008.

[156] G. W. Price, J. W. Lee, C. Garvey, and N. Gibson, "Appraisal of sessional EEG features as a correlate of clinical changes in an rTMS treatment of depression," *Clinical EEG and Neuroscience*, vol. 39, pp. 131–138, 2008.

[157] A. P. Funk and M. S. George, "Prefrontal EEG asymmetry as a potential biomarker of antidepressant treatment response with transcranial magnetic stimulation (TMS): a case series," *Clinical EEG and Neuroscience*, vol. 39, pp. 125–130, 2008.

[158] D. Spronk, M. Arns, A. Bootsma, R. van Ruth, and P. Fitzgerald, "Long-term effects of left frontal rTMS on EEG and ERPs in patients with depression," *Clinical EEG and Neuroscience*, vol. 39, pp. 118–124, 2008.

[159] A. Essali, N. A. Haj-Hasan, C. Li, and J. Rathbone, "Clozapine versus typical neuroleptic medication for schizophrenia," in *Cochrane Database of Systematic Reviews*, ser. Cochrane Reviews. John Wiley and Sons Ltd., 2009, no. 1, Art No. CD000059.

[160] W. Gunther, T. Baghai, D. Naber, R. Spatz, and H. Hippius, "EEG alterations and seizures during treatment with clozapine: a retrospective study of 283 patients," *Pharmacopsychiatry*, vol. 26, no. 3, pp. 69–74, May 1993.

[161] B. A. Malow, K. B. Reese, S. Sato, P. J. Bogard, A. K. Malhotra, S. Tung-Ping, and D. Pickar, "Spectrum of EEG abnormalities during clozapine treatment," *Electroencephalography and Clinical Neurophysiology*, vol. 91, no. 3, pp. 205–211, Sep. 1994.

[162] O. Freudenreich, R. D. Weiner, and J. P. McEvoy, "Clozapine-induced electroencephalogram changes as a function of clozapine serum levels," *Biological Psychiatry*, vol. 42, no. 2, pp. 132–137, July 1997.

[163] V. Knott, A. Labelle, B. Jones, and C. Mahoney, "Quantitative EEG in schizophrenia and in response to acute and chronic clozapine treatment," *Schizophrenia Research*, vol. 50, no. 1–2, pp. 41–53, May 2001.

[164] G. Adler, S. Grieshaber, V. Faude, B. Thebaldi, and H. Dressing, "Clozapine in patients with chronic schizophrenia: serum level, EEG and memory performance," *Pharmacopsychiatry*, vol. 35, no. 5, pp. 190–194, Sep. 2002.

[165] A. Birca, L. Carmant, A. Lortie, and M. Lassonde, "Interaction between the flash evoked SSVEPs and the spontaneous EEG activity in children and adults," *Clinical Neurophysiology*, vol. 117, no. 2, pp. 279–288, Feb. 2006.

[166] R. M. Dunki and M. Dressel, "Statistics of biophysical signal characteristics and state specificity of the human EEG," *Physica A: Statistical Mechanics and its Applications*, vol. 370, no. 2, pp. 632–650, Oct. 2006.

[167] T. Oikonomou, V. Sakkalis, I. G. Tollis, and S. Micheloyannis, "Searching and visualizing brain networks in schizophrenia," *Springer Lecture Notes in Computer Science: Biological and Medical Data Analysis*, vol. 4345, pp. 172–182, 2006.

[168] V. Sakkalis, T. Oikonomou, E. Pachou, I. Tollis, S. Micheloyannis, and M. Zervakis, "Time-significant wavelet coherence for the evaluation of schizophrenic brain activity using a graph theory approach," in *Proceedings Int. Conf. of the IEEE Engineering in Medicine and Biology*, Sep. 2006, pp. 4265–4268.

[169] N. N. Boutros, C. Arfken, S. Galderisi, J. Warrick, G. Pratt, and W. Iacono, "The status of spectral EEG abnormality as a diagnostic test for schizophrenia," *Schizophrenia Research*, vol. 99, no. 1–3, pp. 225–237, Feb. 2008.

[170] V. Knott, A. Labelle, B. Jones, and C. Mahoney, "EEG hemispheric asymmetry as a predictor and correlate of short-term response to clozapine treatment in schizophrenia," *Clinical Electroencephalography*, vol. 31, no. 3, pp. 145–152, July 2000.

[171] A. Gross, S. L. Joutsiniemi, R. Rimon, and B. Appelberg, "Clozapine-induced QEEG changes correlate with clinical response in schizophrenic patients: a prospective, longitudinal study," *Pharmacopsychiatry*, vol. 37, pp. 119–122, 2004.

[172] J. Gallinat and A. Heinz, "Combination of multimodal imaging and molecular genetic information to investigate complex psychiatric disorders," *Pharmacopsychiatry*, vol. 39, pp. S76–S79, 2006.

[173] N. F. Ince, F. Goksu, G. Pellizzer, A. Tewfik, and M. Stephane, "Selection of spectro-temporal patterns in multichannel MEG with support vector machines for schizophrenia classification," *Proc. Annual Int. Conf. IEEE Eng. in Medicine and Biology Society*, pp. 3554–3557, Aug. 2008.

[174] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "COMPARE: classification of morphological patterns using adaptive regional elements," *IEEE Trans. Medical Imaging*, vol. 26, no. 1, pp. 93–105, Jan. 2007.

[175] Y. Guo, F. D. Bowman, and C. Kilts, "Predicting the brain response to treatment using a Bayesian hierarchical model with application to a study of schizophrenia," *Human Brain Mapping*, vol. 29, pp. 1092–1109, 2008.

[176] D. Kim, J. Burge, T. Lane, G. D. Pearlson, K. A. Kiehl, and V. D. Calhoun, "Hybrid ICA–Bayesian network approach reveals distinct effective connectivity differences in schizophrenia," *Neuroimage*, vol. 42, no. 4, pp. 1560–1568, Oct. 2008.

[177] J. Struyf, S. Dobrin, and D. Page, "Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia," *BMC Genomics*, vol. 9, no. 531, Nov. 2008.

[178] C. Lin, Y. Wang, J. Chen, Y. Liou, Y. Bai, I. Lai, T. Chen, H. Chiu, and Y. Li, "Artificial neural network prediction of clozapine response with combined pharmacogenetic and clinical data," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 2, pp. 91–99, Aug. 2008.

[179] J. Kane, G. Honigfeld, J. Singer, H. Meltzer, and the Clozaril Collaborative Study Group, "Clozapine for the treatment-resistant schizophrenic: A double-blind comparison with chlorpromazine," *Archives of General Psychiatry*, vol. 45, pp. 789–796, Sept. 1988.

[180] S. R. Kay, A. Fiszbein, and L. A. Opler, "The positive and negative syndrome scale (PANSS) for schizophrenia," *Schizophrenia Bulletin*, vol. 13, pp. 261–276, 1987.

[181] S. Leucht, J. M. Kane, W. Kissling, J. Hamann, E. Etschel, and R. R. Engel, "What does the PANSS mean?" *Schizophrenia Research*, vol. 79, no. 231–238, 2005.

[182] G. Okugawa, G. C. Sedvall, and I. Agartz, "Reduced grey and white matter volumes in the temporal lobe of male patients with chronic schizophrenia," *European Archives of Psychiatry and Clinical Neuroscience*, vol. 252, pp. 120–123, 2002.

[183] S. F. Faux, M. E. Shenton, R. W. McCarley, M. W. Torello, and F. H. Duffy, "P200 topographic alterations in schizophrenia: evidence for left temporal-centroparietal region deficits," *Electroencephalography and Clinical Neurophysiology Supplement*, vol. 40, pp. 681–687, 1987.

[184] Z. J. Daskalakis, B. K. Christensen, P. B. Fitzgerald, B. Moller, S. I. Fountain, and R. Chen, "Increased cortical inhibition in persons with schizophrenia treated with clozapine," *Journal of Psychopharmacology*, vol. 22, no. 2, pp. 203–209, Feb. 2008.

[185] G. Winterer, M. Smolka, J. Samochowiec, and *et al.*, "Association of EEG coherence and an exonic GABA$_B$R1 gene polymorphism," *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, vol. 117B, no. 1, pp. 51–56, Feb. 2003.

[186] V. Kusumakar, "Antidepressants and antipsychotics in the long-term treatment of bipolar disorder," *Journal of Clinical Psychiatry*, vol. 63, no. Suppl 10, pp. 23–28, 2002.

[187] A. J. Niemiec and B. J. Lithgow, "Alpha-band characteristics in EEG spectrum indicate reliability of frontal brain asymmetry measures in diagnosis of depression," in *Proceedings of Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, Sep. 2005, pp. 7517–7520.

[188] P. Coutin-Churchman, Y. Anez, M. Uzcategui, L. Alvarez, F. Vergara, L. Mendeza, and R. Fleitas, "Quantitative spectral analysis of EEG in psychiatry revisited: drawing signs out of numbers in a clinical setting," *Clinical Neurophysiology*, vol. 114, no. 12, pp. 2294–2306, Dec. 2003.

[189] F. Ferrarelli, R. Huber, M. J. Peterson, M. Massimini, M. Murphy, B. A. Riedner, A. Watson, P. Bria, and G. Tononi, "Reduced sleep spindle activity in Schizophrenia patients," *American Journal of Psychiatry*, vol. 164, no. 3, pp. 483–492, Mar. 2007.

[190] Q. Hongzhi, W. Baikun, and Z. Li, "Mutual information entropy research on dementia EEG signals," in *Proceedings of Int. Conf. on Computer and Information Technology*, Sep. 2004, pp. 885–889.

[191] D. Abasolo, R. Hornero, P. Espino, D. Alvarez, and J. Poza, "Entropy analysis of the EEG background activity in Alzheimer's disease patients," *Physiological Measurement*, vol. 27, no. 3, pp. 241–253, Mar. 2006.

[192] J. A. Turner, S. G. Potkin, G. G. Brown, D. B. Keator, G. McCarthy, and G. H. Glover, "Neuroimaging for the diagnosis and study of psychiatric disorders," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 112–117, July 2007.

[193] Y.-J. Li and F.-Y. Fan, "Classification of Schizophrenia and depression by EEG with ANNs," in *Proceedings of Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, Sep. 2005, pp. 2679–2682.

[194] A. Deslandes, H. Veiga, M. Cagy, A. Fiszman, R. Piedade, and P. Ribeiro, "Quantitative electroencephalography (qEEG) to discriminate primary degenerative dementia from major depressive disorder (depression)," *Arquivos de Neuro-Psiquiatria*, vol. 62, no. 1, pp. 44–50, Mar. 2004.

[195] S. Ghosh-Dastidar, H. Adeli, and N. Dadmehr, "Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection," *IEEE Trans. Biomedical Engineering*, vol. 55, no. 2, pp. 512–518, Feb. 2008.

[196] S. Sun, C. Zhang, and Y. Lu, "The random electrode selection ensemble for EEG signal classification," *Pattern Recognition*, vol. 41, no. 5, pp. 1680–1692, May 2008.

[197] B. Saletu, P. Anderer, G. M. Saletu-Zyhlarz, and R. D. Pascual-Marqui, "EEG mapping and low-resolution brain electromagnetic tomography (LORETA) in diagnosis and therapy of psychiatric disorders: evidence for a key-lock principle," *Clinical EEG & Neuroscience: Official Journal of the EEG & Clinical Neuroscience Society (ENCS)*, vol. 36, no. 2, pp. 108–115, Apr. 2005.

[198] W.-R. Zhang, A. K. Pandurangi, and K. E. Peace, "Yin Yang dynamic neurobiological modeling and diagnostic analysis of major depressive and

bipolar disorders," *IEEE Trans. Biomedical Engineering*, vol. 54, no. 10, pp. 1729–1739, Oct. 2007.

[199] A. Gross, S. L. Joutsiniemi, R. Rimon, and B. Appelberg, "Correlation of symptom clusters of schizophrenia with absolute powers of main frequency bands in quantitative EEG," *Behavioral and Brain Functions*, vol. 2, no. 23, 2006.

[200] D. Heckerman, "A tutorial on learning with Bayesian networks," Microsoft Research, Tech. Rep. MSR-TR-95-06, Mar. 1996.

[201] K. P. Murphy, "An introduction to graphical models," Tech. Rep., May 2001.

[202] M. I. Jordan, Ed., *Learning in Graphical Models.* MIT Press, 1999.

[203] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning.* The MIT Press, 2006.

[204] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, pp. 1–48, 2006.

[205] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *Proceedings 22nd International Conf. on Machine Learning: ICML*, 2005, pp. 825–832.

[206] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings International Conference on Machine Learning*, 1999, pp. 200–209.

[207] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 34, pp. 1–38, 1977.

[208] A. Sinha, C. Huimin, D. G. Danu, T. Kirubarajan, and M. Farooq, "Estimation and decision fusion: a survey," in *Proceedings IEEE Int. Conf. on Engineering of Intelligent Systems*, Apr. 2006, pp. 1–6.

[209] V. Torra, *Information fusion in data mining.* Springer, 2003.

[210] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, Jan. 1997.

[211] G. Fumera and F. Roli, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 942–956, June 2005.

[212] O. R. Terrades, E. Valveny, and S. Tabbone, "Optimal classifier fusion in a non-Bayesian probabilistic framework," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1630–1644, Sept. 2009.